



**CZECH TECHNICAL  
UNIVERSITY  
IN PRAGUE**

**F3**

**Faculty of Electrical Engineering  
Department of Cybernetics**

**Ph.D. Thesis**

# **Analysis of biochemical data**

**Ing. Jiří Anýž**

**Study programme: P2612 Electrical Engineering and Information Technology**

**Branch of study: 3902V035 Artificial Intelligence and Biocybernetics**

**Prague, August 2019**

**Supervisor: Prof. RNDr. Olga Štěpánková, CSc.**



## Acknowledgement / Declaration

I would like to express my thanks to my supervisor prof. RNDr. Olga Štěpánková, CSc. for a careful guidance of my doctoral studies and for helping me with the summarisation of our collaboration into this thesis.

I would like to thank to doc. Ing. Daniel Novák, Ph.D. for introducing me to challenging problems in biomedical data analysis.

I would also like to thank my colleagues Mgr. Tomáš Sieger, Ph.D. and Ing. Eduard Bakštein, Ph.D. for many discussions on the topics of data visualisation and analysis.

And many thanks go to my family for support and patience.

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others is acknowledged in the text and the list of references is provided.

Prague

August 2019

## Abstrakt /

Chemie a biochemie prošly významným vývojem laboratorní techniky, který způsobil, že měřená data jsou mnohem složitější a větší než kdy předtím. Analýza chemických a biochemických dat vyžaduje spolupráci chemiků a biologů s datovými analytiky a statistiky. Tato práce prezentuje několik obecných závěrů o analýze chemických a biochemických dat. Data jsou vysoce dimenzionální, obvykle s méně pozorováními než proměnnými. Z povahy dat jsou proměnné obvykle kolineární. To vymezuje specifické požadavky na metody datové analýzy. Práce nabízí přehled metod vhodných pro tento typ dat a věnuje zvláštní pozornost generalizovaným mixed effect modelům a jejich výhodám při modelování vztahů v komplexních datech. Významná část práce je věnovaná metodám založených na simulacích pro statistickou inferenci a porovnání modelů.

Hlavní část práce se zabývá analýzou dat z experimentu na melanomem trpícím Liběchovským Minipraseti, laboratorním plemeni miniprasete trpícím melanomem, které slouží jako *in vivo* model progresu melanomu. Biologické vzorky byly analyzovány laserovou ablací ve spojení s hmotnostní spektrometrií indukčně vázaného plazmatu, metodou která produkuje prostorové mapy chemických prvků ve tkáňových řezech.

Abychom prozkoumali vztahy mezi prostorovým rozložením bioaktivních chemických prvků (zinek, měď), histologicky konzistentní tkáňové zóny identifikované v sousedícím tkáňovém řezu byly zpropagovány pomocí vytvoření vrstvené reprezentace dat provedené registrací obrazů. Vrstvená reprezentace dat umožňuje použití standardních metod analýzy dat. Shlukování na vrstvených datech ukazuje, že

i po zpracování vzorku pomocí laserové ablace jsme schopni rozlišit obecné kategorie tkání jako je melanomová tkáň a vazivová tkáň. Tento postup nám také umožnil provést analýzu prostorové struktury melanomové tkáně. Analýza prostorové struktury melanomu nám poskytuje oporu v plánování podobných experimentů – tkáň v melanomu si je podobná až do vzdálenosti 24  $\mu\text{m}$ .

Anotované tkáňové zóny byly využity pro popis rozdílů v mapách chemických prvků zinku a mědi. S použitím navrhovaného postupu zpracování dat, lineárních mixed effect modelů, které umožňují zohlednit individuální rozdíly mezi zvířaty, a neparametrického bootstrapu pro statistickou inferenci, jsme byli schopni ukázat, že obsah zinku v rostoucí melanomové tkáni a tkáni s počínající spontánní regresí je signifikantně vyšší než při pozdní spontánní regresí a vazivové tkáni.

Navrhovaný postup zpracování dat otevírá nové možnosti pro zhodnocení experimentů s laserovou ablací s hmotnostní spektrometrií indukčně vázaného plazmatu a podobnými metodami. Použití komplexních modelů zpřesňuje pozorované vztahy v datech z přítomnosti nezávislé i závislé variability. S využitím na simulacích založených metod zlepšujeme spolehlivost statistické inference stejně tak jako reprodukovatelnost.

**Klíčová slova:** mapování chemických prvků, LA-ICP-MS, generalizovaný lineární mixed effect model, simulační metody, histology, melanom, supervizovaná PCA, metabolomika, NMR spektroskopie

## Abstract /

The chemistry and biochemistry witnessed substantial instrumental advances, which resulted in a measurement of ever more complex and bigger data. The analysis of data from modern chemical and biochemical laboratory analyses requires dedicated support from data scientists or statisticians. In this work, we present several general findings of the chemical and biochemical data analysis. The data are high-dimensional, usually have small number of observations and a large number of variables. Due to the nature of the data, the variables are generally collinear. This property imposes specific limitations on the data analysis methods. We present and discuss the recommended methods and pipelines. We emphasize the generalised mixed effect models as an efficient tool for modelling relationships in complex data. An important part of the work is devoted to simulation-based methods for statistical inference as well as for model comparison.

The main part of the work deals with the practical analysis of data from an experiment on melanoma-bearing Libechov minipigs. This animal is a laboratory breed of pigs suffering from melanoma used as an *in vivo* model of melanoma progression. The biological samples of melanoma were examined by laser ablation inductively coupled plasma mass spectrometry that produces spatial maps of elements in tissue sections.

We integrated the elemental maps and histology scans using image registration, creating the layered data representation. The layered data representation enables using standard data analysis procedures to examine the relationships between the annotated tissue zones and the spatial distribution of bioactive metals (zinc, copper). The clustering

on the layered data shows that broad categories such as melanoma tissue and fibrous tissue can be distinguished. The analysis of the melanoma structure provides us with justifications for the planning of experiments in a similar manner – the tissue in the melanoma is similar for areas up to 24  $\mu\text{m}$  apart.

The annotated tissue zones were utilised to examine the differences in the elemental maps of zinc and copper. We showed that the zinc content in growing melanoma tissue and early spontaneous regression tissue is significantly higher than in the late spontaneous regression tissue and the fibrous tissue. The linear mixed effect model takes the individual differences among the animals into account, and the non-parametric bootstrap for the statistical inference provides more reliable results than the standard approach.

The data processing pipeline opens new possibilities for the evaluation of experiments involving spatial data such as the elemental maps produced by the laser ablation inductively coupled plasma mass spectrometry. The use of complex models refines the observed relationships in the data among the influence of independent and dependent variability. The simulation-based non-parametric bootstrap improves the reliability of statistical assessment as well as the reproducibility.

**Keywords:** chemical elemental mapping, LA-ICP-MS, generalised linear mixed effect model, simulation-based methods, histology, melanoma, MeLiM, supervised PCA, metabolomics, NMR spectroscopy

# Contents /

<b>1 Introduction</b> .....	1
1.1 Goals of the thesis .....	2
1.2 Structure of the thesis .....	2
<b>2 Problem statement</b> .....	4
2.1 Introduction .....	4
2.2 Example – Gel electrophoresis ..	4
2.3 The biochemical data .....	6
2.3.1 Brdička curve .....	6
2.3.2 Electrophoreogram im- age .....	7
2.3.3 Nuclear Magnetic Res- onance spectroscopy .....	8
2.3.4 Elemental mapping by LA-ICP-MS .....	10
2.3.5 Additional information supplemented to the biochemical data .....	12
2.4 Problem formulation .....	13
2.5 Summary .....	14
<b>3 Specific processing of bio- chemical data</b> .....	15
3.1 Feature extraction .....	15
3.1.1 Data-dictated feature extraction .....	16
3.1.2 Data-non-adaptive fea- ture extraction .....	17
3.1.3 Statistical models for feature extraction .....	19
3.1.4 Summary .....	19
3.2 Combination of data from various sources .....	20
3.3 Covariance, correlation and collinearity .....	23
3.3.1 Variability .....	24
3.3.2 Covariance .....	24
3.3.3 Correlation .....	24
3.3.4 Collinearity .....	25
3.3.5 Summary .....	29
3.4 Data analysis methods for biochemical data .....	30
3.5 Machine learning approach for the analysis of data in biochemistry .....	30
3.5.1 Unsupervised methods...	31
3.5.2 Supervised methods .....	33
3.5.3 Summary .....	37
3.5.4 Statistical approach to the data analysis in biochemistry .....	38
3.5.5 Multiple comparisons ....	41
3.5.6 Statistical modelling – regression .....	43
3.5.7 Summary .....	45
<b>4 Generalised mixed effect mod- els</b> .....	47
4.1 LMM model description .....	47
4.2 LMM parameters estimation ..	48
4.2.1 Maximum likelihood estimation of LMM pa- rameters .....	48
4.2.2 Restricted maximum likelihood estimation of LMM parameters .....	49
4.2.3 Numerical optimisa- tion methods for the LMM parameters esti- mation .....	49
4.3 Statistical inference of the LMM .....	50
4.3.1 Standard LMM infer- ence .....	50
4.3.2 LMM inference based on simulations .....	51
4.4 Generalised LMM .....	53
4.5 Bayesian approach to the LMM and GLMM .....	54
<b>5 Comparison of methods in NMR spectroscopy</b> .....	57
5.0.1 Projection to latent structures .....	58
5.0.2 Supervised principal component analysis .....	58
5.0.3 Model validation .....	59
5.0.4 Model interpretation and assessment .....	59
5.1 Methods .....	60
5.1.1 Hypotheses formulation .	60
5.1.2 Description of data .....	60
5.1.3 Alteration by known signals .....	61
5.1.4 NMR spectra process- ing .....	62

5.1.5	Methods comparison.....	62	7.3.1	Publications related to the topic of the thesis ...	98
5.1.6	Single iteration of the comparison procedure ...	62	7.3.2	Publications unrelat- ed to the topic of the thesis .....	100
5.1.7	Evaluation of results from repeated compar- isons .....	64	<b>References</b> .....		102
5.2	Results .....	64	<b>A List of abbreviations</b> .....		113
5.2.1	Data description .....	65	A.1	Biochemical and other mea- surement methods .....	113
5.2.2	Simulations.....	65	A.2	Biological and biochemical terminology .....	113
5.3	Discussion .....	66	A.3	Machine learning and statis- tical methods.....	113
5.4	Conclusion .....	68			
<b>6</b>	<b>Element mapping</b> .....	69			
6.1	Introduction.....	69			
6.2	Data collection .....	69			
6.2.1	Histology .....	69			
6.2.2	LA-ICP-MS .....	70			
6.3	Data description .....	73			
6.3.1	Histology .....	73			
6.3.2	LA-ICP-MS .....	74			
6.3.3	Additional information ..	75			
6.4	Methods .....	76			
6.4.1	Spatial covariance.....	76			
6.4.2	Data integration .....	79			
6.4.3	Spatial properties of the melanoma tissue .....	83			
6.4.4	Unsupervised analysis of histological zones and elemental maps.....	84			
6.4.5	Statistical analysis of the differences in the metals spatial distribu- tion .....	84			
6.5	Results .....	85			
6.5.1	Data integration .....	85			
6.5.2	Spatial properties of the melanoma tissue .....	86			
6.5.3	Unsupervised analysis histological zones and elemental maps .....	87			
6.5.4	Statistical analysis .....	88			
6.6	Discussion .....	91			
<b>7</b>	<b>Conclusion</b> .....	94			
7.1	Achieved goals of the thesis ...	96			
7.2	Future work .....	97			
7.3	List of author's publications...	98			

## Tables / Figures

<p><b>3.1.</b> Summary of ML methods..... 38</p> <p><b>5.1.</b> Metabolites used in comparison ..... 63</p> <p><b>5.2.</b> F1 score on simulated sets..... 65</p> <p><b>5.3.</b> Area under receiver operating characteristics on simulated data..... 66</p> <p><b>5.4.</b> Sensitivity on simulated data.. 67</p> <p><b>5.5.</b> Specificity score from simulated data..... 68</p> <p><b>6.1.</b> The relationship between the average Dice coefficient and the distance of the tissue sections for images registered by the elastic transformation..... 87</p> <p><b>6.2.</b> Modelled differences in Zn content among tissue zones .... 90</p> <p><b>6.3.</b> Modelled differences in Cu content among tissue zones .... 90</p> <p><b>6.4.</b> Modelled differences in Zn content among revised tissue zones (GMT x SR&amp;FT) ..... 91</p> <p><b>6.5.</b> Modelled differences in Zn content among revised tissue zones (GMT&amp;ESR x LSR&amp;FT)..... 91</p> <p><b>6.6.</b> Modelled differences in Cu content among revised tissue zones (GMT x SR&amp;FT) ..... 92</p> <p><b>6.7.</b> Modelled differences in Cu content among revised tissue zones (GMT&amp;ESR x LSR&amp;FT)..... 92</p>	<p><b>2.1.</b> Electrophoresis apparatus .....5</p> <p><b>2.2.</b> Example of Brdička curve .....7</p> <p><b>2.3.</b> Example of electrophoreogram ..8</p> <p><b>2.4.</b> NMR spectrometer .....8</p> <p><b>2.5.</b> NMR spectrum .....9</p> <p><b>2.6.</b> LA-ICP-MS apparatus ..... 11</p> <p><b>2.7.</b> Pipeline of data integration.... 12</p> <p><b>2.8.</b> General pipeline of biochemical data processing ..... 13</p> <p><b>3.1.</b> Electrophoreogram, band and extracted curve ..... 16</p> <p><b>3.2.</b> Brdička curve - a decomposition of the curve into waveforms..... 17</p> <p><b>3.3.</b> Example of Fourier transform . 18</p> <p><b>3.4.</b> Time-frequency resolution of STFT and WT ..... 19</p> <p><b>3.5.</b> Arguments for image registration ..... 22</p> <p><b>3.6.</b> NMR spectrum of simple compound ..... 26</p> <p><b>3.7.</b> PCA scores ..... 33</p> <p><b>3.8.</b> Decision tree..... 34</p> <p><b>3.9.</b> Illustration of kernel method .. 35</p> <p><b>5.1.</b> Simulation flowchart ..... 61</p> <p><b>6.2.</b> Parameters of laser ablation... 74</p> <p><b>6.4.</b> Arguments for image registration ..... 80</p> <p><b>6.5.</b> SSD criterion of image dissimilarity ..... 83</p> <p><b>6.6.</b> Clustering based segmentation of histology image..... 85</p> <p><b>6.7.</b> Average content of Zn and Cu in all samples..... 89</p> <p><b>6.8.</b> Zn and Cu content in different samples ..... 89</p>
--	---



# Chapter 1

## Introduction

Chemistry and biochemistry witnessed substantial instrumental advances. The improvement in laboratory techniques helped to understand many chemical and biochemical phenomena. However, it also led to the generation of large amounts of data [1]. Let us take the structure of the deoxyribonucleic acid (DNA) discovered by Watson and Crick [2] as an example. In the case of the decoding of the information in the DNA, the breakthrough came with the Human genome project and the techniques of DNA sequencing [3] which identified 14.8 billion base pairs of the human genome over nine months. Nowadays the DNA sequencing is cheaper and faster in order of magnitude with next-generation sequencing methods [4].

And in other sciences, the amount of data followed the same pattern [5]. The new generation measuring devices can generate vast amounts of data that many times overreach the abilities of the laboratory staff to process the data with the contemporary approaches relying heavily on a manual treatment of the data or in the utilisation of simple programs (usually supplied by the producer of the measuring device) for processing small batches of data. This development inevitably led to the establishing of new scientific disciplines combining chemistry and biochemistry with computer science, machine learning and data mining. These interdisciplinary sciences are among other the computational biology [6] and the biostatistics [7]. In the context of medicine, it is the biomedical informatics [8].

All fields of study deal to a certain extent with similar problems when they try to understand and interpret the measurements. Few topics summarise the issues common to all the sciences dealing with data. The data collection and preprocessing, feature extraction, unsupervised and supervised analysis, interpretation of results are the problems to be encountered during analysis of data originating in biochemistry [9]. The new trend in the data sciences is to develop data analysis pipelines performing all the necessary preprocessing steps specific to the data and producing the statistical assessment. A well designed pipeline offers the scientist the opportunity to detach himself from the tedious data analysis and to have results that point towards promising directions of further research or that are ready for publication of a new set of findings. There are generally two approaches:

- the time consuming manual processing of the data by a team of data scientists, which check the consistency of the data, the potential confounding factors, sources of variability, patterns of missing data and other important questions;
- a fast throughput analysis by a pipeline, which produces the desired results in fraction of time of the first approach, however the the pipeline may not fit all the data.

Each of these approaches has its supporters and opponents, its advantages and drawbacks. In a team of specialists, the enthusiastic data scientists would be in favour of the first approach, because it is their expertise. They are aware of the many issues related to any data analysis. On the other hand, other team members are not always convinced that careful data analysis is better than an automated pipeline. The presented view is



data modelling common in nuclear magnetic resonance spectroscopy. **Section 5.1** describes the pipeline for a simulation-based comparison of methods. **Section 5.2** summarises the results of the comparison.

**Chapter 6** is a case study for the methods and approaches introduced in the previous chapters. **Sections 6.4.3 and 6.4.4** explore the structure of the melanoma tissue. **Section 6.4.2** introduces data integration method suitable for elemental mapping and digital microscopy. **Section 6.4.5** provides a statistical framework for the testing of differences in developmental stages of melanoma tissue.

**Chapter 7** concludes the thesis. The **Section 7.1** summarises the achieved goals of the thesis. The **Section 7.2** follows with the perspectives for future work. The **Section 7.3** lists the author's contributions and scientific publications.

# Chapter 2

## Problem statement

### 2.1 Introduction

The data originating in chemical and biochemical laboratories can be very diverse. It is not the aim of this work to provide an exhaustive list of possible outcomes of chemical experiments. On the other hand, the data share many properties. The measurements are usually vectors, matrices or multidimensional arrays. Specific preprocessing techniques have to be utilised to preprocess the measurements. And in the data, there are relationships among the variables, therefore the data are collinear.

The common issue with the data originating in chemistry and biochemistry are the need to preprocess the data to remove the uncertainties introduced by the measurement process [11–13]. In connection with the need to preprocess the data comes the need to effectively extract features of the original data for following steps of the analysis. After the preprocessing stage, the exploratory visualisations and tables help data inspection and examination. This step can help to discover various sources of variability in the data. These can be interactions among the variables, differences between groups of observations, patterns of missing data, or anomalies. The variability in the data is often of two types:

- variability correlated to the experiment design
- variability orthogonal to the experiment design

The experiment outcome often expressed as a response variable (e.g. level of a metabolite in body fluid), classification of observations (e.g. classification of patients as suffering from a disease or healthy) [14]. The analysis of the data employing unsupervised and supervised machine learning methods can exploit the variability related to the study design.

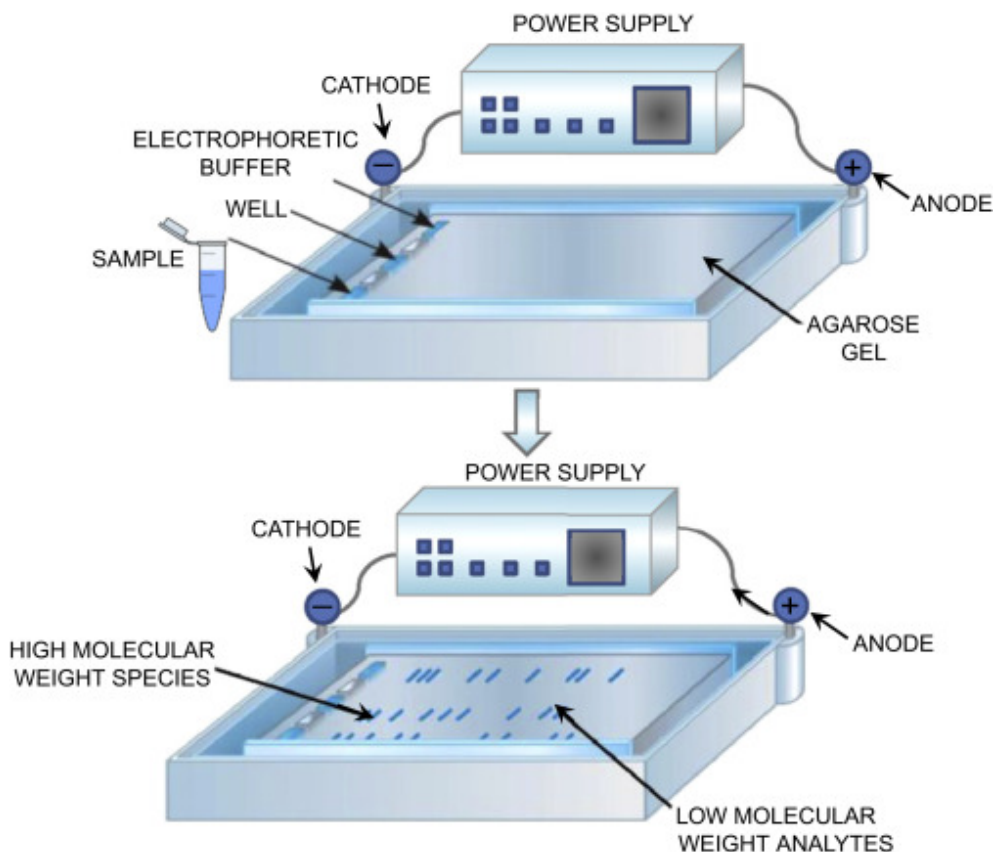
On the other hand, the orthogonal variability is often a nuisance during the data analysis and specific methods were developed to remove or suppress it [15, 14].

### 2.2 Example – Gel electrophoresis

As an example of the problems dealt with during the data analysis may serve the study of electrophoreograms. The electrophoreograms are the results of electrophoresis. A digital image obtained by camera serves for the examination. The process of taking the photography alone is a source of several severe problems. The illumination of the electrophoreogram is often not even. The person taking the photo often does not use any apparatus to support the camera, which leads to distortions in the image. The misalignment between the camera and the photographed electrophoreogram leads to a perceived perspective in the resulting image. The second source of uncertainty in the data is the measurement process itself. The gel matrices used to the separation of

the examined (biological) material into its compounds are often inhomogeneous, and when inserted into the apparatus, they distort the electrical field. The design of the equipment also makes the electrical field uneven, see Fig 2.1. When combining a series of experiments, the alignment of the electrophoreograms by image transformation is necessary to eliminate the lack of standardisation.

The electrophoreograms are always equipped with a standard band to provide a reference for matching between diverse electrophoreograms. Image processing methods can solve the issues connected to the taking of the image as well as with the differing results in repeated measurements. The elimination of these problems does not lead to data suitable for processing by machine learning or statistical methods. A feature extraction procedure has to be employed to obtain variables describing the actual interesting characteristics of the electrophoreograms. The individual bands representing the analysed (biological) material have to be recognised and subjected to a feature extraction algorithm. In the case of electrophoreogram, the average brightness curves [J6, J5, J4] can represent the distribution of molecules according to their specific mass. A collection of brightness curves provides information in a form that can be utilised by data analysis methods for time series data. The most suitable are methods performing further feature extraction and distinguishing correlated and orthogonal variability such as [14] in [J5].



**Figure 2.1.** Electrophoresis apparatus. The analysed samples are loaded into the wells in the gel matrix. The power supply generates an electrical field which attracts/repels the compounds. The mobility of the compounds depends on the molecular weight and the electric charge. The low molecular weight compounds travel farther than the high molecular weight compounds. [16]

The data analysis of the collection of the brightness curves may benefit from the second application of feature extraction procedure. In the first case, the feature extraction was a part of data preprocessing stage of the analysis. An algorithm developed specifically for the task of detection of bands performed the feature extraction. The second application of feature extraction creates a new representation of the data that helps to filter useful information and attenuate the nuisance variation in the data (random noise). Such a method can be the wavelet transformation, which proved very powerful in representing time series data for an analysis of time series data on various scales [17].

We also observe the collinearity [18] in the average brightness curves. The resolution of the digital camera is very high. The adjacent pixels in the electrophoreogram image strongly correlate. The value of the pixels is very similar and providing we have several neighbouring pixels we can estimate the value of the pixel with high precision. Even though we collapse one dimension of the electrophoreogram image, the very detailed sampling (due to the high resolution of the photography) still affects the average brightness curves. The feature extraction methods such as the wavelet transform can provide a data representation that is less affected by the colinearity. To summarise the example of the analysis of electrophoreogram, the processing of biochemical data relies on procedures that are in general common for any data [11]. However, the data suffer from several issues that have to be resolved by specifically developed methods.

## 2.3 The biochemical data

This chapter aims to present the specifics of the processing of biochemical data, to identify the issues common to all the data types that we dealt with as well as the specific properties concerning specific to data from certain biochemical analytical methods. This chapter is intended to describe the following data shortly:

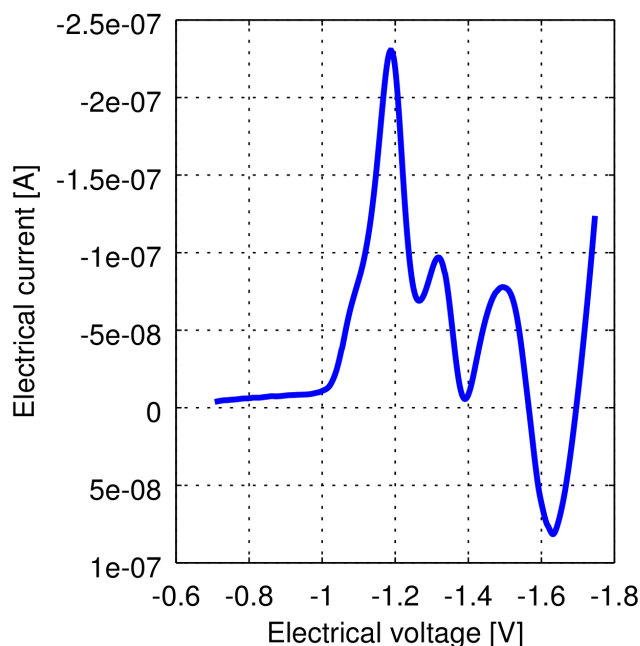
- Brdička curve
- Electrophoreogram image
- Nuclear Magnetic Resonance spectroscopy
- Elemental mapping by LA-ICP-MS
- Additional information supplemented to the biochemical data

### 2.3.1 Brdička curve

The Brdička curve is a result of Brdička reaction [19]. The Brdička curve is measured by differential pulse voltammetry [20]. The Brdička curve can be used to measure the electrochemical properties of various (biological) materials [C4, J7]. The information carried by the Brdička curve and its features can predict several medical conditions or composition of various organic solutions.

The Brdička curve is a vector of value pairs of electrical current  $i$  and voltage  $u$ ,  $c_j = [u_j, i_j]$ ,  $u_j, i_j \in \mathbb{R}$ , for  $j = 1, 2, \dots, n - 1, n$ , where  $n$  is the length of the vector. In addition the curve is usually associated with a set of variables  $y = [y_1, y_2, \dots, y_n]$  (treatment group, severity of disease, concentration of a chemical compound, etc.).

The Brdička curve suffers from the usual problem encountered in time series data, and it is the collinearity among the values of the series (see Section 3.3). The very fine steps in electric voltage, which is varied during the measurement, produce very finely

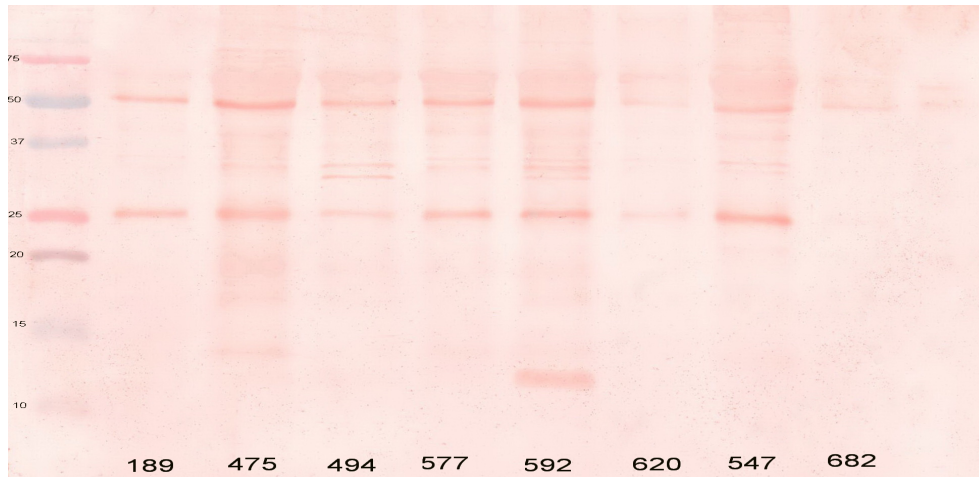


**Figure 2.2.** Example of Brdička curve of liver tissue [21]

sampled Brdička curve, but these adjacent data points are highly correlated and do not carry much of useful information. The relationship governing the shape of the Brdička curve can be non-trivial [22] and the processing of Brdička curve is covered in the following publications [C4, J7, C3, C2] of the author. A visualisation of a Brdička curve originating from rat's liver is presented in Fig.2.2.

### ■ 2.3.2 Electrophoreogram image

The short description of the electrophoreograms was provided as an example in Section 2.2. To continue the running example of the processing of electrophoreograms we will present the problem in a more general way. A digital image  $I : \Omega \rightarrow \mathbb{R}^m$  with  $m$  colour channels defined in a rectangular  $\Omega \subseteq \mathbb{Z}^d$ , where  $d$  is the grid dimension. The digital image represents the electrophoreogram containing the results of an analysis of several actual samples of (biological) material. A band represents each sample, and a simple algorithm (adaptive thresholding) can extract the band and complement it with features estimating the specific weights of the analysed samples of (biological) material. Formally, we transform the image  $I \rightarrow c$  a curve from a set  $c \in \mathbf{C}$ . Each curve is represented as a pair of molecular weight  $m$  and image brightness  $b$ , so  $c_j = [m_j, b_j]$ ,  $m_j, b_j \in \mathbb{R}$ , for  $j = 1, 2, \dots, n - 1, n$ , where  $n$  is the length of the vector. As a result, an electrophoretic curve represents each band. A coordinate system estimated from the standard band provides the reference for indexing or interpolation. Following this procedure, the result is a set of curves aligned according to reference information from the standard band. Again, additional information (treatment group, disease severity, etc.) supplements the experimental data. The curves suffer from collinearity (as was described in more detail in the Section 2.2). A whole electrophoreogram is shown in Fig.2.3 and an illustration of extraction of brightness curve from one band in an electrophoreogram is in Fig.3.1.

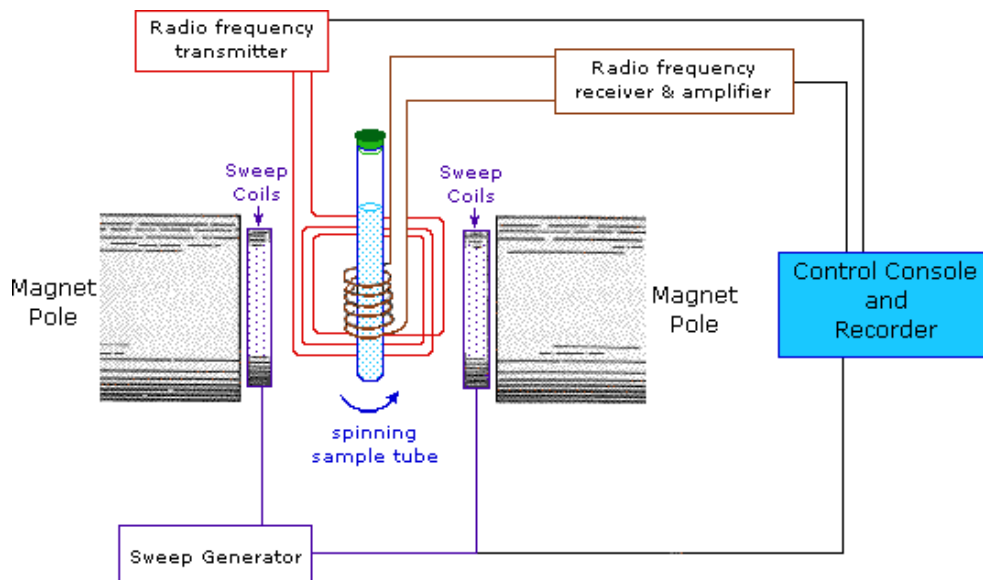


**Figure 2.3.** Example of electrophoreogram image of MT proteins [J6]

### 2.3.3 Nuclear Magnetic Resonance spectroscopy

The nuclear magnetic resonance (NMR) spectroscopy measures the properties of hydrogen or other active elements in a sample of material which is usually some liquid. The active elements are characterised by non-zero magnetic spin. In the case of metabolomics, the examined samples are biological fluids such as plasma, blood, urine or saliva) [J2].

The measurement is very similar to the nuclear magnetic resonance imaging, see Fig. 2.4.

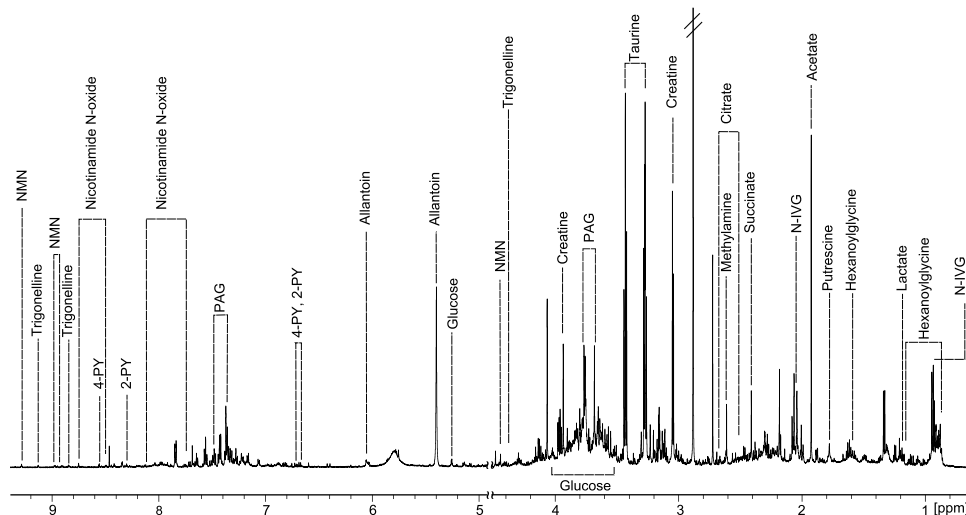


**Figure 2.4.** Schematic depiction of the NMR spectrometer [23]

A radio-frequency pulse stimulates (excites) the analysed sample which is in strong magnetic field, and the recovery of the system to the equilibrium is measured. The recovery describes the free induction decay signal. The interaction of hydrogen nuclei in chemical bond and specific structures modulates the resonance frequency of the nuclei. The modulation results in the form of shifts in the measured frequency. The



frequency shifts help to figure out the molecular structure of analysed compounds or recognise known compounds. The data analysed in NMR spectroscopy are the frequency spectra obtained by Fourier transform of the free induction decay signal. An example of NMR spectrum of mice urine is in Fig. 2.5. [24]



**Figure 2.5.** Example of NMR spectrum of urine with annotated peaks belonging to selected metabolites. [J2]

The spectra of the examined samples are vectors of frequencies and respective magnitudes. The spectra often do not need any specific preprocessing because the shifts in the frequencies related to the compound structure are very distinct [24]. However, a minor and non-linear shift in the frequencies may be present. The magnitude of these minor shifts is smaller in order of magnitude than the major shifts due to compound structure. The processing usually includes steps that eliminate minor shifts. For example, the aggregation of the spectra by averaging magnitudes over frequency intervals with specified width greater than the minor shifts eliminates the minor shifts. This procedure called binning [12]. Another problem of NMR spectroscopy data is the often unknown dilution of the biological samples, which affects the magnitudes of the peaks in the spectra and the area under the curve of the spectra. The unknown dilution of the sample results in the inability to establish a relationship between the magnitude of in NMR spectra and the actual concentration of the observed chemical compounds. The dilution varies among the samples and cannot be estimated. That is why specific processing procedure has to be employed to alleviate this problem. These techniques are denoted as the normalisation of NMR spectra [12]. The simplest normalisation method is the constant sum normalisation that normalises the sum of the spectrum to a constant value. The constant is the same for each spectrum in the analysed set. The sum of the spectrum approximates the content of diluted compounds. The normalised spectra are comparable among each other, however we lose the information about the actual content of the diluted compounds.

After the processing steps set of vectors of pairs  $s = [f_j, m_j]$ ,  $f_j, m_j \in \mathbb{R}$ , for  $j = 1, 2, \dots, n - 1, n$ , where  $n$  is the length of the vector, represents the NMR spectroscopy data. The representation of the data as vectors tends to suffer from collinearity. In the case of NMR spectroscopy data, the relationships among the samples of the spectra are

not just the relationships among neighbouring samples. Due to the interactions between compounds in examined samples, the relationships may exist among any pair of values of the vector, as can be seen in Fig. 2.5. In Fig 2.5 for example, four annotated peak represent the metabolite Nicotinamide N-oxide. Each of the four peaks is proportional to the content of the Nicotinamide N-oxide in the analysed biological sample. These relationships make the analysis of NMR spectroscopy data difficult.

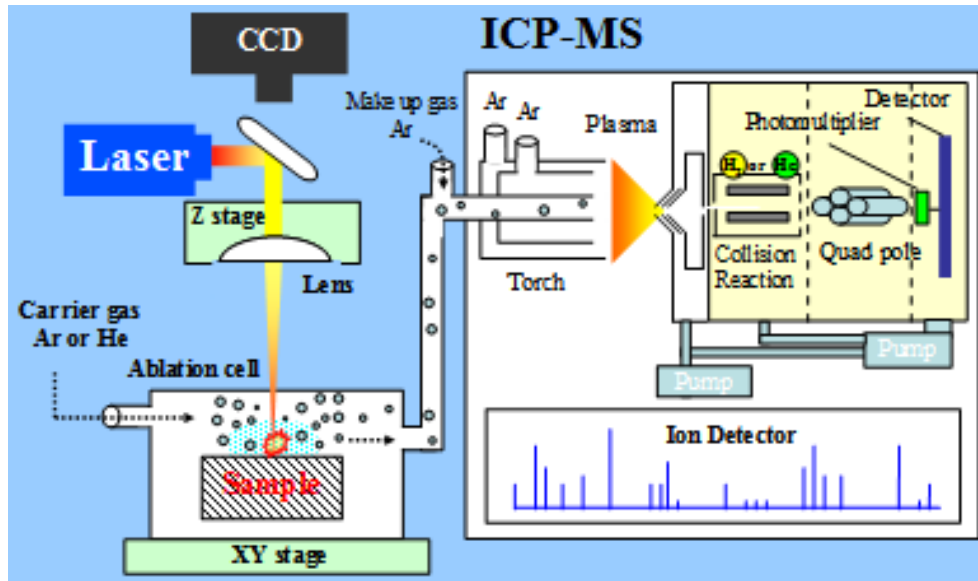
The NMR spectroscopy is used mostly in metabolomics [25]. Metabolomics is a study of metabolism. The metabolomic experiments sometimes suffer from small sample sizes (especially in the case of studies performed on animal models, which are expensive). Small sample sizes may result in data having more variables (frequency bins) than observations (measured samples taken from objects of study) [26]. Small sample sizes and a large number of variables is another challenging property of NMR spectroscopy data calling for specific analytic tools. An example of an NMR spectrum with annotated peaks belonging to important metabolites is in Fig.2.5.

### ■ 2.3.4 Elemental mapping by LA-ICP-MS

The elemental mapping carried by laser ablation induction coupled plasma mass spectrometry (LA-ICP-MS) [27] is a sophisticated chemical analysis, which combines several laboratory techniques. The analysis is performed by continuous evaporation of the analysed sample by a high energy laser beam. The resulting gas is transported, mixed with inert medium and subjected to analysis by a mass spectrometer to identify the presence of specific chemical elements. Schematic depiction is in Fig 2.6. The laser ablation moves across the analysed (biological) material row by row in an orthogonal grid and destroys and evaporates the matter of the sample. The laser beam has a finite cross-section, which limits the width of the rows in the analysed sample. The speed of the laser beam movement limits the resolution long the row. The practical limit is the duration of the scanning, which is inversely related to the speed of the laser beam movement [28]. Due to the nature of the scanning procedure, the resulting data are inherently isotropic. The transportation of the evaporated sample to the detector of the mass spectrometer causes the gas to mix. The gas mixing contributes to a decrease in the spatial focus and results in an observed blur in visualisations. Another source of uncertainty in the resulting signals is due to the detector imperfections. The technical details of the detector are out of the scope of this work, but for example, a simultaneous appearance of two silicon  $^{28}\text{Si}$  atoms at the detector can falsely indicate an iron  $^{56}\text{Fe}$  atom signal.

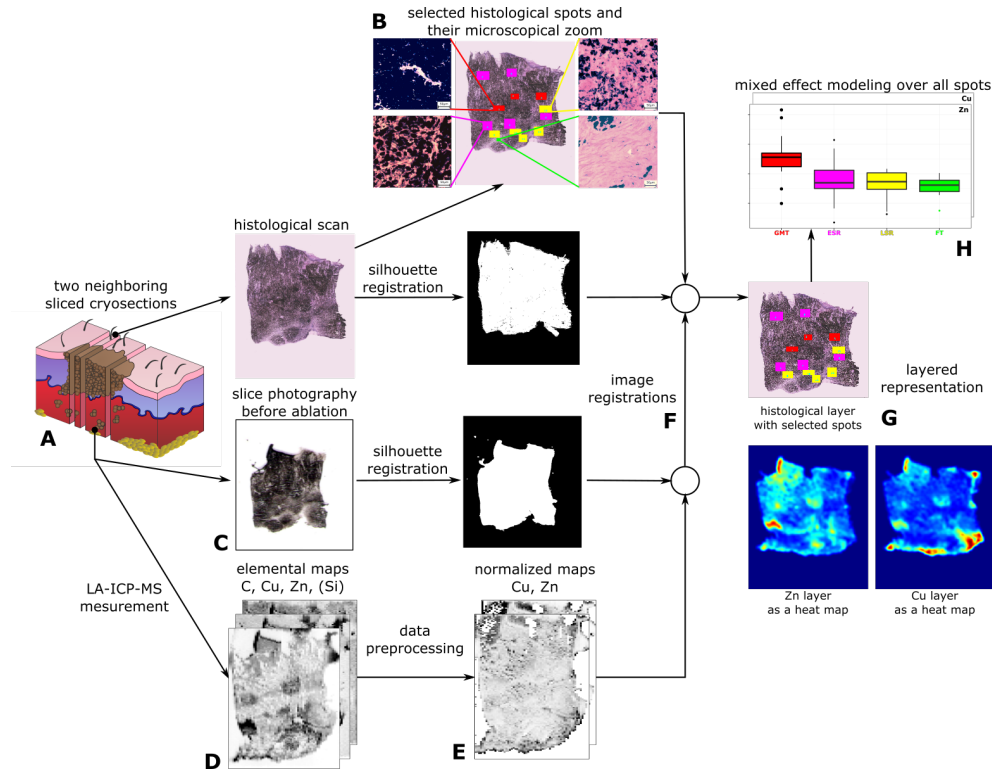
An important role in the measurement of chemical elements' content in biological sample analysis plays the inhomogeneity of the matter in the sample. The sample is thicker in some places and thinner in other places. The intensity of the laser beam is constant during the measurement. In case of relatively low intensity, the laser beam may not evaporate all the matter in the sample. This so-called soft ablation results in low intensity of detected signals. The hard ablation uses laser beam intensity high enough to evaporate all the matter of the sample and even a substantial amount of the supporting material, which is often glass. The observation of strong silicon  $^{28}\text{Si}$  signal indicates the hard ablation. The silicon is present in glass in the form of silicon dioxide  $\text{SiO}_2$ . The silicon signal indicates evaporation of glass slide supporting the analysed sample.

The result of the biological sample analysis by LA-ICP-MS is a set of matrices of signals of the analysed chemical elements, which can be represented as a multi-channel



**Figure 2.6.** Schematic depiction of LA-ICP-MS [29]

digital image A digital image  $I : \Omega \rightarrow \mathbb{R}^m$  with  $m$  colour channels corresponding to different elemental maps defined in a rectangular  $\Omega \subseteq \mathbb{Z}^d$ , where  $d$  is the grid dimension. These may be for example elements carbon  $C$ , silicon  $Si$  or metals copper  $Cu$ , zinc  $Zn$  and iron  $Fe$ . A sample with a known constitution has to be measured to establish a relationship between signal magnitude and the element content. The relationship may be in the form of a calibration table or a mathematical formula. This relationship serves for estimation of element content in the analysed sample. The actual thickness of the (biological) material may vary significantly, and it is not possible to measure. In that case, the calibration relationship does not hold. The information about the measurement parameters (the physical size of the sample, laser beam intensity, speed of the scanning movement of the laser) as well as additional information of the biological properties of the sample is part of the resulting data. From the description of the measurement procedure, it should be clear that the data are highly collinear, and the relationship among the measurements differ in each direction. The evaporated sample does not travel through the device in separated quanta, but as a steady flow in which the compounds mix, however, the amount of mixing can be very small. The data properties call for the analysis of the spatial distribution of the elements according to other spatial properties of the samples. An example may be the attempt to find a relationship among clusters of cancerous cells in a tissue sample and the content of chemical elements [C1], i.e. the biologically active metals. To relate the clusters, those have to be first identified in the sample beforehand. The identification of cancerous clusters requires staining the sample. The dye used for this procedure contains metals that would affect LA-ICP-MS measurement. A parallel tissue sample has to be used to identify cancerous tissue [J1]. The tissue sample is scanned by a microscope and represented as an digital image A digital image  $I : \Omega \rightarrow \mathbb{R}^m$  with  $m$  colour channels defined in a rectangular  $\Omega \subseteq \mathbb{Z}^d$ , where  $d$  is the grid dimension. To relate the sample examined for the presence of cancerous cells and the LA-ICP-MS measurements have to be related. For this purpose a method known as image registration can be used [30][J1]. The image registration tries to find a relationship between images in the form of linear mapping by minimising a criterion of image similarity. The image registration allows for indexing among the samples and reasoning about the element spatial distribution and

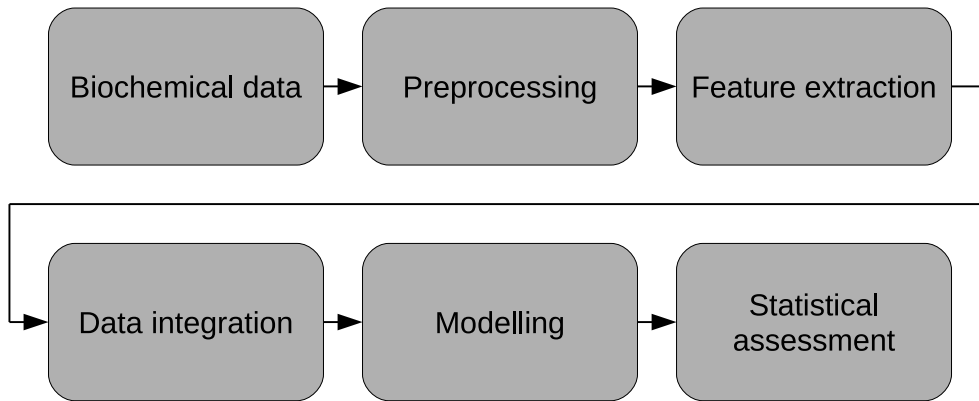


**Figure 2.7.** Illustration of the whole processing pipeline. The subject of the analysis is the biological sample of melanoma (A). An expert annotates one tissue section (B). From a parallel tissue section, The LA-ICP-MS produces an image of the tissue section (C) and element maps (D). The carbon element map corrects for the uneven thickness of the tissue sample (E). Image registration enables indexing among the histology and the element maps (F). Layered data representation (G) allows statistical analysis (H). [J1]

presence of cancerous cells. Fig.2.7 illustrates the whole pipeline of elemental mapping data processing.

### 2.3.5 Additional information supplemented to the biochemical data

The description of all of the measurements mentioned above included a reference to additional information associated with the measurements. This additional information can be in any form. For example, it can be information about the membership in a treatment group (the typical patients vs controls design). Therefore it can be called the general data. We use the term general data only to differentiate them from the biochemical data, which we have to treat differently. The general data are the type of data that is usual in the standard statistics. The analysis of general data usually does not involve any inventive work (in the sense of the development of new data analysis methods). The analysis follows the standard procedure of data exploration, understanding, and cleaning in the data preprocessing step and is completed by an appropriate statistical analysis, for example, a statistical test [9]. The methods statistics provide for various data types are out of the scope of this work. The only exception is the generalised linear mixed effect models (GLMM). The GLMMs are of immense importance in the statistical analysis of data originating in biology, chemistry, and biochemistry calls for a brief description. The GLMMs will be discussed in the section 3.5.4.



**Figure 2.8.** General pipeline of the biochemical data processing

## 2.4 Problem formulation

Let us assume, we perform a series of experiments that produces a collection of data necessary for verification of a hypothesis. We process and integrate the measured (input) data. We combine the experimental data with external information about the experiments to find a relationship modelling our hypothesis tested by the experiment. The general data processing pipeline is in Fig. 2.8.

The data we deal with are generally vectors or matrices  $D : \Omega \rightarrow \mathbb{R}^m$  with elements represented as vectors of length  $m$  defined in a rectangular  $\Omega \subseteq \mathbb{Z}^d$ , where  $d$  is the grid dimension. The elements in the vector of length  $m$  constituting the data  $D$  relates to all the measurements resulting from the experiment, for example the elemental maps of different chemical elements (Cu, Zn, Si and C).

The data preprocessing and feature extraction are generally transforms  $f$  of the data  $D' = f(D)$ . Generally,  $D' : \Omega' \rightarrow \mathbb{R}^l$  is a matrix whose elements are vectors of length  $l$  in a rectangular grid  $\Omega' \subseteq \mathbb{Z}^e$ , where  $e$  is the grid dimension.

These can be various transforms of the elements of the data  $D$ , the transforms of the rectangular grid  $\Omega$ , or both of them. An example of a transform of the elements of  $D$  is a conversion of an image from RGB to greyscale colour space, the original vectors of length three are simplified into one number. The transformation of the grid  $\Omega$  is for example image rescaling, where the dimensions of the image change. When the data  $D$  are digital images, common image transformation can represent the preprocessing. In the case of vectors, the methods are generally signal processing methods. The preprocessing of the elemental maps represents the step G in Fig. 2.7. The image registration performs image transformation of the grid, which is depicted in the step F of the pipeline in Fig. 2.7.

The results of preprocessing and feature extraction serve for the selection of the objects of interest from the point of view of the original hypothesis. The data integration step describes a strict transform of the input data  $D$  into the 2 dimensional data matrix  $X$  with  $n$  rows and  $p$  columns. The rows in the data set  $X$  represent the objects of interest in the measured data  $D$ , for example the spots annotated in the histological images as presented in the Fig. 2.7B. The columns of the matrix  $X$  represent the properties of the objects of interest, for example the information about the spot's tissue type also depicted in Fig. 2.7B. The resulting data set  $X$  is in a common data format accessible by any machine learning or statistical method.

We define the matrix  $Y$  of  $n$  rows and  $q, q < p$  columns as a subset of the matrix  $X$ ,  $Y \subset X$  as the matrix of responses. These responses are derived from the input data

$D$  and serve for the modelling purposes. An example of the response variable can be the labelling of the objects of interest in the data; labelling of the spots according to their tissue type, see 2.7B. The matrix  $Y$  contains the external and internal information about the data that describes our modelled hypothesis. The external information can be for example the age of the animals weeks, while the internal information is the tissue type labelling. We model the data in the matrix  $Y$  with a model represented by a function  $h$  that generally provides an estimate of  $Y$ ,  $\hat{Y} = h(X)$ , in case of regression models the estimate is real number  $h : \mathbb{R} \rightarrow \mathbb{R}$ , in the case of classification  $h : \mathbb{R} \rightarrow l$ , where  $l$  is a label from a set of labels  $\mathbf{L}$ .

The final part of the biochemical data analysis is the statistical assessment that estimates the variation in the modelled relationships represented by the model function  $h$ . The standard methods use the classical statistical test which compare the test statistics to a test statistic distribution. Simulation-based methods generate the test statistic distributions by modelling the randomness of the data generation process. The applied method of non-parametric bootstrap is described in Section 3.5.5 and also in Section 4.3.2.

## 2.5 Summary

In this chapter, we introduced the data dealt with and presented some of their properties. The first significant problem with data originating in biochemistry and related fields is the need for reliable extraction of useful information from the original measurements, which may be very diverse. We describe the methods of feature extraction in the Section 3.1. The second problem is the combination and merging of data from sources, which differ in the measured signals, but are necessary for the completion of the full description of the problem and proper understating of the complexity of the data. Methods of data integration are discussed in Section 3.2. The third common problem of all of the presented data is the inherent relationships in the measured signals. These relationships call for methods that are capable of discriminating between the variance in the data correlated with the response, underlying concepts hidden in the signals, and the orthogonal noise, possibly inflated by collinearity. We discuss the collinearity in Section 3.3. The collinearity poses problems for the data modelling when we take the additional information about the data into account. The modelling methods suitable for the biochemical data are reviewed in the Section 3.4.

## Chapter 3

### Specific processing of biochemical data

In this chapter, the three main problems connected with the biochemical data, which were introduced in the Chapter 2 – the feature extraction (Section 3.1), the combination of data from different sources (Section 3.2) and the inherent dependency problem (Section 3.3) – we will examine the problems in more detail and we will present several possible approaches and solutions for the problems. The data analysis approaches are reviewed in the Section 3.4.

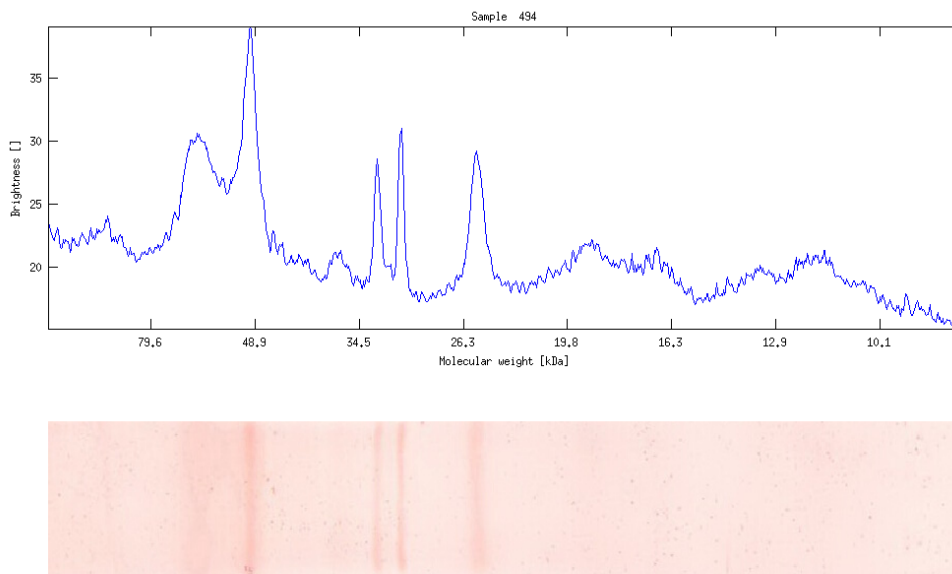
#### 3.1 Feature extraction

The fields of data mining and machine learning deal with the problem of feature extraction. One of the many classifications of feature extraction methods is:

- a) model-based
- b) data-adaptive
- c) data-non-adaptive
- d) data-dictated,

This classification was adopted from [31] and is by no means the only possible or the best of all possible classification of feature extraction methods. However, it is reasonably simple and covers most of the feature extraction methods relevant to our work. Naturally, the choice of method for feature extraction depends on the type of data. The biochemical data share many similar properties, and therefore general feature extraction methods can be used as the starting point for the modelling of the relationships.

Let us review the most important ones for each category. From the category of **model-based feature extraction methods**, the statistical models such as GLMMs, hidden Markov models are the most prominent. To the category of **data-adaptive methods** belongs the principal component analysis (PCA) [32] and in broader sense even methods such as partial squares regression (PLSR) [33] and artificial neural networks (ANN) [34], which perform the feature extraction in order to estimate an outcome variable. The category of **data-non-adaptive feature extraction methods** consists of variety of methods, it is traditionally the Fourier transform and related methods of frequency spectrum estimation [35–38], the bountiful wavelet transform [17] and its variations as time-frequency representation of the data and at last various data aggregation methods in time domain. The **data-dictated feature extraction** is in accord with the name particular to the data. Therefore, it is difficult to name any general methods belonging to this category. On the other hand, the data from the field of biochemistry offer excellent examples of data with clear interpretation. The electrophoresis for example provides a signal with peaks, that can be attributed with chemical compounds of specific molecular weight [J6, J5, J4]. The amount of a chemical compound in the studied sample could be estimated from the area under curve of the peak. Such representation of signals is comprehensible for chemists and biochemists. However, it is not suitable for use on



**Figure 3.1.** Electrophoreogram, one of the bands from Fig.2.3 (down) and an extracted curve of the separated compounds (up)[J6]

larger sets of signals with subsequent application of statistical, data mining or machine learning methods. The reason behind this claim is poor reproducibility of the measurement conditions across several runs of experiments. The variation in experiment condition introduces uncertainties in the data. Feature extraction methods attempting to mimic the manual assessment of the data by an expert are problematic. An expert can easily decide many ambiguous cases. Such behaviour (a case-specific treatment of peak assignment by an expert) is extremely difficult to model reliably. Developed methods usually cannot perform with the accuracy comparable to an expert. In the scope of the data that have been presented, the relevant approaches are:

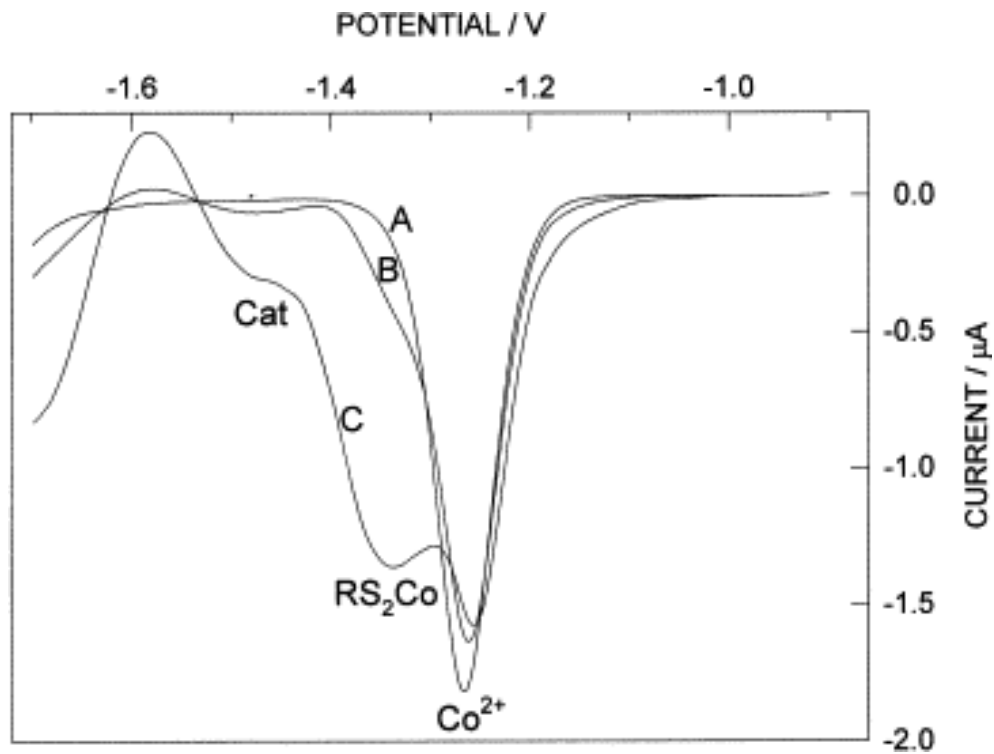
- a) process the data in a way specific to the measurement method and established in the community for the data interpretation – the data-dictated approach to feature extraction,
- b) design new features that are not necessarily relevant to the exact interpretation of the of the data – the data-non-adaptive approach and
- c) to utilise statistical or other methods which are capable of dealing with the specific features of the data in their original form.

### ■ 3.1.1 Data-dictated feature extraction

The data dictated approach is applicable to Brdička curves [C4, J7], electrophoreograms [J6, J5, J4] and NMR spectroscopy [39]. All of the resulting signals are consisting of several peaks, that relate to a specific part of the mixture that constitutes the studied sample. Therefore, the processing methods are very similar for all the data types. Unfortunately, the problems faced in the endeavour to decompose the signals into meaningful descriptors of the mixture are also the same. In the field of chemistry and biochemistry, the method of signal decomposition is called deconvolution. Deconvolution relies on estimating the parameters of individual peaks. The meaning of the term deconvolution in chemistry and biochemistry differs from the understanding of



the process in signal processing or other fields. Assumptions about the peaks' specific shape (bell curve, gaussian or lorentzian) simplify the signal decomposition. The algorithms for the decomposition include the expectation-maximisation algorithm (in case of approximating the signal as mixture of gaussian bell curves, or other statistical distributions) [40], the heuristic approaches (in case of unspecified peaks shapes) [C3, C2], or the simulation of the actual mechanism taking place during the chemical reactions [41] as seen in Fig. 3.2.



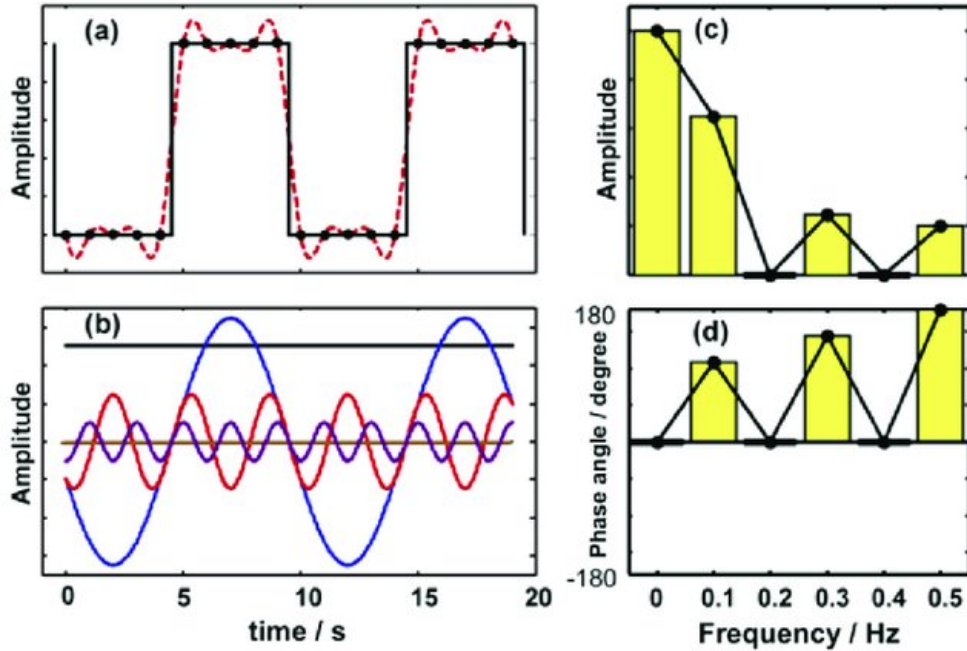
**Figure 3.2.** Brdička curve - a decomposition of the curve into waveforms [41]

The problem is how to estimate the number of peaks present in the signal. The common property of the signals is a considerable peak overlap. The peak overlap in combination with the corruption of the signals by noise and artefacts does not allow for full automation of the process. The literature provides many examples of algorithms for signal deconvolution [22, 42–43]. To our best knowledge, none of them established itself in the field as the standard data analysis procedure. These methods may prove to provide better results than data-non-adaptive feature extraction methods when used by chemists or biochemists. The need for manual checking of the feature extraction process results and making corrections may outweigh the benefits.

### ■ 3.1.2 Data-non-adaptive feature extraction

The inability to automate the process of data dictated feature extraction makes the data-non-adaptive feature extraction methods a viable alternative. The methods mentioned in the introductory paragraph were Fourier transform [35], wavelet transform [17] and time-domain aggregations [31]. The properties of the studied signals and the representation of the signal in the frequency domain provided by Fourier transform are not a good approximation of the data. The position of the peaks and other waveforms in the signals is very important for the interpretation. Unfortunately, the frequency

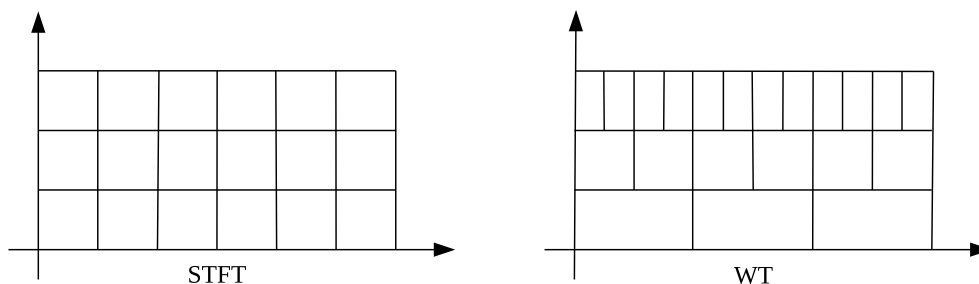
spectra do not preserve the time information. On the other hand, the shape of the frequency spectrum of the signal describes very well the shape of the waveforms. Fig. 3.3 shows a simple illustration of the Fourier transform.



**Figure 3.3.** Example of Fourier transform of square signal. The square signal approximates an infinite sequence of sine waves with odd periods. [44]

The lack of time-domain information in frequency spectra directs us from the pure frequency domain transforms to transforms that combine time domain and frequency domain. This combined representation provides the short-time Fourier transform (STFT) [45]. The STFT retains the time domain information by splitting the processed signal into short segments and performing the discrete Fourier transform on these pieces of the signal. Different approaches to estimation with different properties exist [36–38]. Each resulting spectra of the shorter segments constitute a column (or a row) in a matrix called spectrogram. It is common to visualise spectrogram as a heatmap. The spectrogram provides extensive representation which can be analysed by machine learning methods. The main disadvantage of the STFT is the inherent trade-off between the resolution of the spectrogram in time and frequency domain [46]. There is an inverse relationship between the time and frequency domain resolution. The resolution depends on the length of the segment, the shorter the segment of the original signal, the better the time domain resolution. The frequency-domain resolution also depends on the segment length, however, the longer the segment, the better the frequency-domain resolution. In the case of the spectrogram, the segment length is constant, and thus the resolution in time and frequency are constant in the whole spectrogram. Using overlapping segments improves the time and frequency resolution to a certain extent, however, the effect is limited.

A similar representation of the studied signals provides the wavelet transform [17]. The wavelet transform combines the time and frequency domain information by its definition. In comparison with the STFT, the wavelet information does not estimate the



**Figure 3.4.** The difference in the time-frequency resolution between short-time Fourier transform (STFT) and wavelet transform (WT). The time-frequency resolution is uniform in both the time and frequency for the STFT, while the WT offers better time resolution for higher frequencies (where shorter signal segments suffice for the estimation of the frequency spectrum) and better frequency resolution for lower frequencies (where the estimation of the frequency spectrum requires longer segments of signal). [47]

frequency spectrum, but the result of the transformation is a response of specifically designed filters. The mother wavelet and its scaling function define the filters. The choice of the mother wavelet allows for customisation of the outcome – the scaleogram. The time-domain and frequency-domain resolution in the scaleogram depend on the scale at which is the signal analysed. The wavelet transform allows for higher time-domain resolution for short scales (high frequencies) and higher frequency-domain resolution on long scales (low frequencies). The palette of wavelets of different properties allows choosing wavelets for specific tasks – which can be for example de-noising, change detection or peak detection. The choice of mother wavelet suitable for feature extraction by peak detection in studied signals is relevant for the interpretability of the resulting description [48]. An example of the use of wavelet transform is described in [J6].

To complete the description of the data-non-adaptive methods for feature extraction relevant for biochemical data, the time domain aggregations [31]. These methods split the signal into equidistant segments and compute an aggregated value from the segments. The length of the segments can be varied to obtain the best results, and the segments may be overlapping. The aggregation function may be as simple as a sum of the elements in the segment or more complex such as mean, median, variance, standard deviation, or designed specific purpose. The aggregation in the time domain is very extensively used in NMR spectroscopy-based metabolomics and is called binning [12].

### ■ 3.1.3 Statistical models for feature extraction

The feature extraction methods falling into the category of the statistical models is an umbrella term for an extensive list of methods for dealing with data with various types of statistical properties. One such category are the generalised mixed effect models which we describe in the Chapter 4. There are specific methods for dealing with data originating in chemistry and biochemistry [49]. The PLS analysis is the most used method in the NMR spectroscopy-based metabolomics. The statistical assessment of these methods needs specific approaches from the category of simulation methods [50]. Therefore, we redirect the reader to the Chapter 5 dealing with the statistical analysis of these data.

### ■ 3.1.4 Summary

In conclusion, there exist several approaches to feature extraction widely used in the analysis of data originating from chemistry and biochemistry. The preferred methods

are data-non-specific feature extraction methods capable of providing a very rich representation of the signals from the perspective of keeping both the time-domain and also frequency-domain information. An example of such a method is the wavelet transform. It is not necessary to perform the feature extraction when using statistical analysis methods capable of dealing with the data in their original form.

## 3.2 Combination of data from various sources

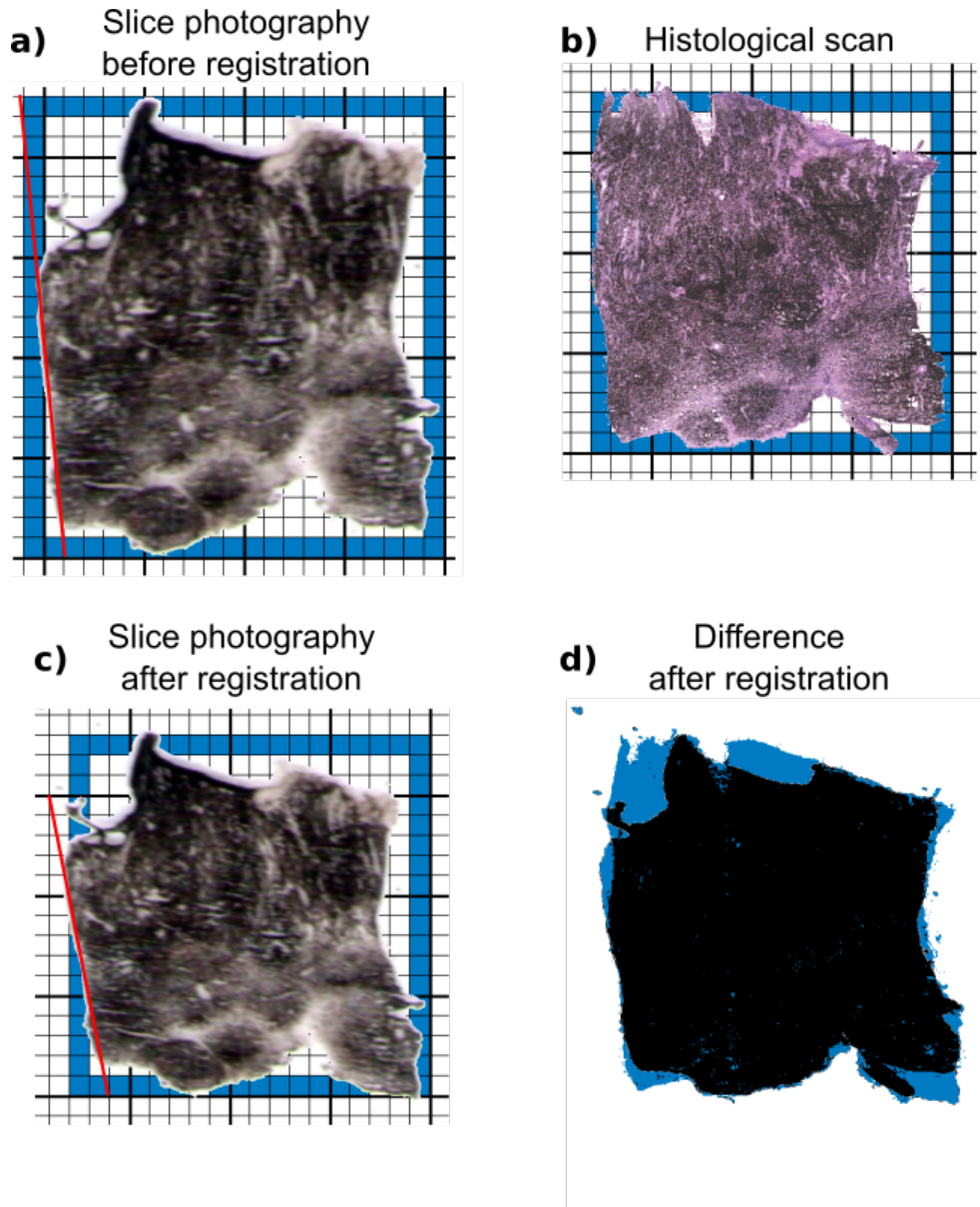
The ultimate goal of advanced analyses is to combine information from diverse sources and combine them all to describe relationships that are new and not obvious to the researchers. The whole field of statistical modelling deals with this endeavour and witnesses constant development of new methods. However, the methods may be advanced; most of them rely on the standard representation of data as data matrices. In order to construct the data matrix, which allows most of the statistical methods to work, the data have to be collected and arranged in a manner enabling exact identification of all observations. Data matrices suit well simple data, where one observation is either one value of one variable or vector of several values of several variables. However, when the observation is a vector, matrix or multidimensional array of values for one variable, the construction of a data matrix is not a simple business of arranging some values in rows and columns. The data matrix is not the best representation method for such data; the relational database outperforms it in all possible measures. However, the statistical methods still require the data to be in the form of the data matrix. Then the easiest approach to construct the data matrix is to unwrap the vectors, matrices and multidimensional arrays into vectors and use each sample as a new variable. This procedure produces the data matrix in the so-called wide form [51]. It is only applicable in the case, where all the vectors, matrices and multidimensional arrays are sampled strictly at the same values of the independent variables constituting the coordinates of the samples. In many experiments, the coordinator of the work bears in mind the analysis of the data and chooses measurement methods, that ensures the absolute consistency among the input vectors, matrices and arrays. In some cases, the feature extraction method produces synchronised features. For example, the NMR spectra are all very well synchronised. The machine processing of the signal utilises fast Fourier transform, which works only for signals of length in powers of 2 [35] and the signal length is directly related to the estimated frequencies. In the case, where neither the coordination of the experiments nor the nature of the multidimensional observations ensure data consistent among observations, several approaches may be utilised to correct these inconsistencies. The procedures for the alignment of the inconsistent data are of two categories. In the first type of problem, the multidimensional observation sampled in a known coordinate system. In the second category, we do not know the coordinate system for multidimensional observation. In the first case, the solution to the problem is straightforward. Simple transformations and interpolations may correct the signals. In the second case, we have to devise the coordinate system and devise the transformations. We have to assume at least partial correspondence among the observations to solve this problem. This correspondence may be in the form of key features, similar shapes or other features of the images [30]. According to the used metrics of the correspondence, a procedure minimising a discrepancy in the correspondence between the images can be used to infer the relationship among the observations and in conclusion perform the correction procedure. [52–53]

The wide data matrix is useful, but there exist what is called a narrow form [51], that

stores data in a more detailed fashion. Instead of having a special variable for each value of multidimensional observations, the narrow form allows for storing the data without the need for any transformations to consistent form. In comparison to the wide form where all measurements were in the form of a 1 to 1 relationship, in case of the narrow form, the relationships in form 1 to n are possible. The obvious advantage of the narrow format is the possibility to retain the original structure and granularity of the data. Additional identification of the data in the 1 to n relationship prevents storing of identical rows, which may cause problems during further analyses. The identifier does not have to be in the form of a special unique key. The indices of the sample in a multidimensional array or any similar value may serve very well, under the constraint, that the value is unique in the context of the object it is describing. As an example may serve a time-stamp for time series data or coordinates of pixels in case of digital images. The narrow data format is not the most appropriate format for all types of analyses. However, the retained structure may be beneficial for data analysis. In such cases, where the observations on different objects result in substantially differing outcomes. In general, the data may differ length or dimension, may be sampled at different, and many more issues may complicate the integration of the data. For this type of data, the application of standard methods of statistical analysis may not be appropriate or even possible (in case of data suffering from a lot of missing values).

The data matrix in the wide or the narrow format, describe only cases where we treat the objects as homogeneous entities. Unfortunately, the reality is often very different. With the development of new imaging methods coupled with chemical analyses, we have the opportunity to process the samples (predominantly biological) and capture their non-uniform properties. The perfect example of inhomogeneous data are the element maps from Section 2.3.4. The inhomogeneity is due to the growing or shrinking nature of the tumour. A single tissue section of the size of approximately 10 x 10 mm may contain several different stages of the tumour tissue. In the case of melanoma [J1], the tissue sections contain growing melanoma tissue, early and late stages of spontaneous regression, fibrous tissue and adipose tissue. The spontaneous regression is a process of organised rebuilding of the melanoma tissue into fibrous tissue. Apart from these tissue types, the tissue sections contain other objects such as blood vessels or bristles. In the case of the tissue sections, the annotation of the tissue histology has to be performed by a specialist. LA-ICP-MS can measure the spatial distribution of metals in the tumour tissue with very fine resolution. It is not possible to perform the histological annotation and LA-ICP-MS measurement on the same tissue section (for more detail refer to Section 2.3.4). Alongside the differences in the examined samples, other issues are hindering the combination of the histological description and spatial distribution of metal content. The resolution of the methods differs not only in the data granularity but also in the measurements uniformity. The histological data have the same resolution in all directions). The spatial distribution of metal content shows substantial differences in the resolution of the data in different directions – the data are anisotropic. The problem of analysing two adjacent samples also may result in data (images) that are not aligned. The data may originate from images that were shifted, rotated, flipped or even deformed – stretched or squeezed, see Fig. 3.5.

In order to perform objective analyses of the data employing statistical analysis, the data have to be first aligned to make the indexing among the different measurement possible. A person can observe similarities in the shapes in element maps and histological annotations. These similarities may point out to the potential relationships among the tissue types (e.g. growing melanoma tissue type and areas with high metal con-



**Figure 3.5.** Illustration of different deformations in images of biological samples of different thickness. The panels a) and b) show the images before registration, the images have different aspect ratios as well as shear deformations. The panel c) shows the image from panel a) after the image registration procedure, and panel d) illustrates the differences in the images after image registration. [J1]

tent). However, without any procedure for indexing among the images, the relationship cannot be quantified and statistically assessed. Therefore the indexing among the data from histology and LA-ICP-MS is of utmost importance. We need a transformation that provides a mapping from one data type to the other to index between the images. Without the unnecessary listing of possible transformation classes applicable for the problem of finding a mapping between two sets, we direct the reader to the work of [30] describing the broad field of image registration methods. The image registration provides a solution to the exact problem of finding a mapping between two images and generally between two sets of points. The affine transformation seems to be appropriate for the problem of two tissue sections. It allows correcting for different scales

(differing resolutions), a shift in position and shear deformations. The transformation parameters are estimated by finding values that minimise a criterion of the agreement between the sets. One of the criteria for the agreement between the images may be a sum of squared differences (SSD, least squares approach). In case of images measured in the same modality (registration between two digital images), this criterion may be sufficient. In case of differing modalities (example from medical imaging grayscale image from computational tomography CT scanner and measured activity of radioactive element from positron emission tomography PET) the simple SSD approach cannot provide reliable information for the estimation of the registration parameters. In general, optimising criterion based on the entropy or mutual information [30] provides a solution to the problem of registration of data sets from differing modalities. The problem of registration of histology and spatial distribution of metal content even a very rough approximation of the data sets by binary masks (image silhouettes) with the SSD criterion may provide surprisingly good results. The best way to deal with the problem of the combination of data from different sources is to plan the experiments so that the need for complicated data integration never arises. In such a case, we avoid the problem at the beginning, and we can bravely proceed with the statistical analysis. Careful planning can only help to avoid the need to combine data from different sources in simple experiments. The more complex the experiment design is, the more difficult it is to align all the data sources and eliminate the need for data preprocessing. From a certain point, the combination of data from various is unavoidable. Several methods exist to align multidimensional observations such as vectors or matrices. The result is a data matrix in wide form. The data matrix in the wide form is suitable for most of the statistical methods as well as machine learning approaches. In cases where the data integration fails, it may be better to use narrow form. The data matrix in the narrow form needs specific methods of data analysis as the standard methods usually do not work well. In case of substantial interactions among the multidimensional measurement, it is better to transform the data to avoid confusion caused by the different measurement methods. The fields of machine vision and image processing provide methods for estimation of relationships among the multidimensional measurements, namely the image registration.

### 3.3 Covariance, correlation and collinearity

This chapter discusses the problem of collinearity [18]. We will present the problem of the correlated data and propose several approaches on how to deal with it. All the presented terms refer to very similar phenomena that are often interchanged or have a different interpretation in different fields. As an example of the different meaning of a term in different contexts may serve the correlation. In the statistics, the correlation is a measure of a 'strength' of a relationship between two variables. If the relationship is linear, the correlation measured Pearson's correlation coefficient [54] is sufficient. Non-parametric estimates of correlation coefficients based on ranks such as Spearman's or Kendall's correlation coefficient [55–56] provide good estimates for certain classes of non-linear relationships between variables. Mutual information [57] and other approaches from information theory measure strength of general relationships between variables. In the field of signal processing, the term correlation has a quite different meaning. In signal processing, the correlation is a signal describing the similarity between two signals, correctly referred to as the cross-correlation of the two signals [58]. One signal is often substantially longer than the other. The shorter signal is called

the pattern. The computation of the correlation in the signal processing context lies in shifting the pattern signal over the longer and compute the similarity at each value of the relative shift (delay). The issue with nomenclature only arises when the two fields meet. With the use of proper names, even the statisticians and signal engineers can communicate without confusion. In the end, it is often complicated to differentiate among the meanings for one simple reason. The statistical analyses often deal with time series data which are signals – the object of study of signal processing. Therefore the meaning of a term should always be explicitly stated in the given context and use.

### 3.3.1 Variability

In this work, we will try to stay in the field of statistics. We cannot rely solely on the statistical definitions, because the studied data are multidimensional (signals, images) and we often have to cross the boundaries. To understand the concepts, the best starting point is probably the **variability**. The variability is a measure of scale or spread of values of a random variable. The common way to estimate the variability is the variance.

$$\text{Var}(x) = (E[x - E[x]])^2 \quad (3.1)$$

The variance  $\text{Var}(x)$  of the random variable  $x$  is the expected value (indicated by  $E[\ ]$ ) of differences of the values of  $x$  from the expected value of  $x$  indicated by  $E[x]$ . There are many estimators of the variability of different properties [59].

### 3.3.2 Covariance

The covariance measures joint variability. For example, if high values in one of the variables coincide with high values in the other, it can mean that the variables covary. Easy way how to estimate the covariance of two variables is the estimation of the expected value of the product of the variables pair.

$$\text{cov}(x, y) = E[(x - E[x])(y - E[y])] \quad (3.2)$$

The covariance  $\text{cov}(x, y)$  of two random variables  $x$  and  $y$  is the expected value (indicated by  $E[\ ]$ ) of the product of the the random variables' differences from their expected values. The high values of the covariance indicate a relationship between the variables. It is not easy to decide which values of covariance indicate strong relationship and which are just caused by high variance in the variables themselves.

### 3.3.3 Correlation

This problem solves the correlation by normalisation of the variables variances.

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \quad (3.3)$$

The Pearson correlation coefficient  $\rho$  of the random variables  $x$  and  $y$  is the covariance of  $\text{cov}(x, y)$  normalised by their variances  $\text{Var}(x)$  and  $\text{Var}(y)$ . Because of the variance normalisation, the correlation is in the range from -1 to 1. Therefore the correlation can be very well understood and used for the description of the relationship between the examined variables. The linear relationship between two variables  $x$  and  $y$  in form  $y = a \cdot x + b$  with coefficient  $a$  and constant  $b$  between results in correlation attaining the



value 1 for positive values of the coefficient  $a > 0$  and the correlation attaining value -1 for negative values of the coefficient  $a < 0$ . If there is not any relationship between the variables, the correlation will be reaching the value 0. Apart from these extreme cases, the absolute value of the correlation coefficients may indicate the strength of the linear relationship. For example, the absolute value of the correlation coefficient greater than 0.5 or 0.8 may indicate a strong relationship. These simple interpretation rules are only approximate and depend on the scientific field. It is usually better to perform a statistical inference and estimate the value of the so-called null hypothesis. The normalisation strips the resulting value of the original scale and units. That makes the correlation usable only for the description of the characteristics of the relationship. For more on the topic, refer to any statistics textbook, which will provide the reader with more details. It is important to notice that the covariance and Pearson correlation coefficient are measures of linear relationships. The methods are not easily extensible to non-linear relationships. The linearisation may help to deal with non-linear relationships. Applying a transform that is inverse to the supposed transform that altered the original variable (e.g. logarithm or square root). The transformation pair is often not clear from the exploration of the data, and some experimenting with the data transformations is necessary. Another approach is to revert to non-parametric methods, which are in case of correlation coefficient the Spearman's or Kendall's rank coefficients. The covariances and correlations are very good for describing the observed relationships. The less obvious thing is the identification of the relationship source.

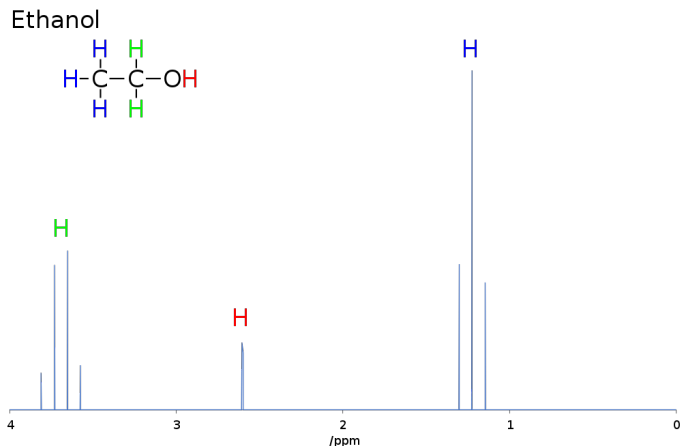
### ■ 3.3.4 Collinearity

One of the difficult problems often faced in statistics is to determine whether the observed relationship indicated by correlation is a direct relationship between the variables. Formally for two variables  $x$  and  $y$  it is true that  $y = f(x)$  and  $x = f^{-1}(y)$ , or the relationship is more complicated, to follow the more formal description, assume variables  $x$ ,  $y$  and  $z$ , then more complicated relationship can be  $y = f(z)$  and also  $x = f(z)$  and even though there is no actual relationship between the  $x$  and  $y$ , the correlation coefficient may indicate otherwise. It is often not possible to decide the relationship type (direct, complicated) without the collection of new data.

The problem of the relationships among data is the point where the collinearity comes to play. By collinearity we understand linear relationships among variables caused by complicated relationships to another variable that is not observed [60, 18]. We are dealing with the multicollinearity when the relationships include more than one variable, that we do not observe directly.

Multicollinearity often affects observations, that have more dimensions (vectors, matrices and arrays). The multicollinearity may be viewed as the most general example of the relationships having an effect on the variables we observe. The relationship may be among any possible group of variables.

One class of data that suffers from collinearity are the time series data. The time series data are sequences of samples collected consecutively one after another from the same studied object (e.g. biological sample). Therefore, we can expect a strong relationship between the adjacent samples. The time series often describe the behaviour of real-world systems. Any disturbance that propagates through any physical system does that in a finite time. The system changes its states continuously. The actual value is a function of the previous values as well as of the external inputs. More accurately, if we adopt the methodology of systematic description of the real-world phenomena,



**Figure 3.6.** Illustration of NMR spectrum of ethanol. Differently coupled  $H$  nuclei produce different spectra, which form the compound (ethanol) spectrum. [61]

we automatically assume the dependence among the samples of time series. The definition of a system of time-related relationships on an object completely describes its properties from the statistical point of view. Unfortunately, in data analysis, we are very far from defining any systems on objects. We deal with (time series) data usually without any model of the system producing the data, but we are interested in the 'content' of the time series data. We usually look for specific patterns in the time series, and the question can be what is the cause of the observed behaviour. A simple question may be which of the patterns are independent on each other and which are linked together. Alternatively, when a trend in the data is among the inherent variation and when the trend indicates a change in the model parameters [62]. In general, we would want to decompose the time series into few parts, most often a random part and a signal of disturbances which caused the observed behaviour. That is the moment when we can utilise the methodology of time series models and the covariance or correlation among the samples. The auto-regressive (AR) model, the moving average (MA) model or their combination, the auto-regressive moving average (ARMA) model treat the signal as a random process (AR) with disturbances (MA). [63] These models do not have to estimate the full covariance among each pair of samples (variables). It is enough to estimate the covariances among adjacent samples. Another motivation for the use of specific methods for dealing with dependency among samples in time series data may be the effort to analyse the trends in time series statistically.

The power of inferential statistical methods is proportional to the sample size – the number of observations, that was collected in order to support a hypothesis. A fundamental assumption which affects almost any classical statistical method is the need of the data to be independent and identically distributed (usually abbreviated as i.i.d.). When we would not know or be willing to respect this assumption, we could (un)intentionally inflate the number of observations and commit a very severe error in our reasoning. Let us take the interpolation of the data as an example. We can very easily change the time resolution of the signal by interpolation. Applying the standard statistical methods on data which were interpolated to double the sampling frequency will provide us with a double of the number of observations. The information hidden in the interpolated data is the same as that in the original signal. By inflating the number of samples, we would very significantly boost our power of statistical reasoning. Therefore we should be cautious about estimating the actual 'number of

independent samples' of the time series. In the case of the time series data, we can rely on the ARMA model to take care of the variability connected with the nature of data. Similar reasoning about the collinearity applies to other types of dependent data. The data with more dimensions cannot benefit from the extensive methodology of ARMA models in the same way as the time series. The approach of using ARMA models for investigation of time-related dependencies has been the standard for a long time. New methods based on simulations [64] are now feasible because of the increase in computing power.

A good example may be an image or any measurement in the form of a matrix with inherent spatial properties [65]. In such data, we may estimate the local covariance. As in the case of time series, it is not probable for samples far apart to be linked together by a relationship, but the adjacent samples are very much alike.

The presented simple examples should provide us with the general idea of relationships among variables. The relationships are either in the form of covariance or correlation. The correlation directly measures the relationship between the variables. The collinearity indicates that there is an unobserved variable causing the relationship. In the context of our data, many of the general comments and recommendations presented in the previous paragraphs are true. Some of the data have very specific properties regarding the relationships among the variables. The Brdička curves and electrophoretic signals are similar to the time series data, and the independent variable is in the case of these data an electrical potential. From the point of the data analysis, it usually does not matter whether the independent variable is time, potential, position, or another variable. In order to analyse these types of data, we may rely on the standard approaches for time series. The usual processing of Brdička curves and electrophoretic signals comprises of variable extraction methods which decompose the signals into specific parts relevant to the chemical content. These parts describe the properties of chemical compounds they represent. This approach eliminates the need for further considerations about the collinearity in the data. In contrast with the simple data, the collinearity plays an important role in the NMR spectroscopy-based metabolomics. The collinearity belongs to the properties of the spectroscopy data and stems from the physical properties of the measurement procedure. In short, the NMR spectrum is a result of the response of several hydrogen  $H$  nuclei to excitation by radio-frequency pulses in very strong magnetic fields. In general, the  $H$  nuclei respond on one specific frequency. In the case of  $H$  nucleus bound in chemical structure, the interactions between the  $H$  nucleus and other nuclei in the chemical structure of the compound modify the resonance frequencies. These modifications are referred to as chemical shifts in the resonance frequency. In chemical compounds, the number of  $H$  nuclei can range in several orders of magnitude (dozens of  $H$  nuclei in anorganic compounds, thousands and more in complex organic macromolecules such as proteins and lipids) and each of the nuclei may have its specific resonance frequency. These shifts in resonance frequency make the spectroscopic data of a chemical compound unique to the compound, see Fig. 3.6. That enables the identification of a chemical compound and its structure from the characteristics of its NMR spectrum. Even in mixtures of several (hundreds) chemical compounds, these are identifiable, because the NMR spectrum is a linear combination of individual compounds spectra. The less optimistic property is that each peak in the NMR spectrum of a compound is proportional to the concentration of the compound. From the data analysis perspective, this property makes the resulting spectra collinear.

Due to the specific nature of NMR spectra, the collinearity does not affect only the

adjacent samples, but the relationships are possible among any group of frequencies (values in the time series data representing the NMR spectra). The source of the collinear behaviour is the measurement procedure. We cannot solve the multicollinearity with a simple application of ARMA models, but we have to resort to another class of data analysis methods, which will be discussed in the next section 3.4.

The fourth example of data from biochemistry was the elemental mapping overlaid over histological images. The properties of this data were briefly discussed several times, to sum up, we have to differentiate among the data sources. The collinearity has very different properties in each of the source of data. The histological image is an image, and thus, we may expect the general behaviour regarding the relationships among the pixels. The adjacent pixels are highly correlated. However, the further apart are the pixels, the less they are influenced by each other. It may seem unnecessary in the histology image, but the collinearity affects the precision of segmentation of the relevant tissue into specific categories. The collinearity emerges when we index and relate corresponding areas in the images. We measure the elemental maps with resolution smaller in order of magnitude in comparison to the resolution of the histological image. The resolution in the elemental maps is not homogeneous. The resolution along the laser path is finer than in the perpendicular direction. The collinearity in the elemental map may be considered to be as of a vector (arranged into a matrix) and analysed accordingly. The resolution in the elemental map is the limiting element for the complete analysis.

We may assume two (or three) approaches. First, a specialist may process the histological image, annotate the specific tissue types which will be used for indexing. In this case, the annotated spots may be large enough for the histological image, but small for the elemental map. After indexing the map according to the histological image, the resulting area may be smaller than a pixel in the map.

The second approach is the inverse of the first approach. We annotate the interesting areas in the elemental maps and index the histological image (however such an approach may be frowned upon by the members of the team concerned with the interpretation of the results from the perspective of biology and histology). The resulting areas would be so large that they would cover inhomogeneous parts of tissue and would not be representative for any further analyses.

The third approach prevents the pitfalls of both the previous approaches. We estimate the collinearity in the elemental maps and decide, which pixels are independent. This parameter may be then used to infer the minimal size of an annotated spot in the histological image from the parameters of the registration transform. The specialist may be instructed to find sufficiently homogeneous areas in the histological image [J1]. This approach ensures safe indexing among images of differing resolutions. The last but not the least important issue is the type of data termed in this manuscript as the general data. The general data are the data closest to the standard statistical datasets. With the general data, we do not have to deal with overly correlated data or repeated measurements of one subject. However, the collinearity may still be very strongly present, as was presented in the introductory paragraphs. The collinearity in general data may originate due to various reasons. One typical example is having one variable measured twice and leaving the copy in the data or similarly having one variable in different units in the data. In the context of biomedical data, we may measure the height of a patient and put down the readings in the metric and the imperial system, one variable in meters and the other in feet. The variables are exact copies and with the Pearson correlation coefficient almost 1 (there may be round off

Another example may be a purely mechanistic relationship among variables: Typical measurement concerning the patients for physical examinations is the measurement of patient's weight. Although the relationship between the height and weight of a patient is not trivial, a very basic idea may govern the relationship. Everybody who ever attended any course in physics knows the relationship among the weight, specific weight and volume of an object. The weight  $[kg]$  of an object is equal to the product of its specific weight  $[kg \cdot m^{-3}]$  and its volume  $[m^3]$ . The object dimensions determine its volume. In the case of a human, the height may very well serve for computation of the volume of a person. The height is only one of three main dimensions, but the shape of the human body is proportional to a certain extent, all three dimensions are closely related. Due to the human body proportions, the body volume of an individual is a cube of the person's height corrected by coefficient reflecting the ratio between the dimensions (height, width, depth) and the fact that the body is not a cube. The second step is the relationship between volume and the weight; the human body is mostly water, and the weight is proportional to the volume. The result is that there is a strong non-linear relationship between the height and weight of a person.

**Example 3.1.** Collinearity in physical objects' dimension

errors when converting the units). For another example, see Ex. 3.1

A third common cause of collinearity in general data is the inappropriate collection of data or pure chance, which may lead to spurious correlations among variables without any apparent cause [66]. The problem with the collinearity in general data is that it is often not reasonable to transform or modify the variable, because we want to retain the exact meaning of the variables. The solution is usually in the application of methods tolerant to this properties.

### ■ 3.3.5 Summary

In summary, the section discussing the covariance, correlation and collinearity covered the intuition and provided examples of correlated data. The covariance is a measure of a linear relationship between variables and the correlation as normalised covariance. The collinearity indicates linear relationships among variables unrelated to the modelled relationship. The presented examples conveyed the idea that the linear nature of the relationships is a very strong assumption and the real-world data may be affected by strong relationships independent on the objective of the analysis. The collinearity has a strong effect on the efficiency of the statistical evaluation. We may interpolate the data in the form of vectors, matrices and multidimensional arrays (time series, images, integrated data sets) and artificially inflate the sample size. Specific procedures are needed to correct the statistical evaluation for the collinear measurements. The last part covered the typical structures in collinearity in data. The most common causes of collinearity among data is the time-dependent covariance among samples in time series data. ARMA models may estimate the time-dependent relationships among samples. In the case of images, the collinearity is mostly local covariance among the pixels. The collinearity in biochemical data in the scope of this work significantly differs from the general cases. The NMR spectroscopic data although they are vectors, cannot be processed by ARMA models, because the collinearity is not constrained to adjacent samples, but may affect any pair or group of samples in the spectrum. The concentration maps obtained by LA-ICP-MS are characteristic by a very big difference

in the resolution in different directions, which makes the estimation of local covariance among pixels in the map complicated, especially in case the map is transformed and overlaid onto other images. To sum up, these properties are the reason for which we cannot directly apply the standard statistical procedures to biochemical data. We have to use specific methods.

### 3.4 Data analysis methods for biochemical data

We divided the section dealing with the methods for the biochemical data analysis into two parts. It follows two standard approaches to the analysis of any data. The classification may be arbitrary because both the classes overlap to a certain extent. However, the approaches stem from different fields of science, and it is better to treat them separately. The first approach is the data analysis using machine learning, artificial intelligence and similar methods. The second approach is the 'traditional' analysis by methods of mathematical statistics. In both of the approaches, we will describe the basic ideas, the necessary data preparation steps and the learning procedure, as well as the model interpretation and the assessment of the findings importance.

### 3.5 Machine learning approach for the analysis of data in biochemistry

The field of machine learning witnessed tremendous development in recent years. The development shows ever greater leaps in the process of creating a complex system (let it be autonomous robots, drones or artificial intelligence) capable of solving difficult problems. Problems which we consider extremely difficult or impossible when humans attempt to find the solution. These new systems often surpass human skills not only in computation but also in activities that were for a long time viewed as being only possible with human intelligence. The breakthroughs, such as the victory of the artificial intelligence system AlphaGo over the human champions in the game of Go, were covered by international media very extensively. It was not only another lost battle for human intelligence, but the victories in the games also mark the incredible development in the field. In many areas, these systems are being deployed and provide many benefits. Regardless of these considerations, the data analysis of biomedical data may benefit very much from these new models. Generally, the projects common in biochemistry do not generate the amount of data necessary for the application of the cutting edge research. In our effort, we have to rely on simpler methods.

The specific properties of biochemical data call for special data analysis methods. The method has to be capable of dealing with multicollinearity. It is practical to use methods that can specifically treat repeated measurement. In many cases, it is necessary to use methods that can provide results for ill-conditioned problems (in the sense of the number of observations and variables). We will first consider unsupervised methods. The **principal component analysis** (PCA), the basic method of unsupervised data analysis with far-reaching applications, especially in fields dealing with collinear data [32]. Other methods for unsupervised data analysis are the **clustering** methods (the most appropriate seem the hierarchical clustering methods) [67]. The unsupervised analysis is mandatory in almost any exploratory analysis of biochemical data. The problem of finding a structure in the data is substantial for the consequent analyses; the structure may imply specific methods for the supervised analysis.

The most common problem in supervised data analysis is the **classification** of observation into classes based on the expert's annotation. Methods for the assignment of a new observation into one class of a given set of classes are the decision trees [68] and related methods (bagged trees [69], random forest [70] and AdaBoost [71] with trees), linear classification methods such as ridge [72], LASSO [73] and elastic net [74] regression, the classification artificial neural networks [75] are closely related to the linear classification methods. Lastly, the multivariate methods, namely partial least squares discriminant analysis [33] or supervised principal component analysis [76].

Another large group of methods are **regression** methods. The regression problem is important in calibration analyses, and in understanding the systems of chemical reactions. In the regression, the first choice is the linear regression and related regression techniques [67]. However, we usually classify the regression among statistical methods, and therefore, we will describe it in more detail in the section 3.5.4. In the case of high dimensional data, the partial least squares regression and supervised principal component analysis are good methods for general analyses with no prior knowledge about the possible relationships in the specific case. We will provide a brief description of the listed methods in the following paragraphs.

### ■ 3.5.1 Unsupervised methods

There are two main types of unsupervised data analysis - the clustering and the principal component analysis.

#### 3.5.1.1 Clustering

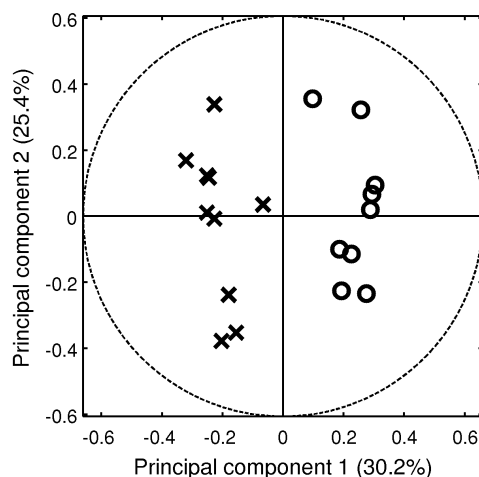
The **clustering** is a method of classification of observation into clusters without any knowledge about any structure in the observations and based only on the observed values of variables. There are several classes of clustering methods. The non-hierarchical clustering methods divide the observations into a given number of clusters based on the similarity among the observations. The typical example of non-hierarchical clustering methods is the **k-means** method [67], in case of the k-means algorithm the similarity measure among the observations is a distance, typically euclidean, but any function meeting the criteria for distance measure may be used. The algorithm assigns an observation to a cluster to which it is most similar – has the minimal distance. Another important example of non-hierarchical clustering method is the model-based clustering – the **Gaussian mixture model**. In this case, the similarity to a cluster determines the mixture probability model. The parameters of the clusters are in both cases estimated by iteratively assigning observations to clusters and subsequent updating of the clusters parameters according to the newly assigned observations. Both these methods have their place in the biochemical data analysis. The **hierarchical clustering** methods provide the structure of similarity among the observations. The hierarchical methods do not need the number of clusters supplied before the analysis. The hierarchical methods perform the bottom-up procedure of aggregating observations into clusters. The procedure starts by finding the closest observation (similarly to k-means in the sense of similarity measured by distance) and creates a link between the pair. The pair now forms a cluster of two observations. In the second step, we look for the smallest distance between a pair of observations. We now treat the pair of observations aggregated in the first step as one entity. The procedure continues until all observations are aggregated [67]. The structure of similarity among observations, which is called dendrogram, is very practical for data exploration because it does not assume any number of clusters in advance. We obtain the clustering in the form of a label by cutting the dendrogram. The issue with the hierarchical clustering is the obvious need

for not only defining the similarity measure for the observations but also the formed clusters. We can compute the similarity between a cluster and an observation in a variety of ways. We can represent a cluster by the average of all observations in the cluster. We then compute the similarity as a simple distance between an observation and the cluster average. Another approach is to compute the distance for each pair of observations and then represent the similarity as aggregated value. Unfortunately, the choice of the aggregating parameters very substantially affects the results. An average or a median result in compact clusters, minimum as an aggregation function tends to results in chains of small clusters. The many choices reduce the reproducibility and validity of the results of hierarchical clustering. Therefore the resulting clustering has to be carefully interpreted and related to the understanding of the data. There are many different approaches to address this problem, interesting measures of the quality of the resulting clusters are the separation of clusters [77] and the cluster stability [78].

### 3.5.1.2 Principal component analysis

The unsupervised **principal component analysis** solves a different problem. The PCA tries to find a set of new variables (principal components) which may represent the original data. The principal components are linear combinations of the original variables and are equivalent to the original representation (lossless representation). The important property of the PCA is that the principal components are orthogonal, uncorrelated, but not necessarily independent in the statistical sense. The principal components combine correlated variables by finding the common component in all the correlated variables. The advantage of the PCA is, therefore, the elimination of covariance among variables – the principal components represent independent components in the original data. More importantly the principal components may be evaluated by the variance which they explain in the original data, and consequently, the original data may be represented a by a set of principal components whose number is lower than the number of variables in the original data. The principal components may be very useful for data summarisation, for visualisation of high dimensional data, for exploration of the relationships among variables and to a certain extent for modelling. The principal components may not be interpretable in the sense of the original data. The intuition behind the PCA analysis is to find principal component as a projection of the data which minimizes the maximum of variation in the data, estimate the contribution of the original variables to the principal component and subtract this from the data; repeat the procedure with remaining data until only uncorrelated noise is left. There exist algorithms which follow this idea, namely the NIPALS algorithm [33]. It is more common to estimate the principal components by an algorithm based on singular value decomposition of the covariance matrix of the data [32]. Differences in variables' variance affect the results. Variables are usually scaled by dividing by standard deviation and also centered by subtracting the mean. The scaling ensures that each variable has equal weight during the estimation of the parameters of the principal components and makes the principal components uncorrelated. The general PCA has many extensions, for example, for nonlinear analysis. The PCA is one of the most utilised exploration methods in data analysis, especially for high dimensional data, which typically the data originating in biochemical analyses, predominantly the NMR-based spectroscopy. For an example, see Fig. 3.7.





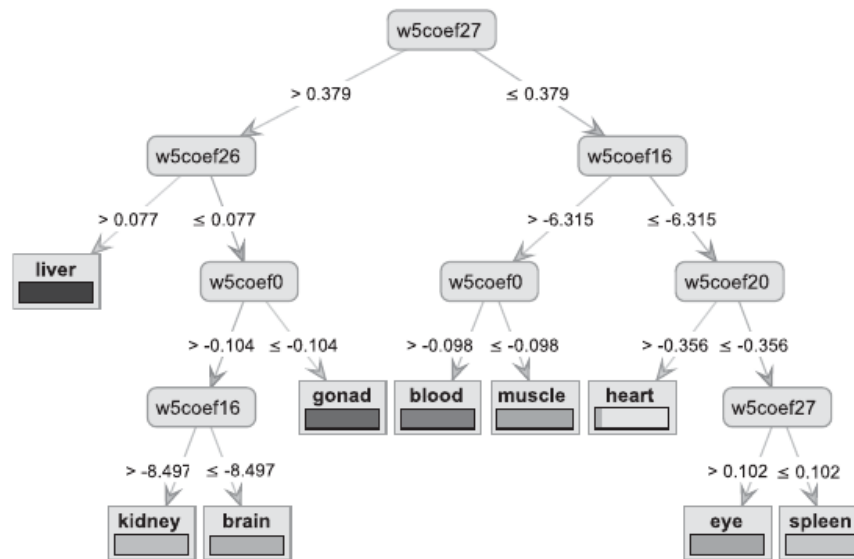
**Figure 3.7.** Example of PCA score plot used in NMR metabolomics for data exploration. Here a difference in metabolic profiles of two groups of laboratory mice is apparent even in the unsupervised analysis of the spectral data. The group denoted by 'X' are mice suffering from a condition similar to diabetes mellitus type 2. The control group indicated by 'O' are healthy mice. [J3]

## 3.5.2 Supervised methods

We divide the supervised data analysis into two main parts – the classification and the regression. The distinction is in the response type. In the case of a continuous variable, the problem is called the regression; if the variable is an indicator of several groups, the problem is called the classification. Classification and regression is a very rough distinction. The properties of the response variable may be more diverse. For the general discussion of the other types of response variables, refer to [79]. The classification problem is more common in the field of biochemistry and thus it will be discussed in more detail.

### 3.5.2.1 Rule-based methods

The **decision trees** are very versatile classifiers applicable to a variety of problems. The main advantage of the classification trees is in the intuitive interpretation of the resulting model. The decision tree represents a set of rules that are understandable to a wide audience. On the other hand, the simple nature of the rules is limiting the number of types of decision boundaries they can model. In general, the decision trees can handle any decision boundary. In order to approximate linear and non-linear boundaries, the decision trees have to utilise piecewise constant segments. The trees have to consist of a substantial number of rules for each curvature in the decision boundary. Such decision trees lose their parsimony. Therefore the best type of data for decision trees is independent variables which imply simple decision boundaries. It may seem that this is in contradiction with the basic properties of most types of biochemical data. Many methods of the data preprocessing provide uncorrelated data suitable for the classification by decision trees. In case there is not any preprocessing procedure providing uncorrelated data, we can use the PCA for creating a new set of uncorrelated variables perfect for the classification with decision trees. An example of



**Figure 3.8.** Decision tree discriminating tissue types in laboratory rats constructed from features devised from the Brdička curves using the discrete wavelet transform. [21]

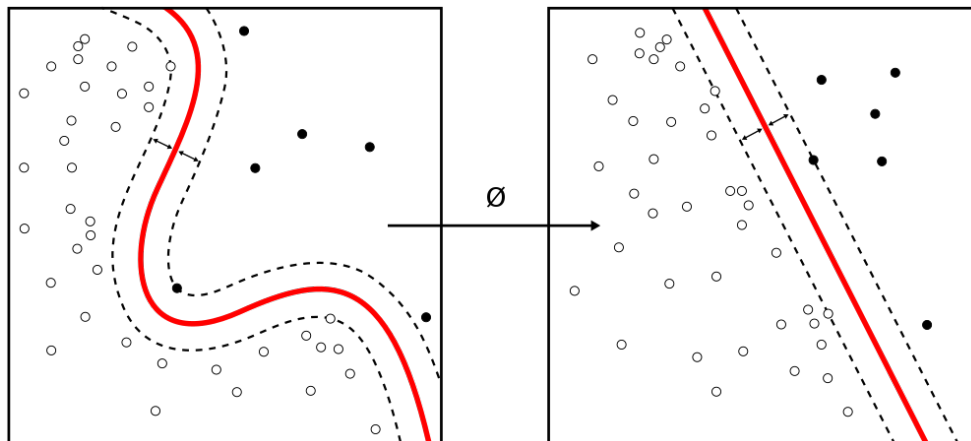
the use of decision tree for discrimination among tissue types is in Fig. 3.8.

### 3.5.2.2 Linear classification methods

The **linear classification methods** [67] work better with correlated data because they can combine predictor variables and model more complex decision boundaries more concisely. They can perfectly fit the constant (simple threshold) and linear boundaries. With the help of **extensions of the original data by polynomial features or spline-based features**, we can model non-linear decision boundaries with linear classification methods. The linear methods may yield more parsimonious models in comparison with decision trees, but the interpretation is not as straightforward as in case of decision trees. People interpreting the model have to have a basic understanding of linear combinations and translation of the classifier coefficients into decision boundaries. Although the interpretation of linear classifier is difficult, it is feasible. Also, linear classifiers are capable of solving a wide variety of classification problems. In case of biochemical data linear classifiers are better at dealing with correlated data than the decision trees, however, at a certain level of collinearity in the data, the linear classifiers would fail to estimate the decision boundary. Some help may provide regularised linear classification methods. These methods can mitigate the collinearity in the data and help to choose the best predictor variables. The class of linear classifiers is very rich in methods. The linear classification methods include the linear discriminant analysis, logistic regression, perceptron and support vector machines. From the given list of linear methods, the logistic regression is interesting for its statistical background [80]. Special attention also needs the **support vector machine method** (SVM) [67]. The SVM optimises a slightly different criterion than the other linear classifiers. The linear classifier typically minimises the classification error of the classifier.

There can be many decision boundaries with the same error rate. Not all of them may prove a good generalisation of the actual decision boundary and be efficient in prediction. The criterion optimised by SVM minimises not only the classification error but also the width of the decision margin. The wider decision margin sup-

posedly means better predictions. The large margin may seem to be a competitive advantage, but for a well-collected data, it may not be the biggest advantage over standard linear classification methods. The wide margin classification is not the only one strong advantage of SVM. The definition of SVM makes the method very easily modified to utilise **kernel transformation** of the variables [67]. The kernel method allows for transformation from the space of original variables into new variables describing the similarities between observations. The linear classification on kernel features is equivalent to non-linear classification boundary in the space of original variables. To employ the SVM with kernel transformation, we have to choose the type of the kernel and its parameters. An illustration of the kernel method is in Fig. 3.9.



**Figure 3.9.** Illustration of the kernel method that transforms the non-linear decision boundary in the original feature space to a linear decision boundary in the kernel space. [81]

### 3.5.2.3 Artificial neural networks

The **artificial neural networks** (ANN) [75] may be viewed as a modification of the linear classification methods (a combination of perceptrons). The ANNs are a diverse category of methods inspired by the neural system of animals. The ANNs reach from the multilayer perceptron [82] on one side of the spectrum to the deep networks [83] on the opposite end. The ANNs are a subject of substantial development, and the most recent breakthroughs in machine learning produce the ANNs. The ANNs have a long history, the initial methods which may be classified as ANN appeared around half of the 20th century. The methods very much differ in their capabilities as well as in their complexity. The capabilities of deep convolutional ANNs are impressive, but the application of deep neural networks on the biochemical data would not be appropriate due to several reasons. The biochemical data from typical experiments are too small (usually tens to hundreds of observations) to learn such complex model (which require at least thousands or tens of thousands of observations). Second, the model is highly complex and rather a 'black box' - the **interpretation of a trained model would be extremely difficult**. We usually want to interpret the model behaviour in the data analysis of scientific experiments. Regardless of the cutting edge research in machine learning, the ordinary ANNs are suitable for applications in data analyses of biochemical data. The advantages of ANNs are that they can approximate any decision boundary without any modifications to the original data as in the case of linear classifiers, where the nonlinear boundaries could be approximated with enhanced data sets by the polynomial expansion of variables or by kernel method. The ANNs, which

consist of artificial neurons – perceptrons – arranged into layers (typically input layer, several hidden layers and output layers) create new variables as linear combinations in the hidden layers of the network and classify them in the output layer. The simple ANNs are more powerful for the classification, but it makes the interpretation of their behaviour difficult. The interpretation cannot rely on the understanding of the interpretation of network coefficients. The interpretation requires feeding the network with artificial data. The neurons 'activation' can then tell us what the model does. However, this approach of presenting specific patterns is not anything that can be easily presented to collaborators or in scientific papers. Therefore even the simple ANNs are considered to be rather 'black boxes' for data analysis. Another use for the ANNs in data analysis may be for comparing other methods on the same data. The ANN may help to set an upper bound of classification performance, and simple easily interpretable classifiers can be trusted if they score similarly to the ANN.

#### 3.5.2.4 Multivariate methods

The last but not the least important in the analysis of biochemical data are the **multivariate methods** – the partial least squares discriminant analysis (PLS) [33] and supervised principal component analysis (SPCA) [76]. Both the methods are very similar, and as the name of the second implies, they are related to the unsupervised method, the principal component analysis. The advantages of the regular PCA were that it could create a new set of variables, which is a combination of the original variables. The PCA creates the new variables so that it minimises the covariance in the data – under special conditions (centered and scaled variables) the new variables are uncorrelated. The idea behind the multivariate models is the following – we can create a set of variables which extract only the variance that is related to the response. In other words, combine the original variables in a way to maximise the classification performance on the new variables. The simplest approach which may be successful to certain extent is the **principal component discriminant analysis** (PCDA, the **principal component regression**, PCR, for continuous outcome) [84]. The idea of PCDA is to use the estimated principal components for classification. This approach does not use any specific procedure to make the new variables any good for the classification. What may be less obvious is, that if there is the relationship allowing for good classification, it has to translate into the principal components, because it is a source of variability in the data. In case that the difference between the classes is not a substantial source of variation in the data, the principal component related to the classification problem may be any of the principal components and the number of the principal components is in general equal to the number of original variables. In comparison of the naive procedure of PCDA [84], the PLS-DA [33] and SPCA [76] actually create new variables with the classification objective. The new variables are devised to provide the best possible classification. The advantage of these multivariate models are partially the same as in case of PCA, the new variables are uncorrelated and summarise several correlated original variables. This characteristics of the multivariate methods are ideal for data, which are inherently highly correlated. The disadvantage of this multivariate methods is that they estimate the full covariance matrix – they expect that relationship may exist among any pair of variables. There are types of data, where this is an actual problem. These are typically the NMR-based spectroscopy data. Other types of data may not need such a general model of collinearity – for example time series data and their relatives. Another **advantage of the multidimensional methods is that they are**

**capable of dealing with ill-posed problems** [85]. By ill-posed, we mean a data which consist of fewer observations than variables (the data matrix has fewer rows than columns). Standard estimation methods like ordinary least squares [67] rely on matrix inversions. In the case of ill-posed problems, the matrix is non-invertible. PLS and SPCA methods can overcome this problem. In the case of PLS by employing the iterative estimation of one new latent variable and its subtraction from the data, the PLS-DA can overcome the problem with matrix invertibility. In the case of SPCA, the solution relies on the SVD decomposition, which the method uses for the parameters' estimation of the new variables. The properties of both algorithms will be more clear from the algorithms. The methods are described in more detail in Chapter 5 or in the original articles - for the PLS refer to [33] and for the SPCA refer to [76].

### ■ 3.5.3 Summary

The list of methods capable of dealing with the specific properties of biochemical data could be longer. The exhaustive list would not provide the reader with any substantial additional knowledge; several works provide exhaustive reviews, refer to [12], which deals mostly with NMR spectroscopy. However, the methods apply to other fields. In this section were reviewed the most prominent methods for data analysis typical for machine learning approach. The first mentioned methods were the decision trees. Decision trees are advantageous for their simplicity and interpretability. The simplicity of the method, unfortunately, implies that the ability of the method to estimate complex decision boundaries and the performance in classification of heterogeneous and collinear data is poor in comparison with other methods.

Second, we mentioned methods from the wide family of linear classifiers. The extent of the linear methods means that there is also wide variability in the properties and classification performance. Although the methods may differ, the common denominator is the linear classification boundary. The estimation procedure gives the resistance of the methods to the collinearity. Some linear classifiers can deal with the collinearity quite well; others may fail to provide reliable results. The linear classifier is reasonably simple to be interpreted and understood by researchers without rigorous education in mathematics and statistics. The linear decision boundary is more flexible than the crude thresholding of individual variables in case of decision trees, but it is not able to estimate more complicated decision boundaries without the use of polynomial or kernel features.

The third mentioned classifier was again a group of methods called the artificial neural networks. The properties of ANNs are such that the classification ANNs can estimate any decision boundary and their nature is such that the collinearity does not pose any problem. The main drawback of the ANNs is the very limited possibility to interpret the coefficients of the network and their relationship to the classification of observations. From the point of the person analysing the data, the ANN classifier behaves more like 'black box' type system. This properties of ANNs substantially limit the class of problems, where they may be applied. There is plenty of problems where we demand the perfect classification, and the structure, coefficients and other details of the classifier are not the main interest of the effort. However, the **research in biochemistry is usually not the field, where we can neglect the inner workings of classifiers**. In the biochemistry, we aim the data analysis at the understanding the phenomena relating the variables to the response in the form of classification. Thus the possibilities to use ANNs in the field of biochemistry research are only a few.

The fourth and the last mentioned were the multivariate methods related to the princi-

pal component analysis, the partial least squares discriminant analysis and the supervised principal component analysis. These methods were developed to solve problems with highly collinear data and problems with fewer observations than variables (ill-posed problems). Therefore these methods are perfect for most of the problems dealt with in the analysis of biochemical data; in some areas, such as NMR spectroscopy, these methods are the methods of choice, in other areas they may be less effective. These models are complex. The models are hierarchical. There is a set of estimated latent variables from the original data, which form the first level of the model hierarchy. The latent variables serve on the second level of the model hierarchy for the final classification. [86]. The models are interpretable; the loading vectors tell us which variables constitute the latent variables. The role of latent variables is obvious from the classifier in the second stage. The estimation of latent variables deals with collinearity in the data by summarising the data by a small set of variables that are linear combinations of the original variables. The problem connected with the multivariate models is in the data, on which we usually apply these methods. The data with more variables than observations are susceptible to overfitting, and it is very difficult to decide whether the classifier is reliable. Whether the model is not a result of pure chance. There are many approaches for testing model for overfitting and as useful reference provides [87–88], specifically for metabolomics [26].

The regression refers to models and methods of relating variables to a response variable that is of numeric (continuous) character. A regression modification exists to each of the method that was mentioned in the classification section. There exist regression trees, linear regression, ANNs for regression and partial least squares regression as well as the supervised principal analysis (able to solve regression problems). In biochemical data analysis, the resistance to collinearity of regression is the same as for the classification. Regression is a wider concept in statistical modelling. In statistics the regression is the umbrella term for all modelling approaches – for all types of response variables – continuous, integer, count, binary [80]. We will discuss the statistical approach to regression in the following section.

Method	Collinearity	Relationship type	Feature selection	Feature extraction
Decision tree		constant	•	
Logistic reg.		linear		
SVM		any		
Ridge reg.	•	linear		
LASSO reg.	•	linear	•	
ANN	•	any		•
PLS	•	linear		•
Sup. PCA	•	linear		•

**Table 3.1.** Summary of machine learning methods concerning their ability to deal with various properties of the data

### 3.5.4 Statistical approach to the data analysis in biochemistry

The long history of development in statistics provides ample methods for various types of data and problems. This section provides a list of statistical methods capable of dealing with collinear data and other problems of the biochemical data. The section will discuss the extent of statistical approaches from simple tests to more complex modelling techniques. The statistics relies on the probability theory; the probability theory

concentrates at the prediction of the probability of phenomena based on a probability model. Typically, an example of the use of the probability theory is various games such as throwing dice. The probability theory provides means for the estimation of results of dice throwing. The statistics is interested in the description of the probability models from the results of experiments. In the simple example of dice throwing, we might have reasoned about the probability model of the dice, whether it is a 'fair' dice or whether any of the values on the dice is more or less frequent than would comply with the assumed probability model. In the data analysis performed by the statistics, we may be interested in **description of the results of the experiments**. The methods for description and summarisation of the data provides the descriptive statistics. The statistical inference allows for estimation of properties of the descriptive statistics. Using the statistical inference, we may examine the descriptive statistical parameters with estimates of variation, confidence intervals or for example, the distribution of the parameters. With the introduction of assumptions about the so-called null hypothesis, we may test, whether the descriptive values are in accord with the distribution of the null hypothesis or whether they differ significantly from the distribution of the null hypothesis [89]. The statistical tests can estimate how much the observed descriptive statistical parameters deviate from the expected values following the null hypothesis. The field of statistical modelling offers methods for estimation of parameters between variables, and the parameters may be tested to estimate the statistical significance of their values.

#### 3.5.4.1 Standard statistical inference

We divide the statistical inference methods into three main approaches. The first is the **classical statistical inference**, which uses the analytical approach to the computation of the parameters and properties of the null hypothesis distributions. The classical statistical inference historically preceded the other approaches, which use computational machinery such as computers. The classical statistics relies on strong assumptions about the data, which consequently allow for the computation of the parameters of the inference by standard statistical distributions directly or as asymptotic approximations. In classical statistics, the standard distributions are either approximated or summarised in tables, which allow for the finding of test characteristics with reasonable precision. The advantage of the classical statistical inference is its well-developed theory. The assumptions the classical statistics relies on may be very easily violated, and in consequence, the inference may not be correct. The classical statistics provides a wide variety of inference methods and tests for most of the problems dealt with in data analysis. To name an example, the typical problem in statistics is testing whether two groups differ in specific parameters – most often whether a sample mean is different between two samples. For such a problem the classical statistics offers **Student's two sample test** [89]. The Student's test assumes the samples to be drawn from the normal distribution and also assumes the variance equal among the samples. With this assumption, the difference between groups may be assumed to follow the normal distribution, and this may be used to compute characteristics such as critical values or p values.

#### 3.5.4.2 Non-parametric statistical inference

The second approach to statistical inference is the **nonparametric approach** [31]. While in the case of the classical statistics the inference relied on standard statistical distributions and thus forming the parametric approach, the non-parametric statistics tries to lift some of the assumptions about the underlying statistical distributions and use other means to perform the statistical inference. The non-parametric statistics provides 'workarounds' for the assumptions of classical statistics, the statistical power (the ability to reject a truly invalid null hypothesis) is lower than in the classical statistics. The non-parametric statistics can provide us with correct inference for cases where the classical statistics would fail. The non-parametric statistics is very close to the classical statistics. Even the non-parametric statistics uses approximations similar to the classical statistics for large sample sizes. An example of a non-parametric test for the problem of the comparison of the sample mean is the Wilcoxon's rank sum test [90], the Wilcoxon's test does not assume the data to come from any specific statistical distribution. Ranks represent the samples; the ranks are used to compute the rank sum. For small sample sizes, the rank sum determines the inferential characteristics; for large sample size, the test uses the normal approximation. Apart from the lower statistical power, the Wilcoxon test is also sensitive to the empirical distribution of the sample. If the distributions differ substantially, the Wilcoxon's also tests the difference in the location, but also the difference in the samples empirical distribution's shape.

#### 3.5.4.3 Simulation-based statistical inference

The third approach is the **bootstrap and other simulation based methods** of statistical inference. The bootstrap is the most recent contribution to the statistical inference and its development by Efron revolutionised the statistics [50]. The idea of bootstrap is very close to the intuition behind the nature of probability, chance and statistics. With few simple assumptions the bootstrap can simulate the stochastic processes which generated the data. The idea of bootstrap is the analogy between the population and sample, and the sample and bootstrap sample. In classical statistics the sample drawn from the population is used to study the statistical properties of the population. In bootstrap the bootstrap sample can be used to study the statistical properties of the sample. Thus what is the sample for the population is what is the bootstrap sample for the sample. The bootstrap is similar to the permutation tests, which are established in the classical statistical inference. For simple systems of discrete phenomena, we can exactly compute all outcomes and perform the inference exactly. These are the so called exact test and the most famous example is the Fisher's tea tasting experiment [91], which led to the development of the Fisher's exact test for binomial distribution. Similarly, the permutation test evaluates the complete set of possible outcomes; by permutation of the outcomes (typically in case of two sample tests), the distributions of null hypothesis may be estimated and used for the inference. The problem with permutation tests (and also with bootstrap) is that set of all possible permutations may be effectively so large, that the evaluation of all the permutations would not be possible (the number of permutations is the factorial of the sample size). For such cases, the Monte Carlo methods provide a good solution. With Monte Carlo methods we evaluate only a reasonable number of permutations.

And finally to introduce the bootstrap sample – the bootstrap sample is a sample with replacement of the size of the original sample drawn from the observations in the original sample. The sampling with replacement allows for very extensive statistic inference;



we can use the bootstrap for the estimation of the distribution of null hypothesis for any one, two or more sample tests as well as for inference in statistical models. **The lack of any strong assumptions for the bootstrap makes the method of choice for any problem where the classical approach to the statistical inference cannot be applied or would not be appropriate.** The apparent disadvantage of the bootstrap method is that in order to estimate the distribution of the null hypothesis, we have to repeat the bootstrap sampling and the number of random repetitions depends on the precision we need. In case of simple tests we may settle with several thousand repetitions, for cases where correction for multiple testing come into play, the number of repetitions may be several orders higher. This need for a large number of repetitions also clearly points to the weaknesses of the bootstrap approach – the repeated estimation of statistical parameters needs reasonable computing power. For simple tests any contemporary computer is sufficient, but with the increase in repetitions of the computation of the examined statistical parameter, the considerations about the efficiency of the computations and efficient use of the resources (multicore processors, computer grids) become a serious problem. The borderline between simple and difficult problems regarding computer time is narrow. Simple problems take several minutes, intermediate problems run for several hours, and serious problems can take months. A simple problem can become easily intermediate when we decide to control our simulations for additional effect. We can usually roughly estimate the total time that the complete number of repetitions will take. The other problem with bootstrap is that apart from several standardised bootstrap test (for example test for a difference between two samples, estimation of confidence intervals for coefficients of a linear model), the **bootstrap test has to be programmed and run with code developed specifically for the given problem.** We cannot advise programming bootstrap from scratch in any programming language. In order to ensure that the bootstrap samples are random and independent, the program needs a good random number generator. The statistical software such as R [92] or programmes for scientific computation come with the utilities to support for pseudo-random computations with guaranteed independence among the (bootstrap) samples.

#### 3.5.4.4 Summary

To sum the statistical inference techniques, the classical statistics and non-parametric approach are readily applicable for a variety of problems, but in order to perform the inference correctly, we have to check the assumptions of chosen techniques. The advantage of these methods is that they are well-established and supported by the probability theory. The bootstrap approach is beneficial for problems, where the classical techniques are not applicable due to their assumptions or where the methods do not exist. The bootstrap allows for statistical inference even in very complicated designs but may be very computationally intensive.

### 3.5.5 Multiple comparisons

In the analysis of multidimensional data, it is common to pose more questions during the data analysis concerning one collected set of data. Testing more than one hypothesis on data collected from one experiment may lead to overly optimistic results. In statistics, the hypotheses are considered valid, when the probability of an error made by refuting the null hypothesis is lower than a threshold chosen before carrying out the analysis. By

performing several tests on the same data, the probability of making an error increases. The intuition of increasing the risk of making an error may be illustrated on a simple example with throwing a coin (see Ex. 3.2).

The probability of throwing head with a fair coin is  $1/2$ . In case of throwing the coin twice, the probability of at least one head is intuitively higher, it is exactly  $3/4$  (the probability may be easily computed by basic rules for computing with probabilities or by following the formula for probability of binomial distribution). The same applies in case of statistical hypotheses, only the error is no more a fair coin. The probability for the standard threshold of 0.05 will result in 0.0975 for 2 hypotheses, but it will be 0.226 for 5 hypotheses, 0.401 for 10 hypotheses, 0.641 for 20 hypotheses and 0.923 for 50 tested hypotheses.

**Example 3.2.** The probability of making the type I error when performing statistical tests illustrated by dice throwing.

The presented number of hypotheses may seem large; however, in several fields of research, the number hypotheses may be even many times higher. Obtaining reliable results in case of testing multiple hypotheses therefore poses a substantial issue for statistical analyses. In literature the problem is referred as **corrections for multiple testing or as family-wise error rate**. The simplest approach to eliminate the problem of multiple testing is lower the threshold for refuting the null hypothesis in way that would ensure the overall error rate not to exceed the desired level (typically 0.05). This may be achieved by dividing the threshold by the number of tested hypotheses. This approach is called the **Bonferroni correction for multiple comparison** [93]. By figuring out the probabilities of the error in this scenario, it is obvious, that the Bonferroni correction is conservative, the actual error rate maintained by the Bonferroni correction is lower than the 0.05 error rate. Better corrected threshold may provide the **Šidák's correction** [94]. The two presented methods for multiple comparison corrections are only a small sample from a variety of methods applicable for multiple comparison correction. The advantage of the mentioned methods is in their simplicity, which makes them easily applicable. In contrast to manipulating the threshold, we may work with the actual probabilities of making an error – the  $p$ -values estimated by the statistical tests and apply the notion of **false discovery rate**, which tries to estimate the number of false discoveries rather than the probability of making one error [95]. By accumulating the probability of the error, we may refute more hypotheses and obtain more interesting results. From these methods we may name the **Holm's procedure** [96] and **Benjamini-Hochberg-Yekutieli method** (BHY) [97]. These methods are more complex and not as easy to use because we have to estimate the  $p$ -values, order them and sequentially process. Standard programs for data analysis contain functions for carrying out these procedures. The BHY method is interesting for correlated tests. Imagine two cases – in one case we test two hypotheses on one set of data, each on a different uncorrelated variable. For the second case imagine, that we tested a hypothesis by mistake twice on the same variable. In the first case, the error rate increases; in the second case, we can hardly tell that the error rate would change; it is the same. The BHY procedure accounts for the correlation among tests and allows for more sensitive treating of the  $p$ -values and better results (in the form of correct refusing of the null hypothesis under correlation).

In conclusion, multiple testing procedures form a substantial issue in statistics. There exist several methods. Simple methods such as Bonferroni or Šidák correction for

multiple comparison work by modifying the threshold value for refusing the null hypothesis. These methods are easy to use but may prove to be overly conservative. Another group of methods works directly with  $p$ -values and allows for a more complex approach to hypotheses' testing. These methods are more difficult to apply.

### ■ 3.5.6 Statistical modelling – regression

As mentioned before, we would not list the statistical models applicable to (biochemical) data with only one exception – the **generalised linear mixed effect models** (GLMM) [98]. The GLMMs are the state of the art method for most data types. It is good to start with just generalised linear models (GLM) [80] to understand the GLMMs. And to understand the GLM, it is best to start with an ordinary linear model (LM) [99]. The ordinary LM is not just the linear regression which most researchers associate with LM. The LM covers several methods, usually denoted by different names. An LM relating (continuous) response to a logical (indicator) predictor is equivalent to a two sample Student's test with equal variances. In such a case, the model coefficients are the mean the reference group and a difference between the means. We can easily extend this model to a categorical predictor represented by a matrix of indicators (indicator for each category of the predictor; meaning the original predictor is either equal to a given category or not). This case in LM is equivalent to testing the difference in the means of the response variable to a baseline value (one of the categories) by two sample Student's test with equal variance. The case of continuous predictor is the well known linear regression. The previous cases presented the simplest LMs with only one predictor. LMs with more than one predictor variable offer a wide variety of relationships between the variables (predictors to the response). The description of all possible cases is too long to be presented in this work. We direct the reader to [99]. When using the LM we are restricted to cases where the response is a continuous variable. There exist models for other types of response variables. There is the Poisson regression for count data, the logistic regression for the binary response variables and the ordinal (not ordinary) regression for the categorical variables. All of these regressions can be used in the same manner as the LM and used to assess the same types of hypotheses. However, there does not have to be any relationship to any existing statistical test.

The idea of GLMs is that all these specific regression methods are special cases of an LM with a transformed response variable. For example, the Poisson regression modelling the integer variable, the natural logarithm of the response can be taken and used for the modelling. The natural logarithm transforms a non-negative integer variable into a continuous real-valued variable. When estimating the model, we have to consider the relationship between the mean and the variance of the distribution. We can formulate the logistic regression as an LM relating predictors to binary-valued response variable transformed by logit function. The transformation is called a link function which linearises the response in order to be able to work with as in an ordinary regression (however the details of the optimisation procedure may slightly differ from the ordinary LM). The variance function describes the relationship between the mean and the variance of the distribution. The link function and the variance function define the generalised linear model. Although the specific regressions may not relate to existing tests, the GLM methodology allows derivation of asymptotic distributions of the parameters so that the statistical evaluation can be easily carried out by standard programs for statistics such as R.

And finally, the mixed effect model is a concept of a model for data which suffer from

substantial variation due to the measurement procedure. An example of such data may be the evaluation in sports such as acrobatics, ice-skating or horse riding, where there is no objective measure of the contestants' performance, but a group of judges assesses it. Each judge at a competition would unintentionally have her, or his biases and the assessment of one contestant by different judges may differ. Consider a problem of modelling performance assessed at competitions of one contestant in acrobatics based on the hours per week spent by practising. The judges will never be the same persons. However, few of them may be assessing the performances in several competitions. The differences between the judges' assessments may be so important, that they completely obscure the relationship between the training and the final evaluation – in this case, the standard modelling procedure would fail to recognise an actual relationship. Another extreme case may be an observed relationship between the evaluation and the training caused by chance due to the judges' biases (say the contestant practised the wrong type of exercise). The mixed effect model consists of two parts – a fixed part and a random part. In our example of acrobatics, each judge can correctly evaluate the performance resulting in the same (or similar) order of contestant, but each judge expresses her or his opinion in a different number of points. In order to model the relationship between performance and training, we would have to find a transformation of each judge's points to compare them. The mixed effect model allows to estimate the transformations for each judge as a random effect parameter and eliminating the individual differences. The common variation among the individuals forms the fixed effect – the changes in performance after compensation for the differences in judges' assessment and their relationship to the training. It is clear that the concept of mixed effect model greatly enhances the capabilities of the GLMs. The use of a mixed effect model needs careful consideration of the nature of the problem, e.g. which variables affect the response and whether they should be utilised as a random effect, fixed effect or both. Complicated model design compensating for various effects in the data may not translate well into a working mixed effect model. The estimation of mixed effect models is very complicated due to several limitations implicated by the very concept of the model (the random effects' coefficients are often many and are subjected to be random numbers drawn from  $N(0, \sigma^2)$ ). Similarly to the computational problems, the statical assessment of the model parameters is complicated. The asymptotic distributions for parameters can be derived, but only for specific cases satisfying strict assumptions [98]. When we decide to use a mixed effect model, we can rarely be sure that all the model assumptions hold. Therefore, non-parametric approaches such as bootstrap may provide better results [100].

The GLMMs, as presented in the previous paragraph, suffer from one substantial drawback. Several assumptions of the LMs which translate into GLMs and to a certain extent to GLMMs are very limiting. The first assumption that is often not met is the independence of residuals in the model. We often observe dependencies among model residuals. The causes for the dependencies among the residuals may be due to a wide variety of reasons. In simple problems, the dependency can indicate another linear or non-linear relationship, that is present in the data and is affecting the studied relationship. The reason for the dependency may be the properties of the data. Consider data collected from a group of individuals in time. We refer to this data type the time series data or longitudinal data. In other contexts, it is called repeated measurements or called specifically by the nature of the dependency, e.g. data with spatial covariance structure. The other important assumption of the LMs is the homoscedasticity, which means that the variance in the data is constant and unrelated

to the response. Homoscedasticity is the reason, why the two sample Student's test with equal variances was mentioned as equivalent to the LM with indicator predictor. The equality of variances is a very strong limitation, which allows for a correct assessment of only a small proportion of hypotheses. In the standard statistics, there is the Welch's test – the alternative to the Student's test under unequal variances. In case of linear models, the alternative is the general linear model (not generalised linear model) which enables estimation of models for data which are dependent and also heteroscedastic. These two properties cause substantial problems during the estimation of the general linear model – the optimisation of the parameters. The good news is that the methodology of GLMMs already allows for the utilisation of the dependence and heteroscedasticity – the problem of these assumptions is related to the random effects in the GLMMs, which are on itself an example of dependency, although expressed in different form.

### ■ 3.5.7 Summary

To wrap up the section on data analysis, we repeat and emphasise the most important points. Generally, two approaches exist for the data analyses in biochemistry. The machine learning methodology provides one way to process biochemical data. The machine learning approach to data analysis classifies the methods into supervised and unsupervised methods according to the analysed outcome. In unsupervised analyses of data, the clustering examines, whether the data do not have any hidden structure. Data with structure should be treated differently from unstructured data; we may exploit the found structure with modelling techniques.

Another important method from the class of unsupervised learning is the principal component analysis. The principal analysis can help to assess the relationships among variables. The correlation structure may be used to create a new set of uncorrelated variables which may prove beneficial for visualisations and generally for data exploration. The properties of such variables predestine them for use in models, which are not capable of dealing with correlated variables.

The supervised learning rests in learning to approximate an expert knowledge – typically a class assignment in classification or prediction of continuous values in regression. For the classification of biochemical data, several classifiers are applicable, but only a few are appropriate. The best method for classification of biochemical data is various linear classifiers. In linear classifiers, the benefits and drawbacks are in good compromise. The linear methods provide sufficient complexity to approximate and correctly classify even data, that are not linearly separable. The models are not difficult to interpret. A good example of such a linear classifier is the support vector machine. Second in the line of reasonable classifiers are the multivariate methods, namely the partial least squares discriminant analysis and supervised principal component analysis. These methods are complex enough to fit complicated decision boundaries and are still interpretable. These methods are better at dealing with correlated variables and therefore are prominent in areas, where the collinearity poses a substantial problem, e.g. in NMR spectroscopy. Other methods are less appropriate than the already mentioned for the biochemical data. The decision trees and related methods are easily interpretable, but the data without any feature extraction methods are far too complex for such simple methods. The other extreme is the neural networks, which are extremely powerful at classification but are uninterpretable to the extent of black box system.

The statistical approach to the analysis of biochemical data rests in the rigorous application of statistical methods. We have to choose the appropriate method and assessment

of statistical significance to perform the statistical analysis correctly. The choice of the method relies on careful consideration of the tested hypothesis – for example, for group comparison we can apply a test of difference. The statistical assessment needs an examination of the method's assumptions. For the case of group comparison, several tests are applicable. The applicable methods are the Student's two sample test, the Welch's two sample test, the Wilcoxon's rank sum test or a bootstrap test for difference in means. The first two tests are classical statistical methods, and they assume several data properties in order to provide correct results. The Wilcoxon's rank sum test belongs to non-parametric methods, and so it has less assumptions than the Student's test, however also a lower statistical power and in some cases does not have to be a test of difference at all. The bootstrap imposes the lowest number of limitations on the tested data. However, it may happen that the used statistical problem would not have the test implemented and therefore the testing procedure would have to be programmed.

When performing a statistical test, it is also important to keep track of the number of tested hypotheses. Each tested hypothesis increases the chance of finding anything statistically significant. Unfortunately, the multiple testing comes at a price – the chance of making an error increases accordingly. The corrections for the multiple testing have to be applied to avoid the risk of drawing incorrect conclusions. We can perform the corrections following Bonferroni's or Šidák's methods, which rely on manipulating the  $\alpha$  level for the refutation of statistical hypotheses. More advanced methods work with the estimates of the probability of making an error, these methods are more sensitive, but we need statistical programs for the computations. Lastly, the very important method, the generalised linear mixed effect models were presented alongside their advantages in the analysis of biochemical data.

# Chapter 4

## Generalised mixed effect models

In the preceding section 3.5.4, we provided the intuition behind the generalised mixed effect models (GLMM) and their application. In this section, we will provide a detailed description of the model and its components. The methods of model parameters estimation will follow. An essential part of the reasoning in experimental science is the statistical inference about the model parameters. Therefore we will list the methods available for the statistical inference about the model parameters, as well as a detailed description of simulation methods. We consider the simulation methods to be the methods for appropriate assessment of the statistical significance. We will start with the simpler linear mixed effect model, and subsequently, the GLMM will be devised based on the description of LMM.

### 4.1 LMM model description

The linear mixed effect model is a statistical two-level hierarchical linear model, which combines fixed effects and random effects. The following equation describes the linear mixed effect model.

$$y = X\beta + Zu + \epsilon \quad (4.1)$$

where  $y$  is the  $n \times 1$  vector of the response variable,  $X$  is the  $n \times p$  matrix of predictors for the fixed effect part of the model,  $\beta$  is the  $n \times 1$  vector of the fixed effect coefficients of the model,  $Z = [Z_1, \dots, Z_b]$  are the design matrices, where  $Z_i$  is the  $n \times q_i$  matrix of predictors for the specific random effect part of the model,  $u = [u_1, \dots, u_b]$  is the  $q \times 1$  vector of the coefficients of the random effect part, where the individual  $u_i$  is a  $q_i \times 1$  vector such that  $q = \sum_{i=1}^b q_i$  and  $\epsilon$  is the  $n \times 1$  vector of residuals. Both the random effect coefficients  $u$  and the residuals  $e$  have zero mean  $E(u) = 0$  and  $E(e) = 0$ . [101–102]

The random effect coefficients  $u$  and the residuals  $e$  are assumed to be independent and drawn from a multivariate Gaussian distribution with zero mean

$$\begin{bmatrix} u \\ e \end{bmatrix} = N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} G(\gamma) & 0 \\ 0 & R(\rho) \end{bmatrix} \right) \quad (4.2)$$

The  $\gamma$  and  $\rho$  are  $r \times 1$  and  $s \times 1$  vectors of variance parameters corresponding to the random effect coefficients  $u$  and residuals  $e$ . The variance-covariance matrix of the response  $y$  is then:

$$\text{var}(y) = \sigma^2 (ZGZ^T + R) = \sigma^2 H \quad (4.3)$$

The presented factorisation of residual variance out of the variance matrix of the data adopted from [101] is advantageous for the estimation of the model parameters. Less general factorisation of the variance-covariance of the response  $y$  by assuming the matrix  $R$  to be the identity matrix may be: [101]

$$\text{var}(y) = V = ZG^{\nu}Z^T + \sigma^2I \quad (4.4)$$

The benefit of the factorisation presented Eq. (4.3) is that the variance and covariance in the residuals can be estimated separately.

## 4.2 LMM parameters estimation

The estimation of LMM parameters poses several problems in comparison with the estimation of parameters of the simpler linear model. We can formulate the estimation of LMM model parameters in several ways. The work [101] presents an approach based on solving the mixed model equations. It also lists other formulations of the problem of estimation. The work [102] proposes a more general approach as a minimisation of a loss function. The parameter estimation devised as a solution of mixed model equations is the following:

$$\begin{aligned} \hat{\beta} &= (X^T H^{-1} X)^{-1} X^T H^{-1} y \\ u &= GZ^T H^{-1} (y - X\hat{\beta}) \end{aligned} \quad (4.5)$$

These equations can be used for the estimation only for known variance parameters  $\sigma^2$ ,  $R$  and  $G$ . Usually, these parameters are unknown, and we have to estimate them from the data. It is necessary to use an appropriate estimator of the variance parameters  $\sigma^2$ ,  $R$  and  $G$ . There are two approaches to the estimation of the variance parameters, the full information maximum likelihood method or the residual maximum likelihood method [102].

### 4.2.1 Maximum likelihood estimation of LMM parameters

The maximum likelihood estimation of the LMM parameters rests in devising a likelihood function - generally a function of model parameters and the data. The likelihood function assesses, how well the model with a given set of parameters fits the data, or more exactly, what is the chance, that the model with the given set of parameters generates the observed data. The optimal values of the parameters are those, which yield the biggest value of the likelihood function, hence the maximum likelihood estimation. The maximum likelihood estimation has many applications in statistics. We can devise the closed-form solution for the maximum likelihood estimation for the estimation of some statistical parameters (e.g. sample mean). More often, there is no closed-form solution, and therefore, we have to utilise the methods of numerical optimisation to find the set of parameters yielding the maximum value of the likelihood function. We usually use the log-likelihood function instead of the likelihood function. The advantage of the log-likelihood is the transformation of products in the likelihood function to sums in the log-likelihood function. The log transform does not change the position of the maximum of the function.

In the case of the LMM, the log-likelihood function of the model parameters and the data is of the form:

$$\begin{aligned} l_{ML}(\beta, \phi|y) &= -\frac{1}{2} \left( n \log 2\pi + n \log \sigma^2 + \log |H| \right. \\ &\quad \left. + \frac{(y - X\beta)^T H^{-1} (y - X\beta)}{\sigma^2} \right) \end{aligned} \quad (4.6)$$



where  $\phi = (\kappa^T, \sigma^2)^T$  and  $\kappa = (\gamma^T, \rho^T)^T$ . The gradient optimisation methods can use the analytically devised partial derivatives. For the more detailed description of the likelihood function of the LMM variance parameters refer to [101]. The maximum likelihood approach is known to be downward biased because they do not take into account the degrees of freedom lost when estimating the fixed effects.[101]

### 4.2.2 Restricted maximum likelihood estimation of LMM parameters

The unbiased estimation of the LMM parameters provides the restricted maximum likelihood estimation. In comparison with the maximum likelihood approach, the likelihood is not computed with the original response  $y$  data, but with linearly independent error contrasts orthogonal to the design matrix  $X$ . The linear combination in form  $K^T y$  are chosen so that the rank of the  $K$  matrix is maximised but is free of the fixed effect coefficients  $\beta$ . The linear combinations are obtained from the data after fitting the fixed effect coefficients  $\beta$  and therefore they are function of the residuals. That is why the restricted maximum likelihood estimation is referred to as residual maximum likelihood estimation. The restricted log-likelihood function is: [101]

$$l_R(\phi, K^T y) = -\frac{1}{2} \left( (n-p) \log 2\pi + (n-p) \log \sigma^2 + \log |K^T H K| \right. \\ \left. + \frac{1}{\sigma^2} y^T K (K^T H^{-1} K)^{-1} K^T y \right) \quad (4.7)$$

where  $\phi = (\kappa^T, \sigma^2)^T$  and  $\kappa = (\gamma^T, \rho^T)^T$ . For a more detailed description of the restricted maximum likelihood refer to [101], which also presents the partial derivatives of the log-likelihood function needed for estimation of the coefficients by the means of numerical optimisation. [101]

### 4.2.3 Numerical optimisation methods for the LMM parameters estimation

The estimation of the LMM parameters either by the maximum likelihood approach or by restricted maximum likelihood calls for numerical optimisation methods. The optimisation can be performed by general-purpose optimisers or by optimisers whose properties are best suited for the specific properties of the optimised model. Another criterion by which we can distinguish the optimisation methods is whether they provide only the best estimates of the model parameters or whether they also provide other information about the estimates useful for the inference about the parameters.

To the optimisers used to fit the maximum likelihood and restricted maximum likelihood estimates of the LMM parameters belong the Newton-Raphson method, Fisher scoring, Expectation maximisation algorithm and Average information algorithm [101]. The Fisher scoring and Average information are variants of the Newton-Raphson algorithm. The Newton-Raphson method finds the roots of a function by iteratively refining the estimates. In each iteration the intersection between the first derivative of the function (the tangent) at a certain value  $x_0$  and the  $x$ -axis is found by solving  $\frac{x_1 = x_0 - f(x_0)}{f'(x_0)}$ . The value of the intersect  $x_1$  serves in the following iteration as the value, where the first derivative is used to find the intersect. This process continues until either we reach the maximum number of the iterations or we achieve the convergence to the zero value. For the algorithm to converge, we have to provide a good initial value. In the case of (restricted) maximum likelihood, the Newton-Raphson method can be applied the first derivative of the likelihood function, because the solution of  $f'(x) = 0$  corresponds to

an extreme value of the function  $f(x)$  which is the maximum of the concave function. The paper [101] is a good starting point for the other two similar methods - the Fisher scoring and the Average information algorithm. The advantage of the methods based on the Newton-Raphson algorithm is that they provide the asymptotic standard error of the estimates, which is useful for the inference about the parameters. [101]

Similarly to the Newton-Raphson algorithm, the Expectation maximisation (EM) algorithm is a general optimisation method. The EM method repeats the expectation and the maximisation step. At the  $n$ -th expectation step, the conditional likelihood  $l(\theta_n|y)$  given the parameter estimates  $\theta_n$  and the data  $y$ . At the consequent  $n$ -th maximisation step, the conditional likelihood  $l(\theta_n|y)$  from the expectation step is used to produce a new parameter estimate  $\theta_{n+1}$ . We repeat the expectation and maximisation steps until either we perform the maximum number of iterations or reach the convergence. [101]

## 4.3 Statistical inference of the LMM

The statistical inference of the LMM divides into categories by the model parts - we can perform the statistical inference for the fixed effect part, random effect part and the variance of the model. From a different point of view, we can classify the approaches to the statistical inference methods themselves - we can apply the standard statistical approaches as well as simulation-based methods. Recently, there was criticism about the standard statistical approaches to the inference about various parts of the LMM, and the simulation approaches have been recommended to be the preferred method for the statistical inference in LMM (and also in GLMM). However, we will present both approaches to the problem.

### 4.3.1 Standard LMM inference

#### 4.3.1.1 Standard inference about the fixed effect part of the LMM

The interest in the statistical significance of the LMM is in most cases in the fixed effect coefficients. We usually interpret the fixed effect coefficients and use them for concluding the studied problems. The modelling effort with the LMM should not be, in any case, only focused on the fixed effect coefficients.

In case of the models fitted by the maximum likelihood approach, we can perform the likelihood ratio test of the goodness of fit of a pair of models with equal random effect parts and differing fixed parts. The test statistic is:

$$MLRT = -2(l_{ML_0} - l_{ML_1})$$

where the  $l_{ML_0}$  and  $l_{ML_1}$  are maximum likelihoods of two models that differ by a  $k$  fixed effect parameters. The maximum likelihood ratio test  $MLRT$  asymptotically follows a  $\chi^2$  distribution with  $k$  degrees of freedom. We cannot compare models fitted by the restricted maximum likelihood approach by the presented procedure. The use of different error contrasts  $K'y$  makes the likelihood functions incomparable. [101]. We can use a modified Wald test for the inference about the fixed effect coefficients. The problem poses the estimation of the degrees of freedom for the distribution of the test statistic. The Satterthwaite approximation can be used to estimate the approximate degrees of freedom. The modified Wald test is not the preferred way to perform the inference of the fixed effect coefficients.

### 4.3.1.2 Standard inference of the variance of the LMM

The variance is an important part of the model. The design of the variance-covariance structure should reflect our understanding of the modelled problem as well as provide sufficient means to fit the data properly. Variance structures with either too strict (fewer parameters than necessary) or too loose (too many parameters) result in inaccurate standard error estimates of the fixed effects. The restricted maximum likelihood ratio test can assess the appropriateness of the variance structure of the model. A pair of LMMs estimated by the restricted maximum likelihood approach with variance parameters  $R_0$  and  $R_1$ , we compute the restricted maximum likelihood ratio as:

$$REMLRT = -2(l_{R_0} - L_{R_1}) \quad (4.8)$$

Where the  $l_{R_0}$  and  $l_{R_1}$  are the restricted likelihoods of the two variance parameters. The restricted maximum likelihood ratio test  $REMLRT$  statistics follows  $\chi^2$  distribution with  $k$  degrees of freedom, where  $k$  corresponds to the number of additional variance parameters in the  $R_1$  variance model.[101]

Another approach is the score test, which uses the scores and the information matrix to test the significance of the variance parameters. The advantage of the score test is that we only have to fit the model of the null hypothesis. The score test statistic is:

$$S(\kappa_0) = U(\kappa_0)' I^{\kappa_0 \kappa_0} U(\kappa_0) |_{\kappa_0} \quad (4.9)$$

Where  $\kappa_0$  is a vector of variance parameters,  $U(\kappa_0)$  is the score vector of  $\kappa_0$  and  $I^{\kappa_0 \kappa_0}$  is the portion of the inverse of the variance matrix associated with  $\kappa_0$ . Similarly to the REMLRT, the score test statistic follows  $\chi^2$  distribution with  $k$  degrees of freedom. [101]

Both the restricted maximum likelihood test and the score test are problematic, when the null hypothesis is on the boundary of parameter space (e.g.  $\sigma^2 = 0$ ). For strategies available to address this problem refer to [101].

### 4.3.1.3 Standard inference of the random effect part of the LMM

Some works proposed approaches to perform the standard inference about the random effects as well as the combination of random and fixed effect parameters. Performing any inference about the random effect parameters is a matter of debate. The random effect parameters can be underestimated (in the case of the maximum likelihood estimation), and their actual fitted values are not good estimates of the relationship they model. There are other statistical models, which we can use to model the relationships on the required level of hierarchy directly and which we can test directly. However, in some cases, the inference about the random effect can be useful diagnostics information. The work [101] provides an approach to perform the standard statistical inference about the random effect parameters as well as a reference to more specialised discussions for further study of the problem.

## 4.3.2 LMM inference based on simulations

The simulation approaches to the statistical inference about the LMM mainly utilise the idea of bootstrap. The bootstrap, which was developed by Efron, has become a widely used method for estimation of various properties in statistics (e.g. estimates of variance, estimation of statistical distributions, hypothesis testing). The decisive advantage of the bootstrap approach to statistical inference is its non-parametric nature, which is common to many methods based on bootstrap. However, the bootstrap

has also a substantial disadvantage which is high computation complexity of the simulations, which we need to perform and repeat many times in order to estimate the required properties.

The basic idea of the bootstrap is that the inference about the statistical sample can be performed by sampling the observations from the statistical sample with replacement and estimating the required property. This method applies the general statistical method of inferring the properties of a population by drawing a sample. There are many resources on the bootstrap methods in various fields, the book [50] provides a general introduction.

In the case of LMM, there exist several approaches to the bootstrap. Generally, there are three categories of bootstrap methods, which are:

1. parametric bootstrap
2. residual bootstrap
3. cases bootstrap

#### 4.3.2.1 Parametric bootstrap of LMM

The parametric bootstrap uses parametric estimates of the distribution of residuals  $\epsilon$  and distribution of random effects  $u$  to generate bootstrap samples. Usually normal distribution is assumed for the residuals  $\epsilon \sim N(0, \sigma^2 I)$  as well as for the random effects  $u \sim N(0, \hat{G})$ . The method of parametric bootstrap in LMM is presented in Alg. 4.1. [102]

- 1) Draw a random sample of random effect parameters  $u^*$  from a normal distribution with zero mean and covariance matrix  $\hat{G}$ .
- 2) Draw a random sample of residuals  $\epsilon^*$  from a normal distribution with zero mean and covariance matrix  $\sigma^2 I$ .
- 3) Use the randomly generated random effects and residuals to compute the random sample as  $y^* = X\hat{\beta} + Zu^* + \epsilon^*$ .
- 4) Compute the estimates of all the LMM parameters on the  $y^*$  sample.
- 5) Repeat the steps 1-4 to obtain the required number of simulations.

**Algorithm 4.1.** The parametric bootstrap of linear mixed-effect model

#### 4.3.2.2 Residual bootstrap of LMM

The general idea of the residual bootstrap of the LMM comes from the residual bootstrap of the simple linear model. The residual bootstrap uses the resampled estimated residuals (and in case of LMM the random effects' coefficients) to generate bootstrap samples of the data. These bootstrap samples can be used to estimate the distribution of some model parameters, construct confidence intervals or for example, to estimate standard errors. The algorithm of the residual bootstrap of the LMM is as follows in Alg. 4.2: [102]

#### 4.3.2.3 Cases bootstrap of LMM

The cases bootstrap is probably the simplest approach to the bootstrap in LMM. The cases bootstrap rests in resampling the cases. In a hierarchical model, the resampling can be done individually on each level of the model. The decision about which level of the model to resample relates to the problem that is solved. In some cases, we may want to sample with replacement from the whole subjects (of the indicators identifying

- 1) Draw a sample with replacement of random effect parameters  $u^*$  from the estimated random effects coefficients  $\hat{u}$ .
- 2) Draw a sample with replacement of residuals  $\epsilon^*$  from the estimated residuals  $\hat{\epsilon}$ .
- 3) Use the resampled random effects and residuals to compute the random sample as  $y^* = X\hat{\beta} + Zu^* + \epsilon^*$ .
- 4) Compute the estimates of all the LMM parameters on the  $y^*$  sample.
- 5) Repeat the steps 1-4 to obtain the required number of simulations.

**Algorithm 4.2.** The residual bootstrap of linear mixed-effect model

the random structure of the model). In other contexts, we may want to perform the resampling of the observations within the subjects defining the random effect structure. For the first case, a good example can be a model with repeated measurements; the second case can be a problem of dealing with longitudinal data. However, we can perform the resampling on both levels. The cases bootstrap is a simple algorithm Alg. 4.3. [102]

- 1) Draw a sample with replacement of  $X_i$  subsets of the dataset  $X$  corresponding to the levels of the variable that defines the random effects structure.
- 2) In each of the subsets drawn with replacement from the original draw a sample with replacement of the observations.
- 3) Combine the resampled data on both the levels to form a bootstrap sample of the data  $X^*$ .
- 4) Compute the estimates of all the LMM parameters on the  $y^*$  sample.
- 5) Repeat the steps 1-4 to obtain the required number of simulations.

**Algorithm 4.3.** The cases bootstrap of linear mixed-effect model

In the case of the cases bootstrap, subsequent steps depend on the problem and the resampled levels of the model. For example, in case of resampling of both levels with longitudinal data, we may assume, that we perform the bootstrap of a null hypothesis and use the estimated distribution of the parameters directly to compute a  $p$ -value, construct confidence intervals, etc. In other cases, we assume that we perform bootstrap under the alternative hypothesis, which we can use to estimate standard errors or, e.g. to construct confidence intervals.

#### 4.3.2.4 Other approaches to the bootstrap of LMM

Many methods utilise bootstrap. In the case of the LMM the previous paragraphs presented only the general ideas about the bootstrap in LMM. However, many authors combine these categories to devise new bootstrap methods. An article [100] provides a more detailed list which also provides a comparison of several approaches to the bootstrap in LMM.

## 4.4 Generalised LMM

We introduced the idea of generalised models in the previous sections about the data analysis. The same reasoning which provided for the distinction between the ordinary linear models and generalised linear models applies to the mixed effect models. Thus in order to fit a GLMM, we have to have a link function and a variance function. The link function linearises the response variable and enables to fit the model (almost) as



The application of the MCMC methods consists of many experiments with the model structure and parameters tuning. The experimenting with various model structures is a common problem in any modelling. The parameters' tuning is not very usual in the standard statistical data analysis. In the case of the MCMC methods, the user has to evaluate whether the Markov chains have converged, effectively explored the modelled distribution, and assess the dependency among the samples sampled from the distribution. These are just general problems with the MCMC methods. [106]

The convergence is very difficult to evaluate. There are no general results about the rate of convergence and other properties which would be useful for the setting up of the Markov chains. Usually, we run several Markov chains in parallel with random initial conditions. After a sufficient number of iterations, we compare the values to which the different chains converged. We assess the convergence of individual chains by plotting of the traceplots of the sampled parameters. The traceplots should show a transition period with possibly big changes in the samples' values followed by more stable and less variable values when the chain attains good values of the parameters. In case the user does not observe the convergence, we increase the number of iterations and repeat the whole procedure. When the individual chains converged, it is important to compare the values to which the chains converged. If the values (or their distribution) are identical, there is no problem, and the user can proceed to the next steps of the analysis. In the case of different values, there are one or more problems. Typically, there is a disagreement between the data and the model structure. The user should reconsider the formulation of the model and try to fit a simpler model. On the other hand, there can be a problem with parameters governing the sampling of the chains. In that case, the user has to refer to the specific method and adjust the parameters accordingly.

We usually judge the problems with the distribution possible values exploration by the samples dependency. The plots of the samples from the chains can show dependency. If there are slow transitions between the values, the chain most likely did not explore the distribution sufficiently. Another clue is a high autocorrelation among the samples, which indicates the same - insufficient representation of the distribution by the constructed Markov chain.

The autocorrelation of the sampled values also limits the utilisation of the simulated samples. High autocorrelation means that even though the user allowed the chain to sample a large number of samples, the number of actual random samples is smaller. In order to get independent samples for an inference about the model, we have to perform thinning. When performing thinning, we retain only every  $n$ -th sample and discard the rest.

We can use the samples of the parameters obtained by the MCMC methods for statistical inference. Here we have to remember that the MCMC methods are Bayesian methods and therefore, the inference should be rather Bayesian than frequentist. The Bayesian approach is usually not interested in point estimates (which is common in the frequentist statistic inference), but in the distributions of the parameters. In Bayesian statistics, the prior information about the modelled problem in the form of the prior distribution of the parameters is also an important part of the inference following the Bayes theorem, which is loosely

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The data provide evidence in the form of the likelihood. Our prior beliefs about the relationship are in the form of the prior distribution. The prior belief and the evidence are combined and result in the posterior probability of the parameters. The prior distribution can often be a problem; in many cases, we do not have any prior beliefs or any

other evidence from previous trials to use as the prior distribution. In such cases, we can use an uninformative prior distribution – a distribution that gives approximately uniform probability to all possible values of the parameter distribution. To substitute the distribution of a parameter generated by the MCMC method with a point estimate (for a presentation, or better understanding) we can use a simple expected value of the distribution. In many simpler statistical models (such as linear regression) the expectation of the posterior distribution coincides with maximum likelihood estimates. We usually do not perform the inference in the Bayesian framework by simple p-values. The Bayesian framework uses a credible interval. The credible interval is very similar to the standard confidence interval. The credible interval, which covers the value of an unobserved parameter with a given probability is more intuitive than the confidence interval. The confidence covers the true value of an unobserved parameter in a sufficiently large number of random trials with a given probability. The prediction with the models trained by the MCMC methods is problematic. We can perform the prediction with sampling from the joint distribution of the data and the parameters. The result is a distribution. We can utilise the expected value to obtain the more desirable point estimate. [107]

In conclusion, the MCMC methods provide a versatile tool for fitting of complicated LMMs and GLMMs. The MCMC methods are nowadays available in standard programs for statistical data analysis which provide interface to the samplers that estimate the distributions of model parameters. The MCMC methods belong to the Bayesian statistical inference framework and therefore the standard statistical inference with p-values and confidence intervals cannot be used. In addition, the MCMC methods in general require more computational resources than the standard methods as well as lots of checking of the simulation results for problems with convergence and other computational issues.



## Chapter 5

### Comparison of methods in NMR spectroscopy

The metabolomics [108–109] has become an established approach for qualitative and quantitative description of all metabolites present in the biological sample under study. The detailed knowledge of the metabolic status of an organism seems to be beneficial, especially in the biomedical research providing a better understanding of the molecular background of a disease and its early diagnosis [110]. The metabolomics is a part of *simomics* approaches. The metabolomics is a successor of genomics, transcriptomics and proteomics [111]. The metabolomics deals with problems which are far more complex in comparison with other *-omics* sciences. There are two main approaches used in metabolomics - metabolomic profiling and metabolomic fingerprinting. The metabolomic profiling aims at the identification and quantification of a pre-selected set of metabolites. The metabolomic fingerprinting is an unbiased and global analysis of all metabolites in a sample [25]. The nuclear magnetic resonance (NMR) spectroscopy [112–113] is one of the most prominent analytical platforms in metabolomics, especially in the fingerprinting experiments [114–115].

The complexity of metabolomic data calls for sophisticated methods of the evaluation of the experiments. Many works present and review the basic concepts of data analysis, as the best reference may serve [12]. The most prevalent method for data analysis in NMR-based metabolomics became the partial least squares (PLS) regression [33]. The understanding of the parameters estimated by multivariate model is challenging. The analyses of metabolomic data tend to suffer from small sample sizes and a large number of variables, which results in low power of statistical tests, the ability to discover the actual relationships in the data. We cannot easily devise the statistical significance of parameters of multivariate models, so the usual strategy is to utilise non-parametric approaches such as resampling techniques. Even the use of resampling techniques, e.g. bootstrap and permutation tests, cannot be applied directly without the consideration of various properties of the data and the model. That is why the researchers usually do not attempt to perform the statistical evaluation of the multivariate model and use more heuristic approaches. We propose a procedure which allows testing of statistical properties of multivariate models and other methods in NMR-based metabolomic data analyses to provide reliable results.

The problem for the practitioner is often, how to decide, which data analysis method to use. Each proposition of a new method in the analysis of the NMR spectroscopy data should present a comparison to the established method, so the potential user knows all the benefits and drawbacks. Typically, the new methods are, after careful theoretical examination of their properties, tested on simulated data or benchmark data recognised by the community. In the case of benchmark data, the main problem is the limited number of trials, which they offer for the actual testing of the methods. The limited number of trials stems from the fact that these data are usually results of experiments, and we cannot repeat the experiments indefinitely. The small data sets may lead to overspecialisation of the methods for the given data and spectacular failures in data with different properties. Another problem poses the lack of accurate and correct re-



---

The original  $n \times m$  matrix of predictors  $X$  of  $n$  observations and  $m$  predictors is related to the  $n \times p$  matrix of  $n$  observations and  $p$  responses  $Y$  by the loadings vectors in  $p \times m$  matrix  $U$ . The  $n \times p$  matrix  $E$  contains the residuals. The matrix  $U$  is estimated by singular decomposition (SVD) of matrix constructed from matrices  $X$  and  $Y$ , for detailed description of the estimation of matrix  $U$  refer to [76].

The main difference between the two methods is in the estimation procedure of the model parameters. The iterative procedure used in PLS algorithm does not allow to fully exploit the joint covariance in the  $X$  and  $Y$  matrices, because it does not allow subtraction of the estimated latent variables from both matrices simultaneously. Due to the representation of the matrices  $X$  and  $Y$  in one combined matrix and use of the SVD, the SPCA subtracts the latent variables simultaneously in both matrices. The SPCA allows for detection of non-linear dependencies, due to the use of the stronger notion of statistical dependency rather than the correlation used in PLS. The second difference concerning classification problem is the rank deficiency of the PLS method. In classification into  $c$  classes, the response matrix  $Y$  consists of  $c - 1$  vectors of dummy variables. In PLS the rank deficiency does not allow for estimation of more than  $c - 1$  loadings, which has severe consequences for the classification performance.

### ■ 5.0.3 Model validation

The bilinear models have parameters whose values have to be estimated. In the case of both PLS and SPCA, the parameter is the number of estimated latent variables. In case of SPCA, the problem with validation of model parameters is in non-linear analyses exploiting kernel transformations and concerns the parameters of kernels. The standard approach for parameter estimation is the cross-validation procedure [121, 87]. The problem, which usually arises with metabolomic data is that there is not enough data to form distinctive sets for training, validation and testing of the model. In this case, the  $N$ -fold cross-validation is suitable for the estimation of the parameter [87]. In situations where the number of instances of data is not sufficient, only the leave-one-out validation remains [26]. Both of these methods may suffer from over-fitting. Various techniques were suggested to attribute for the over-optimistic results of these techniques [26].

Possibly the best solution is to estimate the variation of the values of performance criterion by permutation or bootstrap test. These tests may significantly decrease the uncertainty about the value of the criterion. This non-parametric test estimates the distribution of criterion on permuted data, which then serves for the examination of the case of no relationship between the groups – the null distribution. Consequently, this distribution serves as a reference for the actual value of criterion measured on the original data. The advantage is that these tests are simple, but may be computationally intensive [87, 122].

The criterion of the model performance serves to assess the quality of the model from various perspectives. We use a different set of criteria for classification models and regression models. In the classification setting, we usually use the model performance measured by classification accuracy or classification error, sensitivity, specificity and F1 score [123]. In the regression setting, we use the coefficient of determination, the percentage of explained variation, and the goodness of fit for the model quality assessment [124].

### ■ 5.0.4 Model interpretation and assessment

The logical conclusion of all the modelling effort is the interpretation and assessment of the model. The interpretation of the model goes hand in hand with the understanding

of the model parameters. The least informative for this purpose are the scores, which are the estimate of the outcome; however, they may be useful for the detection of outlying observations. The loading matrices of the model carry the most important information for the model interpretation. The value of the loadings is directly related to the importance of the variables in the high dimensional space of NMR data to the resulting model. The greater the absolute loading value, the greater the importance of the variable in the model, assuming some scaling procedure was utilised before the model training. However, this does not provide any clues about the relevance of the contribution and significance of the variable. The much needed decisive indicator for loading values provides the **variable importance in projection score (VIP score)** [125]. We usually consider the VIP score of one as a threshold value for selecting variable as significant. The threshold value may vary from 0.83 to 1.21 to yield more relevant results. If we employ the statistic evaluation of the model, the most prominent approach is the simple bootstrap. The main problem with the bootstrap approach is that the null distribution of the loading value is (due to the PLS regularisation) affected by the remaining values of the given loading. We have to control the uncertainty in the sample to a higher degree to limit these unwanted effects. Following the work [126] the stability of the loading's distribution is most effectively estimated by permutation tests for each loading's value examined separately. Using the re-sampling methods in the context of multivariate models such as ordinary PCA affect few issues. We have to pay attention to the direction of loadings and the order of the principal components. These problems do not concern the supervised methods such as PLS and SPCA, where the outcome ensures the same orientation and order of all the loadings.

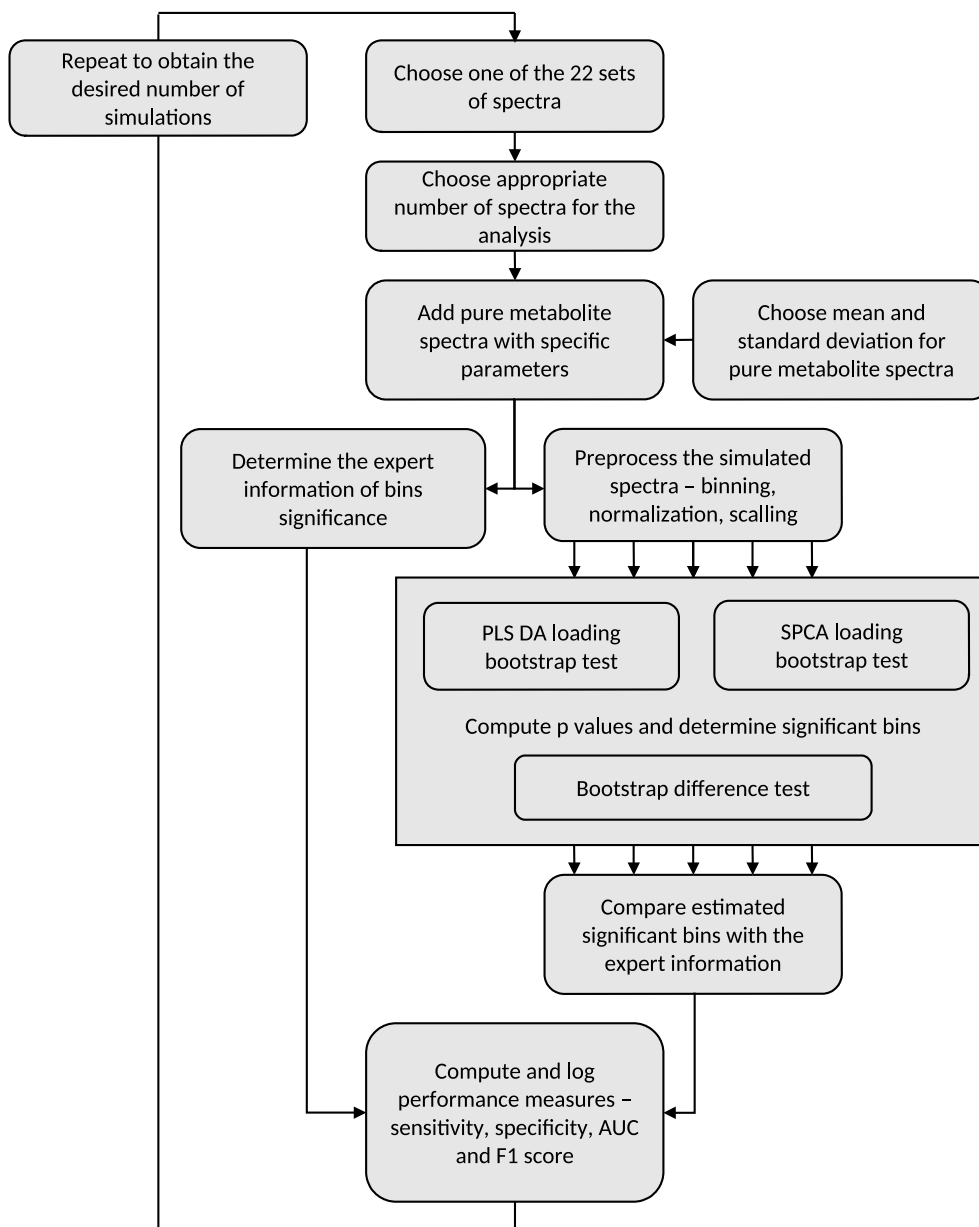
## 5.1 Methods

### 5.1.1 Hypotheses formulation

The authors of the SPCA claim and theoretically show that the method outperforms the PLS. The simultaneous subtraction of variance in SPCA is a benefit over the PLS algorithm in the form of model parsimony. The SPCA can handle classification into more than two classes problems better. These differences can have practical consequences for experimenters. In order to compare the methods, we devised a simulation scheme allowing testing of various hypotheses regarding properties of the algorithms. We decided to follow the recommended procedure for the analysis [121] of data in NMR-based metabolomics and test the methods by criteria related to the typical problems in the analysis of NMR spectra. The process consists of analyses of simulated datasets, where known signals of simpler compounds alter a set of template signals. The identification of added compounds serves for estimation of performance measures. By repeating this procedure, we obtain the distribution of the performance measures and use them for the comparison.

### 5.1.2 Description of data

To compare the PLS and the SPCA, we used a set of NMR spectra of mouse urine samples collected during various experiments. We divide the spectra into subgroups according to the experiment settings (various groups of animals differing in treatment) and NMR experiments applied for proton spectra acquisition (1D-NOESY, CPMG).



**Figure 5.1.** The flowchart describing the treatment of the NMR spectra, generation of simulated data and statistical procedures used for methods' comparison.

### 5.1.3 Alteration by known signals

To simulate the specific characteristics of NMR spectra of urine in metabolomic study we chose a set of metabolites based on previous results in [J3] and in one other independent work [127]. We added the proton spectra of pure metabolites from Birmingham Metabolite library [128] to the selected spectra from our set, to mimic the usual process of the NMR spectra generation and analysis. For the list of used metabolites refer to Tab. 5.1. We divided the selected spectra into treatment groups for the purposes of the methods comparison. We altered the spectra following the formula:

$$S^{alt}(f) = S(f) \sum_{i=1}^N c_{i,j} \cdot S_i^{met}(f - \delta) \quad (5.4)$$

Meaning that we shift the metabolite spectra  $S_i^{met}(f - \delta)$  in position by  $\delta \sim N(0, \sigma_{shift}^2)$  to simulate the positional shift, we estimated the  $\sigma_{shift}^2$  parameter from the data. The metabolite spectrum  $S_i^{met}$  added to the spectrum to be altered  $S(f)$  constitutes the altered spectrum  $S^{alt}$ . The parameter of addition simulating concentration is  $c_{i,j} \sim N(\mu_{i,j}, \sigma^2)$ , randomly distributed numbers from normal distribution with the mean  $\mu_{i,j}$  and standard deviation  $\sigma^2$ . The mean  $\mu_{i,j}$  of the addition parameter  $c_{i,j}$  was specific for each treatment group, indicated by the index  $j$ , and the metabolite, indicated by the index  $i$ , to simulate the differences among groups. Moreover, to make the comparison more credible, we chose the values of means  $\mu_{i,j}$  independently for each iteration of the comparison. The distribution of the means  $\mu_{i,j}$  was a gamma distribution  $\Gamma(a, b)$ , because the values of gamma distribution are non-negative, which also applies for the NMR spectra. We estimated the parameters  $a$  and  $b$  of the Gamma distribution from the original unaltered spectra. In the comparison, we kept the actual addition parameters of pure metabolites spectra sampled from distribution  $c_{i,j} \sim N(\mu_{i,j}, \sigma^2)$  and we subjected the values to an appropriate test estimating the differences between the sampled  $c_{i,j}$  by the class membership indicated by  $j$ . If the sampled addition parameters significantly differed in this comparison, we considered the metabolite as significant and its peaks frequencies used for methods comparison as a reference in the classification task. This allowed the construction of the confusion matrix.

### 5.1.4 NMR spectra processing

The altered spectra from were uniformly binned into intervals of the width of 0.04 ppm. We excluded the parts of spectra exceeding the interval -1 to 10 ppm and parts related to water, urea and reference compound (the intervals 4.6 - 4.9, 5.6 - 5.97, -0.12 - 0.12 ppm). We normalised the spectra by probabilistic quotient normalisation method [129].

### 5.1.5 Methods comparison

We decided to establish our comparison on criteria describing the confusion matrix of a classification problem of metabolites according to their typical frequencies. We chose this approach as a representative of the metabolomic fingerprinting. The criteria related to the confusion matrix were the measures of classification performance. We used sensitivity, specificity, the area under receiver operating characteristics (AUC), and F1 score as the tested criteria. These measures are widely known, but for reference can be used [123]. The distributions of these measures were estimated in simulation repeating individual iterations with random parameters.

### 5.1.6 Single iteration of the comparison procedure

For each iteration, we chose a set of spectra of appropriate size according to the structure of the data. We divided the selected spectra into groups (two, three and five according to simulation setting) and altered by known signals in the way described in the section 5.1.3 and processed by standard pipeline following the work of [12]. We performed the Student's t-test on the parameters altering the template spectra and corrected by the Bonferroni correction for multiple comparison [93] to obtain the reference information about significant frequency bins for the construction of confusion matrix. We considered the bins of the spectrum, which contained the significant frequencies of chosen metabolites acquired from the Human Metabolome Database [130] as true positives. We denoted all the other bins as true negatives.

Metabolite name	Frequency [ppm]	Metabolite name	Frequency [ppm]	Metabolite name	Frequency [ppm]
Acetate	1.90	Histidine	3.14	Malate	2.34
			3.18		2.38
Acetoacetate	2.26		3.22		2.66
	3.42		3.26		2.70
			3.98		4.30
			7.10		
Alanine	3.78		7.90	Methylamine	2.58
	3.74				
	1.46	Isoleucine	0.90	Ornithine	1.70
			0.94		1.74
Citrate	2.66		0.98		1.78
	2.54		1.02		1.82
	2.50		1.22		1.86
			1.26		1.90
			1.30		1.94
Creatine	3.94		1.42		3.02
	3.02		1.46		3.06
			1.50		3.78
			1.94	Oxoglutarate	2.42
Creatinine	3.02		1.98		2.46
	4.06				2.98
					3.02
		Isoleucine	1.98		
Dimethylamine	2.50		2.02	Succinate	2.38
			3.66		
Fumarate	6.50			Taurine	3.26
					3.42
		Lactate	1.30		
Glutamine	2.10		1.34	Trimethylamine	2.90
	2.14		4.10		
	2.18		4.14	Tryptophan	3.26
	2.38				3.30
	2.42				3.46
					3.50
Glycine	3.54	Leucine	0.94		4.02
			0.98		4.06
Hippurate	3.94		1.62		7.18
	3.98		1.66		7.22
	7.54		1.70		7.26
	7.62		1.74		7.30
	7.66		1.78		7.54
	7.82		3.70		7.70
	7.82		3.74		7.74

**Table 5.1.** Metabolites used in the methods comparison.

We processed the data by the PLS and the SPCA methods. For both of these methods, we mean centered and Pareto scaled [131] the data. We tested the significance of the results by the procedure proposed by [126]. This procedure effectively tests each variable (frequency bin) by permutation test separately and without affecting the remaining variables. This procedure thus deals with many issues emerging in rigorous testing of variable significance in models such as PCA, PLS and SPCA. The main benefit is keeping the variance-covariance structure of the data unchanged during the re-sampling process and in stabilising the variance of the tested variable by eliminating the influence of other variables. However, this procedure compared to simple bootstrap

methods needs many times more (exactly the number of variables times more) permutations and repeated the model estimation. The problem gets severe when the correction for multiple comparisons comes into play. The proposed procedure would be infeasible, considering the typical testing procedure using  $p$ -values. The number of tests in the bilinear model in the metabolomics is usually in the order of hundreds. The number of tests with the correction for multiple comparisons lowers the threshold value to such small numbers, that the effective number of permutations to even reach an appropriate resolution in  $p$ -values would have to be in the order of tens or hundreds of thousands. We may try various strategies to decrease the number of permutations. One of them is to test only variables, which show some tendency of being substantially different. We may base this pre-pruning of analysis on an estimate of the true  $p$ -value by a similar test, which may not be the most appropriate, but measures similar phenomenon. In our study, we used the Student's  $t$ -test with no correction for multiple comparisons. The resulting  $p$ -values compared to the threshold value of 0.05 indicated variables for rigorous testing, the remaining remained marked as non-significant.

The preceding procedure helps reduce the number of permutations, though it does not affect the threshold value. We decided to decrease the number of computations by using the relationship between  $p$ -value and confidence intervals. Comparing of  $p$ -value to a predefined threshold value is equivalent to comparing the estimated value of tested statistics to a critical value. In many cases, the critical value coincides with the limits of confidence intervals. The estimation of the confidence interval may be simpler than the exact computation of  $p$ -value. We decided to approximate the critical values for a test by BCa method for estimation of confidence intervals [132] at an appropriate confidence level corrected by Bonferroni correction. The advantages of the BCa method is that the confidence interval can work with non-central distributions, is bias-corrected and thus can provide the confidence interval for skewed distributions and the parameters of the approximations need only thousands of simulations to estimate. Testing the values of loading with confidence intervals means that we declare the value of the loading coefficient as significant in the case that the estimated value of the coefficient lies outside the confidence interval.

The significant frequencies were compared to the reference information and eventually arranged into confusion matrices. We used the confusion matrices for the estimation of sensitivity, specificity, AUC and F1 score.

### 5.1.7 Evaluation of results from repeated comparisons

We repeated the comparison of methods on simulated data, and we kept the results for the ultimate comparison. We do not know the properties of the distributions of the measures for the hypotheses testing, and thus we cannot suppose it would have some of the nice properties such as gaussianity. This directs us to non-parametric statistical tests. The nature of the testing procedure allows for a paired test and thus, the most appropriate statistical test is the Wilcoxon signed-rank test. In all the measures of the performance of the methods, the higher the value, the better the outcome. That means we can test hypotheses with a one-sided hypothesis test. We corrected the tests for multiple comparisons with the Bonferroni correction.

## 5.2 Results



### 5.2.1 Data description

We collected and processed the data using the Matlab Language for technical computing program. We visualised the individual spectra and checked them for anomalies. No missing or outlying observations were present. The final set consisted of 252 spectra divided into 22 subgroups.

We used the whole set of spectra to estimate the parameters for the spectra alteration process – the parameters of the distribution – by a Matlab statistics toolbox procedure [133].

	Sample size	SPCA Mean (SE)	PLS Mean (SE)	SPCA - PLS Mean (SE)	<i>P</i> -value
2 groups comp.	10	0.064 (0.003)	0.170 (0.004)	-0.106 (0.004)	1.0000
	<b>20</b>	<b>0.205 (0.004)</b>	<b>0.198 (0.004)</b>	<b>0.007 (0.002)</b>	<b>0.0003</b>
	<b>40</b>	<b>0.330 (0.004)</b>	<b>0.328 (0.004)</b>	<b>0.002 (0.001)</b>	<b>0.0009</b>
	100	0.442 (0.005)	0.442 (0.005)	0.000 (0.000)	0.1250
3 groups comp.	<b>15</b>	<b>0.114 (0.004)</b>	<b>0.102 (0.003)</b>	<b>0.013 (0.003)</b>	<b>0.0008</b>
	30	0.284 (0.004)	0.281 (0.004)	0.003 (0.002)	0.0297
	60	0.415 (0.005)	0.414 (0.004)	0.001 (0.001)	0.2094
	150	0.492 (0.006)	0.49 (0.006)	-0.001 (0.001)	0.6444
5 groups comp.	<b>25</b>	<b>0.048 (0.002)</b>	<b>0.040 (0.001)</b>	<b>0.008 (0.001)</b>	<b>0.0000</b>
	<b>50</b>	<b>0.264 (0.001)</b>	<b>0.244 (0.001)</b>	<b>0.020 (0.001)</b>	<b>0.0000</b>
	<b>100</b>	<b>0.311 (0.001)</b>	<b>0.301 (0.001)</b>	<b>0.010 (0.000)</b>	<b>0.0000</b>
	250	0.345 (0.000)	0.344 (0.001)	0.001 (0.000)	0.9909

**Table 5.2.** The F1 score on simulated sets of NMR data divided into two, three and five groups presented as means and standard errors of the means by experiment settings. Bold-face indicates the significant differences according to the Wilcoxon signed-rank test – *p*-value less than 0.0010.

### 5.2.2 Simulations

In order to compare the methods, 12 runs of repeated simulations were carried out. We aimed the comparison at the differences between SPCA and PLS in biomarker discovery problem setting. In the runs of the simulation, we varied two parameters. We repeated the runs of simulations with 5, 10, 20 and 50 observations for each treatment group. The second varied parameter was the number of outcome variables. The outcome variable for the biomarker discovery (classification) problem is a binary indicator. The number of outcome variables was chosen to simulate classification into two, three and five groups. For each run of simulations, the number of simulations was 500. In each simulation, we tested the significance of variables in loadings in SPCA and PLS models by univariate bootstrap performing 5000 iterations to construct confidence interval by the BCa method. We compared the actual model parameters to the confidence interval corrected for the multiple comparisons and stored the significant variables. We subsequently compared the significant variables to the reference and computed and stored the performance measures. The resulting performance measures were used to compare the method.

The results of the comparison were summarised in form of tables and are presented in complete form in Tables 5.2, 5.3, 5.4 and 5.5. Examining the results, we may observe few phenomena, some of them obvious, other more interesting. The general tendencies are that the bigger the sample size, the better the results of the methods in identifying the correct variables measured by the performance measures. The explanation for this observation is obvious. The more data is available for the training of the model, the better estimates of the models' parameters and smaller standard errors.

	Sample size	SPCA Mean (SE)	PLS Mean (SE)	SPCA - PLS Mean (SE)	<i>P</i> -value
2 groups comp.	10	0.513 (0.001)	0.542 (0.002)	-0.029 (0.001)	1.0000
	20	0.551 (0.002)	0.549 (0.002)	0.002 (0.001)	0.0014
	<b>40</b>	<b>0.591 (0.002)</b>	<b>0.590 (0.002)</b>	<b>0.001 (0.000)</b>	<b>0.0010</b>
	100	0.630 (0.005)	0.630 (0.005)	0.000 (0.000)	0.1250
3 groups comp.	15	0.524 (0.002)	0.522 (0.001)	0.003 (0.001)	0.0177
	30	0.574 (0.002)	0.573 (0.002)	0.001 (0.001)	0.0280
	60	0.616 (0.004)	0.616 (0.004)	-0.000 (0.001)	0.3004
	150	0.650 (0.005)	0.651 (0.005)	-0.000 (0.001)	0.6613
5 groups comp.	<b>25</b>	<b>0.300 (0.004)</b>	<b>0.269 (0.004)</b>	<b>0.031 (0.003)</b>	<b>0.0000</b>
	<b>50</b>	<b>0.352 (0.001)</b>	<b>0.325 (0.001)</b>	<b>0.027 (0.001)</b>	<b>0.0000</b>
	<b>100</b>	<b>0.413 (0.001)</b>	<b>0.399 (0.001)</b>	<b>0.015 (0.001)</b>	<b>0.0000</b>
	250	0.462 (0.001)	0.461 (0.001)	0.001 (0.000)	0.9909

**Table 5.3.** The area under receiver operating characteristics (AUC) score on simulated sets of NMR data divided into two, three and five groups presented as means and standard errors of the means by experiment settings. Boldface indicates the significant differences according to the Wilcoxon signed-rank test – *p*-value less than 0.0010.

Similarly, the bigger the sample size, the smaller the differences between the compared models. Here the explanation is the same as in the previous case. As the differences get small, we would need many more iterations to prove the difference between the methods.

The third phenomenon is the possibility to estimate a model with an ever-smaller number of observations. This problem originates from the computational characteristics of the algorithms. The PLS algorithm devised for ill-conditioned problems can deal even with the smallest data sets in the comparison. However, the smallest presented set seems to be troublesome for the SPCA, see Tab. 5.2, 5.5. It seems that with the very small sample sizes, the PLS method degenerates to a simple test that does not account for the covariance in the data.

On the other hand, the SPCA utilises the covariance even with this very small sample sizes and the differences are indistinguishable from the other variation in the data. Another explanation for this observation may be the better training abilities of the SPCA. The model can learn the training examples in such a detail that it completely fails on new instances. Projecting this idea into univariate permutation test of loading coefficients, the variability of the loading coefficients is due to the over-fitting very high. High variation of the coefficient results in a very wide confidence interval for the hypothesis testing and poor performance of the SPCA observed on small sets of data (considering classification problems).

### 5.3 Discussion

To sum up the results, the SPCA is better than the PLS method in almost every comparison. The difference is not very high; it depends on the setting of the run of simulations. The most pronounced differences are in the runs with a multinomial response – classification into five classes, which supports the claims of the authors of the SPCA method [76] about the rank deficiency of PLS method. The results show that the SPCA method used for classification has higher sensitivity than the PLS on the same data sets. Using the Wilcoxon signed-rank test, we were able to show that the SPCA is better than the PLS in the classification of samples into more than two classes. The PLS is not capable of using more loadings than the number of classes, and for this reason, it has very limited capabilities of solving classification into more than

	Sample size	SPCA Mean (SE)	PLS Mean (SE)	SPCA - PLS Mean (SE)	<i>P</i> -value
2 groups comp.	10	0.037 (0.002)	0.113 (0.003)	-0.076 (0.003)	1.0000
	20	0.130 (0.003)	0.124 (0.003)	0.006 (0.001)	0.0013
	40	0.241 (0.003)	0.239 (0.003)	0.002 (0.001)	0.0078
	100	0.405 (0.004)	0.405 (0.004)	0.000 (0.000)	1.0000
3 groups comp.	<b>15</b>	<b>0.071 (0.002)</b>	<b>0.062 (0.002)</b>	<b>0.009 (0.002)</b>	<b>0.0000</b>
	30	0.203 (0.003)	0.200 (0.003)	0.003 (0.001)	0.0021
	60	0.359 (0.003)	0.356 (0.003)	0.003 (0.001)	0.0063
5 groups comp.	150	0.556 (0.003)	0.556 (0.003)	0.001 (0.001)	0.1747
	<b>25</b>	<b>0.373 (0.008)</b>	<b>0.306 (0.008)</b>	<b>0.067 (0.005)</b>	<b>0.0000</b>
5 groups comp.	<b>50</b>	<b>0.642 (0.002)</b>	<b>0.585 (0.002)</b>	<b>0.057 (0.002)</b>	<b>0.0000</b>
	<b>100</b>	<b>0.796 (0.002)</b>	<b>0.765 (0.002)</b>	<b>0.031 (0.001)</b>	<b>0.0000</b>
	250	0.911 (0.001)	0.908 (0.001)	0.003 (0.001)	0.0591

**Table 5.4.** The sensitivity score on simulated sets of NMR data divided into two, three and five groups presented as means and standard errors of the means by experiment settings. Boldface indicates the significant differences according to the Wilcoxon signed-rank test – *p*-value less than 0.0010.

two classes problems.

We proposed a method to control the process of comparison that represents an actual metabolomic fingerprinting problem and its correct statistical evaluation. To vindicate the first proposition, we used a process whose parameters are not affected by artificial effects and biases. By choosing the parameters of metabolite spectra to alter the analysed signals randomly, we eliminated confirmation bias (choosing such sets of parameters that yield the desired outcome). The high number of repetitions of simulations experiments allowed us to construct the distributions of the performance measures, which provides more information than point estimates. The possibility to repeat the experiments *in silico* is one of the main advantages of the proposed method. In the laboratory with actual equipment, we cannot repeat an experiment several thousand times to estimate variability in results and would be a complete waste of resources, although it might provide valuable data for method comparison. The use of proper statistical methods and appropriate corrections for multiple comparisons ensures the second proposition of statistical correctness. Certain criticism may consider the choice the reference testing, which does not account for the variability connected with the sampling of the template spectra.

On the other hand, the rigorous statistical approach reflects the current problems of metabolomics. The sample size in metabolomic fingerprinting has to be definitively higher, considering the number of performed tests. The utilisation of corrections for multiple testing helps prune the results and prevent publication of insufficiently founded or erroneous results. The proposed procedure for model comparison proved to be a reliable framework.

Another problem poses the choice of the measures by which to compare the methods. The comparison can be exhaustive in the choice of the measures [134]. However long lists of measures may not be understandable for the potential users. We decided to use a simple and understandable measure for the comparison, which is the F1 score. The F1 score (as well as the AUC) combines the classification errors from all the classes. This helps to avoid biased resulting from shifting a threshold for a classification – the case where an increase in threshold may designate more instances as a correct class, but also increasing the errors in assignments to other classes. In other words, we are comparing the overall performance of the model without the need to check whether we

	Sample size	SPCA Mean (SE)	PLS Mean (SE)	SPCA - PLS Mean (SE)	<i>P</i> -value
2 groups comp.	<b>10</b>	<b>0.989 (0.001)</b>	<b>0.971 (0.002)</b>	<b>0.018 (0.001)</b>	<b>0.0000</b>
	20	0.972 (0.003)	0.974 (0.002)	-0.001 (0.001)	0.9945
	40	0.940 (0.004)	0.940 (0.004)	-0.000 (0.000)	0.9453
	100	0.854 (0.008)	0.854 (0.008)	0.000 (0.000)	0.1250
3 groups comp.	15	0.977 (0.002)	0.981 (0.002)	-0.004 (0.001)	1.0000
	30	0.944 (0.004)	0.945 (0.004)	-0.001 (0.001)	0.9508
	60	0.874 (0.007)	0.876 (0.007)	-0.003 (0.001)	0.9993
	150	0.744 (0.011)	0.745 (0.011)	-0.001 (0.001)	0.9936
5 groups comp.	25	0.228 (0.001)	0.232 (0.001)	-0.005 (0.001)	1.0000
	50	0.061 (0.001)	0.065 (0.001)	-0.004 (0.000)	1.0000
	100	0.030 (0.000)	0.032 (0.000)	-0.002 (0.000)	1.0000
	250	0.013 (0.000)	0.014 (0.000)	-0.001 (0.000)	1.0000

**Table 5.5.** The specificity score on simulated sets of NMR data divided into two, three and five groups presented as means and standard errors of the means by experiment settings. Bold face indicates the significant differences according to the Wilcoxon signed-rank test – *p*-value less than 0.0010.

are only increasing sensitivity at the expense of specificity. The presented sensitivities and specificities only complement the information given by the F1 score and AUC.

## 5.4 Conclusion

We examined the typical approaches to testing and comparing methods of data analysis in NMR-based metabolomics. By **combining the usual testing on simulated and real-world data**, we devised a simple method for extensive testing and comparing of data analysis approaches. The **main benefit is in testing methods in the wide variety of conditions provided by the experimental data** with control over the correct results. The method sufficiently simulates the process of NMR spectra generation and limits the overfitting of the methods to small data set by repeated application to altered datasets. We applied the proposed method on a comparison of methods in metabolomics fingerprinting. We compared the established partial least squares discriminant analysis and the supervised principal component analysis. The SPCA is a new method previously not widely used in the field of NMR-based metabolomics. The results of the comparison clearly show that **the SPCA is better for the assessment of metabolomic fingerprinting**. The differences are the most pronounced in difficult tasks such as classification of individuals into more than two groups.

We were able to show that the actual difference is not in the threshold for classification, which could be a problem if we used simple criteria such as classification accuracy or error. We avoided this problem by examining not only simple classification error but the F1 score, the AUC, the sensitivity and the specificity. We may thus conclude that the SPCA is better than the established PLS method and only the lack of implementations in standard programs for data analysis hinders the use of SPCA in the field of NMR-based metabolomics.

# Chapter 6

## Element mapping

### 6.1 Introduction

We have already presented the idea of the elemental mapping by the LA-ICP-MS method in the introductory chapter. In this chapter, we will focus on a more detailed description of the problems which we need to deal with when analysing the data originating from the LA-ICP-MS. We will use a data collected from an animal study of the spatial distribution of bioactive metals in biological samples of pigs' melanoma alongside some other results from the examination of the variability of the data and their structure. First, we will describe the data collection procedure as well as the resulting data. Second, we will present the statistical properties of the data from various sources, and we will devise a data integration procedure. Third, we will present results from several case studies. This chapter summarises the work presented in the articles [J1, C1], as well as some unpublished work.

### 6.2 Data collection

We can obtain useful information about biological tissue samples through various methods. The standard approaches either use different dyes to colour sample sections and analyse them using microscopy or the chemical constitution of the biological samples is analysed. These two approaches differ mainly in the treatment of the sample. The microscopy analysis is non-destructive. The aim is to prevent damaging the structures in the sample section as much as possible. On the other hand, the analyses of the chemical constitution are destructive. In order to measure the composition, the measurement device liquidises, pulverises or, for example, evaporates the biological sample. Recently developed methods of biological sample analysis allow for the simultaneous analysis of both the spatial structure and the chemical composition.

#### 6.2.1 Histology

The MeLiM (Melanoma-bearing Libechov Minipig) strain of miniature pigs with heritable cutaneous melanoma is an original animal cancer model with histopathological, biochemical and molecular biological similarities to human melanoma [135–138]. Multiple skin melanomas appear at birth or shortly after in approximately half of all piglets. More than 2/3 of the affected minipigs display complete spontaneous regression of tumours, which is usually accompanied by skin and bristle depigmentation. After a short postnatal period of tumour growth, the first signs of spontaneous regression, both macroscopic (flattening and grey colour of tumours) and microscopic (gradual destruction of melanoma cells, reduced expression of collagen IV and laminin, and rebuilding of tumour tissue into fibrous tissue), are observed. Ten weeks of age appears to be a turning point in the transition between tumour growth and spontaneous regression in MeLiM melanoma [139]. The incidence of spontaneous regression of melanoma is

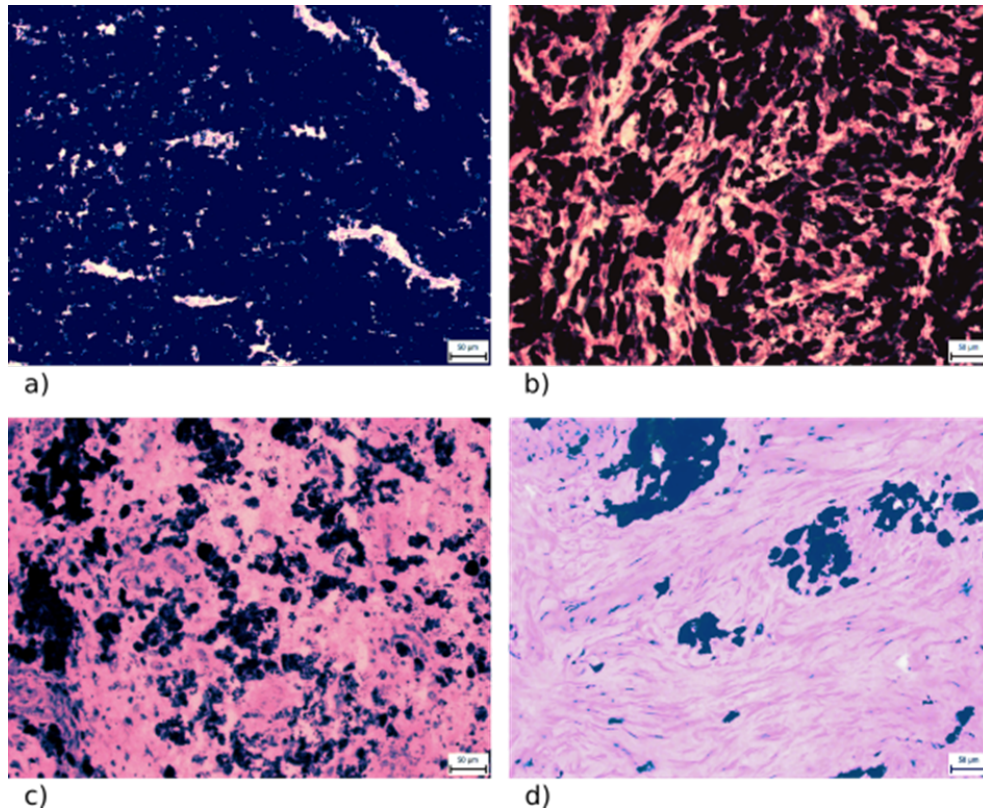
high in the MeLiM model. The immune system destroys the melanoma cells after a short postnatal period of tumour growth, and the healing processes turn the tumour tissue into fibrous tissue. In connection with this process, four structurally different zones were distinguished in the histological samples and marked with various colours Fig. 2.7: GMT (the zone of normally growing melanoma tissue – red rectangles), ESR (the zone of early spontaneous regression – violet rectangles), LSR (the zone of late spontaneous regression – yellow rectangles) and FT (the zone of fibrous tissue – green rectangles). In Fig. 6.1A-D we provide a detailed histological view of the zones and their description. Particularly, using haematoxylin-eosin staining, four histologically different zones were distinguished in the collected melanoma samples Fig. 6.1A-D:

- The zone of normally growing melanoma tissue (GMT) was composed of heavily pigmented, intact melanoma cells, which were distributed close together with narrow extracellular spaces (Fig. 6.1A).
- The zone of early melanoma cell destruction (early spontaneous regression, ESR) included cellular debris from some of the damaged melanoma cells, but a considerable number of melanoma cells were still well preserved (Fig. 6.1B).
- The zone of late melanoma cell destruction (late spontaneous regression, LSR) characterised by extensive damage to the melanoma tissue (forming predominantly cellular debris with small groups or individually dispersed melanoma cells) and its incipient rebuilding in the fibrous tissue (Fig. 6.1C).
- The zone of fibrous tissue (FT) arising by the total rebuilding of tumour tissue. A small number of remaining melanoma cells were occasionally still present (Fig. 6.1D).

We can observe age-dependent changes in melanoma structure. In the melanoma of the youngest (4-week-old) animals, zones of normally growing melanoma tissue were distinctly prevalent compared with zones of early melanoma cell destruction. The other two zones were entirely missing. The number and size of the GMT zones decreased with age, whereas we can observe the opposite tendency in the zones of ESR and late melanoma cell destruction (the latter appeared in 6-week-old animals). Fibrous tissue, which we first observed in the 15-week-old minipigs, gradually replaced the damaged tumour tissue. In the melanoma of the oldest animals (22 weeks old), zones of late destruction of melanoma cells were most prevalent, and fibrous tissue occupied the damaged tumour tissue. In the melanoma of the oldest animals (22 weeks old), zones of late destruction of melanoma cells were most prevalent, and fibrous tissue occupied the areas between the zones. In these minipigs as well as in the 15-week-old minipigs, we no longer observed the zones of GMT. Selected zones were matched with the elemental map (of the neighbouring cryosection) as provided by laser ablation to compare Zn and Cu content in melanoma during melanoma growth and successive stages of spontaneous regression.

### ■ 6.2.2 LA-ICP-MS

Laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) enables the measurement of the metal content in a selected zone ranging from one to several hundred micrometres on the sample surface. LA-ICP-MS provides an ideal, rich source of information because it can match each ablation pixel (the smallest part of the studied sample that can be distinguished using ablation) to the relevant quantified information about the presence of most chemical elements. Laser ablation parameters, such as laser beam fluence, laser spot size, and scan speed rate, determine the time necessary for the analysis of any sample as well as the accuracy of the obtained results. These parameters were optimised to ensure the required performance, namely a low limit of



#### Histologically differing tissue zones

**Figure 6.1.** Four histologically differing zones identified in haematoxylin-eosin stained skin porcine melanoma: (A) growing melanoma tissue (GMT), (B) melanoma tissue with early destruction of melanoma cells (early spontaneous regression-ESR), (C) melanoma tissue with late destruction of melanoma cells (late spontaneous regression-LSR), (D) fibrous tissue (FT) with a few remaining melanoma cells. Scale bar is 50  $\mu\text{m}$ . [J1]

detection (LOD) and low broadening of images within a reasonable time of analysis. This optimisation was performed by ablation of white paper with printed ink lines of 800  $\mu\text{m}$  thickness. For this purpose, we recorded the  $^{63}\text{Cu}$  signal because this element is present in the used ink [28]. Elemental mapping was performed using line scan mode so that each line started on a glass substrate outside the tumour tissue. The laser beam was moved on the sample surface continuously along a straight line with a constant scan rate of 200  $\mu\text{m}/\text{s}$ . The laser beam diameter and the distance between individual straight lines were both 100  $\mu\text{m}$ . The optimisation of the laser beam fluence and the repetition rate resulted in the respective values of 8  $\text{J}/\text{cm}^2$  and 20 Hz; We used these values for all LA-ICP-MS analyses. The high laser beam fluence prevented the influence of different ablation rates and was optimised to reach the glass substrate during laser ablation.

We based the quantification on the calibration performed using agarose gel standards prepared by spiking with known amounts of  $\text{Cu}$  and  $\text{Zn}$ , for  $\text{Cu}$  see in Fig. 6.2A. The prepared calibration standards contained single metal content of 0, 20, 100, 500 and 2000 mg/kg. Each standard was ablated in triplicate using the same ablation parameters used for imaging. We performed the background correction by subtraction of the average signal obtained using a carrier gas blank ( $\text{He}$ ).

In imaging by means of LA-ICP-MS, we can evaluate the broadening of imaged patterns

as described previously [28, 140–141]. The broadening is mainly due to a combination of the laser spot size and scan speed rate. Hence, we have to adjust carefully these parameters concerning the size of the treated samples (due to time of analysis) and the size of the zones of interest (due to the trueness of imaging).

Histologically different zones in minipig tumour tissues were well-defined areas inside the analysed tissue samples with a size of several hundred micrometres in each dimension. Eight scan speed rates were used for the optimisation (80, 100, 150, 200, 300, 400, 500 and 1000  $\mu\text{m/s}$ ). Due to the large dimensions of the imaged tissue samples (approximately  $8 \times 5$  mm), a laser spot size of 100  $\mu\text{m}$  was selected to reach a minimal LOD. An increase in the laser spot size resulted in a lower LOD [28, 142]. We calculated the apparent width  $w_{app}$  to evaluate the broadening caused by the various scan speed rates. We obtain  $w_{app}$  as the difference between the onset of the signal increase and the end of its decrease after the laser spot passed across the testing pattern (ink line). The onset points were the intersections of the trend lines a and b, and the endpoints were the intersections of the trend lines c and d. The trend lines a and d are linear regression fits to the domains of the signal between the printed lines, whereas the trend lines two and three resulted from linear regression in the domains of the signal rise and drop, respectively. We express the trueness of imaging as the relative broadening  $\Delta w_{rel}$  of the image  $w_{app}$  of the printed line with respect to its real width  $w$ .

The dependence of the relative broadening on the scan speed rate shows Fig. 6.2C. The relative broadening increased from 5% to more than 200% as the scan speed rate increased from 80 to 1000  $\mu\text{m/s}$ .

However, the lateral resolution and LOD were not the only parameters considered in developing the LA-ICP-MS elemental mapping method. The duration of analysis is an important parameter because it affects the operating costs. The Fig. 6.2B presents times required for mapping. We calculated the times for typical thin sections of our samples of tumour tissue ( $8 \times 5$  mm). The time required for analysis decreased with the increasing scan speed rate: whereas we need approximately 400 minutes for a scan speed rate of 80  $\mu\text{m/s}$ , we need only 20 minutes for a scan speed rate of 200  $\mu\text{m/s}$ . However, the broadening observed for these parameters was greater than 200%. Hence, we selected a scan speed rate of 200  $\mu\text{m/s}$  as a compromise because it resulted in a relative broadening of 40% and duration of analysis of 150 min.

Laser beam fluence is one of the most crucial parameters for laser ablation. The laser beam fluence mainly affects the ablation rate, the amount of material released during one laser pulse. Variations of the ablation rate complicate the quantification of LA-ICP-MS experiments because each laser pulse releases different amounts of analysed material in the selected range. There are multiple methods to compensate for this uncertainty. The first approach utilises normalisation to the sum of 100% [143–144] and can be successfully used for single-spot analysis or imaging of materials with well-known matrix composition to determine the appropriate multiplication coefficient that results in the whole content of 100%. We cannot use this approach for samples with a complex matrix containing large amounts of non-determinable elements or their groups (e.g., fluoroapatite, in which  $F-$  substitutes  $OH-$ , or biological samples containing  $O$ ,  $N$  and  $H$ ).

In our case, the analysed tumour tissue represents samples containing large amounts of non-determinable elements ( $O$ ,  $N$ , and  $H$ ). Hence, we cannot use the normalisation approach based on the sum of 100%. The second normalisation approach uses the utilisation of an internal standard [145], i.e., monitoring an isotope with a known amount. It is necessary to rely on internal standards such as  $C$ , which is abundant in the sample.



However, when we use carbon as an internal standard, marked systematic error arises due to the production of carbon-containing gaseous species, resulting in high losses of the carbon signal during laser ablation [146].

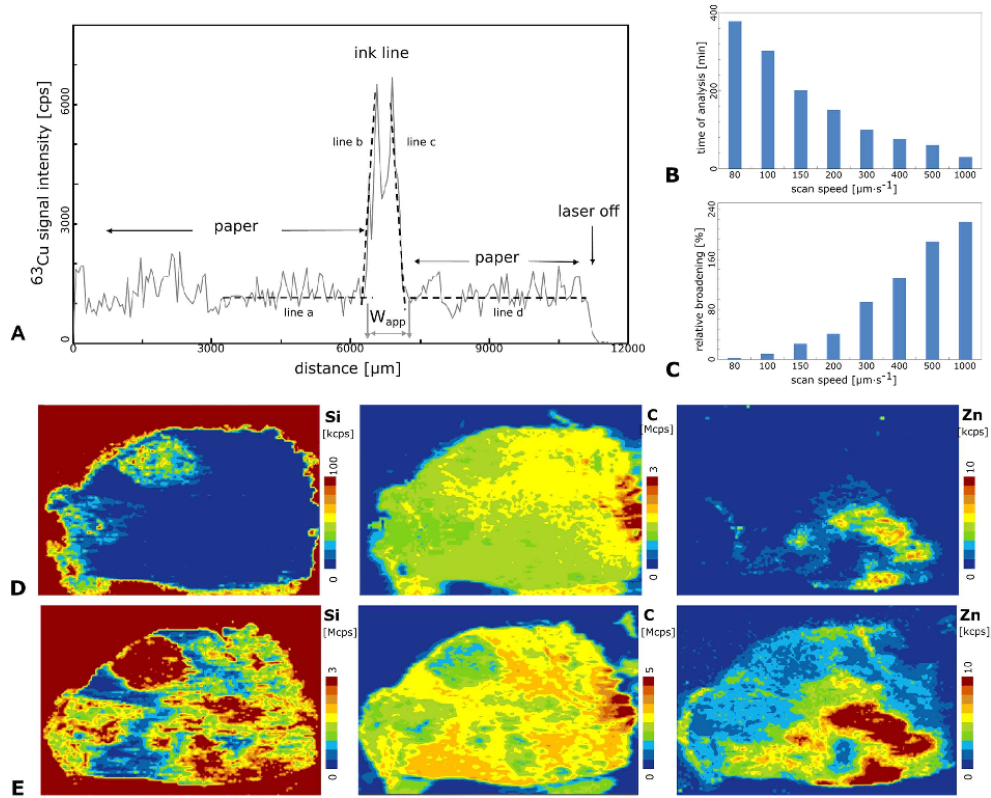
As mentioned above, differences in the ablation rate complicate imaging due to variations in the amount of ablated tissue. Releasing controlled amounts of material during each laser pulse minimise this phenomenon. Hence, we suggest a total mass removal approach when the whole layer of tissue is completely released. The  $Si$  signal indicates removing the whole layer of tissue as well as ablation of the glass substrate. Moreover, if the glass substrate does not contain special-interest elements ( $Zn$ ,  $Cu$  and  $C$ ), there is no danger of contamination from the glass substrate, and the signals of  $Zn$ ,  $Cu$  and  $C$  arise from the tissue only.

We compared the elemental images of two nearby thin sections. We ablated one section at high laser beam fluence ( $8 \text{ J/cm}^2$  – hard ablation), and the second section at low laser beam fluence ( $2 \text{ J/cm}^2$  – soft ablation). We use the terms hard and soft ablation in this text for explanation only. In the case of the soft ablation, the signal of  $^{28}Si$  corresponding to the ablation of the glass substrate under the tissue is not strongly enhanced compared to the gas blank value (Fig. 6.2D). Thus, the laser beam fluence is not sufficient to ablate the whole layer of the tissue, and we do not reach the glass substrate, except for two small regions in the left part of the tissue. We observe significantly higher intensities of  $^{28}Si$  when we apply the hard ablation (Fig. 6.2E). The range of the  $^{28}Si$  scale is 30 times larger than that of the soft ablation image, which indicates that we reached the glass substrate and that ablated the tumour tissue completely. In the case of the  $Zn$  image, we can observe strong enrichment in the lower right corner of tissue. The strongly enhanced  $Zn$  signal does not originate from the glass substrate, as confirmed by comparison with the parts of the image where we analysed only the glass substrate (red part from  $Si$  image and blue part of carbon image). The  $Zn$  signal is close to zero in all of these regions.

## 6.3 Data description

### 6.3.1 Histology

As a result, the collected data are arrays. In the case of the histology, the data are images. A special microscope scans the biological tissue sections. The resulting resolution of the resulting images is very high. The images are usually not provided in any of the standard image formats, because the very high resolution implies that the files are also very high. The data formats are often proprietary – the microscope scanning the biological sections stores the data in such a format that is only readable by the proprietary software provided with the apparatus. This approach, on the one hand, simplifies the work of the laboratory staff with these specific images. On the other hand, it hinders the processing of these images by advanced programs for image analysis. However, this problem common to many research teams solved the ImageJ program [147], and its Bioformats package [148]. The ImageJ program performs standard image processing as well as many more advanced methods. In our work, we only used the ImageJ and the Bioformats package for converting of the histological data in the proprietary format (Olympus .vsi proprietary image format) into a more accessible image format (common .jpeg). A scan of the usual tissue section of the size no bigger than  $10 \times 10 \text{ mm}$  results in an image of the size of several tens of thousands  $\times$  several tens of thousands of pixels. Such big images do not usually pose substantial problems for contemporary computers and standard image manipulation programs and operating systems. The advanced



**Figure 6.2.** A) LA-ICP-MS signal in line scan mode recorded for laser beam passing across one printed line at laser spot diameter of  $100\ \mu\text{m}$ , the scan speed of  $20\ \mu\text{m}/\text{s}$ , laser beam fluence of  $8\ \text{J}/\text{cm}^2$  and repetition rate of  $10\ \text{Hz}$ . B) Duration of scanning of the sample area of  $15 \times 15\ \text{mm}$  at various scan speeds ( $\mu\text{m}/\text{s}$ ). C) The relative broadening of a printed line (expressed in %) with a width of  $800\ \mu\text{m}$  obtained at various scan speeds ( $\mu\text{m}/\text{s}$ ). D) Elemental maps of *C*, *Si* and *Zn* obtained at “soft” ablation parameters ( $2\ \text{J}/\text{cm}^2$ ) for tissue K320/1 (12 weeks old). E) Elemental maps of *C*, *Si* and *Zn* obtained at “hard” ablation parameters ( $8\ \text{J}/\text{cm}^2$ ) for tissue K320/1 (12 weeks old). [J1]

processing of these images very quickly reaches the limits of the system when we try to use improperly implemented methods. One can either optimise the whole pipeline to minimise the system requirements of the processing methods or reasonably decompose the task and move the computations to a system with higher performance, e.g. a computer cluster for scientific computations. Often, we have to follow both paths - we have to optimise the method for processing to minimise the time and memory complexity as well as use computer systems with a higher volume of available memory.

In the histology images, we are usually interested in the objects in the image. These can be various tissue types and other objects of the biological samples (bristles, capillaries, etc.). The description of the objects in the tissue sections is rather complex, and we have to utilise an expert information. In our case, the expert information was the annotation of the tissue zones, as stated in the introductory section 6.2.1.

### 6.3.2 LA-ICP-MS

The results of the LA-ICP-MS analysis are also matrices. These data are not primarily images, even though they can be easily visualised as heatmaps or by similar techniques. The spatial properties of the LA-ICP-MS data are very different from those of the histology data. From the description in the data collection section 6.2.2, it is clear, that we can obtain high-resolution images. However, the resources spent in measuring

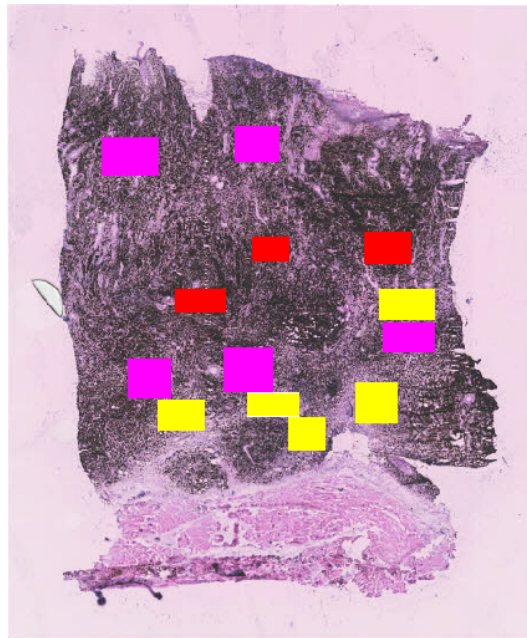
the biological sample with such a fine resolution does not match the gain in the obtained additional information. Where the histology scan of a biological section resulted in an image of the approximate size of  $10^4 \times 10^4$  pixels, the LA-ICP-MS measurement results in a matrix of approximate size  $10^2 \times 10^2$  elements. The difference in resolution is immense. This fact has several implications. First, if we are to integrate the histology and LA-ICP-MS data, we cannot approach this problem directly. We have to either make some heuristic guesses about the initial values for the data integration algorithm or include some intermediate step in the procedure. Second, the low spatial resolution can be a 'bottleneck' in the analyses. We can transform and interpolate the elemental maps. However, we cannot obtain more information from these procedures. Third, the measurement procedure, where the laser beam scans the sample, and we analyse the evaporated material, causes the data to be correlated. The LA-ICP-MS procedure analyses the elements simultaneously, so any anomaly in the analysed material can affect all the maps similarly, which only increases the collinearity in the data. Fourth, the inhomogeneity of the analysed material may lead to incorrect conclusions. For example, the biological material may locally differ in thickness. Various tissue types have different water content. In the preparation of the biological material, the tissue samples' storage at very low temperature causes evaporation (sublimation) of all the water in the tissues and therefore, the uneven thickness of the tissue section. All the problems indicate that we have to pay attention to the properties of the elemental map and assess them from several points of view.

### ■ 6.3.3 Additional information

In order to utilise some methods, additional information has to be available to the already mentioned histology and elemental matrices. The histology and the elemental matrices are spatial data measured in different resolution and by a different modality. In order to perform the data integration which is essential for the further processing and modelling of the data, it is very beneficial to have another a photography of the tissue slice used for the measurement of the elemental matrices. The photography has to be taken before the chemical analysis because the measurement process destroys the biological sample. Even though it is possible to integrate the histology directly and elemental matrices. A rough outline of the tissue sample is visible in the elemental maps. We can use a criterion independent of the actual values of the integrated datasets – such as the mutual information exploiting the joint distribution of the datasets – and perform the matching directly. A good matching between the sets can be obtained much more easily by utilising the information about the tissue sample from simple photography. We can obtain the outline, the shape and the position on the slide from the photography of the tissue sample. Generally, it is more straightforward for the data integration using image registration to combine two transformations, both of them relatively easily identifiable than to try to estimate parameters of complex transformation that combines highly heterogeneous datasets. In our case, it is easy to find a transformation between the histology and the tissue sample photography before the chemical analysis. These images originate from two adjacent tissue slices and are therefore very similar apart from a few local deformations. Similarly, the elemental maps and the photography of the tissue sample are very similar, because they originate from one physical object, the position and shape are identical, and they differ mostly in scaling and resolution. There is no problem in combining the two estimated transforms. They are applied one after the other to the transformed dataset. Of course, the direction of the transform matters very much (the transform is usually estimated to transform the moving image on the

template image). However, we can invert the transforms with a little use of matrix algebra and use them for the opposite direction (a transform from the template image to the moving image).

Another example of the additional information is the expert's annotation of the histology. In the work [J1] an expert annotated the tissue samples of porcine melanoma by the developmental stages of the melanoma. The annotation was in the form of rectangular areas of sufficiently homogeneous parts of the tissue with typical properties of the distinctive stages of the melanoma (see Fig. 6.1). The annotation can be used to select relevant areas in the elemental maps and to perform a comparison of the distributions of the traced elements in the elemental maps. The histological annotation of sample N115 is in Fig. 6.3. These two additional information sources for the analysis of the tissue samples are mentioned because they were used in the work [J1] and were crucial for the successful processing and statistical analysis of the data. However, these additional sources of information are by no means the only possible data, that can complement the processing and the analysis of the data.



Histological annotation of melanoma tissue

**Figure 6.3.** Example of the annotation of the tissue types (see Fig. 6.1) in the sample N115. The colour code: red - GMT, purple - ESR, yellow - LSR, green - FT. [C1]

## 6.4 Methods

### 6.4.1 Spatial covariance

In this section, we will discuss the spatial properties of the data. We will focus on the estimation of the self-similarity of the elemental maps that we use for safe indexing and subsetting of the data. The analysis of spatial data is an important concept in statistics. We can trace the origins of the spatial analysis to the early improvements in the cartography and surveying. One of the famous historical examples of the spatial analysis is the visualisation of the map of the cholera case in London in the 19th century or the visualisation of Napoleon Bonaparte's Russian campaign. These are just

examples showing that spatial analysis has a long history in the statistical analysis of data. The proper utilisation of the spatial information is a problem common to various statistical analyses - for example, the census data are one of the typical cases. The countries, states, and other regional units up to the level of small neighbourhoods are spatial units. The effect of the spatial parameters on various factors is not negligible. The various problems with the spatial properties of the data were mentioned several times throughout the text. The fundamental problems are the scale of the spatial data, the spatial dependence of the spatial data and the correct localisation of the spatial data. These basic problems further affect other statistical problems, for example, the sampling of the spatial data, the modelling of the spatial data and other spatial issues. The scale of the spatial data is problematic for several reasons. Ideally, we would like to analyse data at the proper scale, meaning measure the data in reasonable spatial units that provide us with a meaningful description of the studied phenomena. If we were to measure the data in finer resolution, we would obtain more detailed information, which does not have to be relevant to the analysis and in effect, view it as a noise. Also, if we were to measure the data in a coarser resolution, we would not get the relevant information at all. The domain knowledge may guide the choice of the proper resolution. Take the census data as an example of spatial data. We do not identify the surveyed people, and their actual address is not available due to the anonymisation of the data. However, we store certain spatial information (city district, city, village) and therefore, we can analyse the census data for the location. However, such information about location does not tell us lots about the interesting aspects that we would like to know. For example, people may travel to work in different areas. When analysing such data, we may miss many relationships that are governed by the true (and unrecorded) spatial information. These may be the place where the people work, where they spend their free time, where they travel for holidays, where do they come from, where do their relatives live, where live their friends, etc.

Similar relationships may be present in the biological tissue samples, where the cells are in a spatial structure. Adjacent cells are often similar because similar types of cells often form clusters or bigger structures in the tissue. As well as in cities, there are important places in the spatial structure of the tissues. Analogously to the roads between cities and streets in the between and in cities, the veins, arteries and other types of tubular structures facilitate not only the nutrient transport and exchange the oxygen and carbon dioxide, but the cells of the immune system use it for transportation. Various types of cells performing the functions of the immune system circulate through the bloodstream. This way they get to the places where they are needed, for example to inflamed tissues, cells attacked by viruses or other misbehaving cells.

The spatial dependence can be a multitude of relationships, as is usual in the field of statistics. In the case of analyses of spatial dependence, data collection and representation is of great importance. The form of the data dictates the available methods for the modelling of spatial relationships. We can approach the problem of assessing the spatial dependence in many ways. The choice of the method depends on our goal of the analysis - we may, for example, want only to know a strength of a relationship and in that case we would like to have a measure similar to the correlation coefficient for the spatial data; on another occasion we may want to directly model a spatial relationship in order to predict some useful value, in such a case we would seek after for alternatives of simpler models for spatial data; or we may want to only use certain spatial dependencies as covariates in a model which estimates variables that are not inherently spatial, but a spatial information is necessary to alleviate the target variable from some

unwanted effects.

There are several methods for the measurement of spatial dependence. In some cases, for example in regularly (on a rectangular grid) sampled data, we may model simple relationships directly. In the case where we measure several variables in the same 'coordinates', we may try to assign a direct relationship between the variables directly. If the variables do not completely overlap; however, we may assume relationships between small neighbouring areas, spatial smoothing methods may help to provide appropriate local spatial features for the assessment of local relationships. Distant relationships, where by distant relationships we mean a relationship between consistent areas which are not adjacent to each other, are not that easy to assess and model. In order to corroborate a distant relationship, we cannot rely on a simple comparison of the spatial data. To be sure of a relationship between distant areas additional dimension which contains non-trivial variation has to be present. In that case, we can test the relationship and find support for the hypothesis of the distant relationship between the areas. There are several approaches to measure the spatial relationships in data. The most general measure respecting the assumptions about the data is the Moran's I [149]. The Moran's I is a measure of spatial autocorrelation - a self-similarity of signal (a 2D data). The spatial autocorrelation is more complex than the one-dimensional autocorrelation well-known in the field of signal processing. The Moran's I is defined as follows:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (6.1)$$

where  $N$  is the number of spatial units indexed by  $i$  and  $j$ ,  $x$  is the variable of interest,  $\bar{x}$  is the average of the variable of interest  $x$ ,  $w_{ij}$  is a matrix of spatial weights with  $w_{ii} = 0$  and  $W$  is the sum of all  $w_{ij}$ . Another measure of the spatial autocorrelation is the Geary's C [150], which is defined as follows:

$$C = \frac{N-1}{2W} \frac{\sum_i \sum_j w_{ij} (x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \quad (6.2)$$

where again  $N$  is the number of spatial units indexed by  $i$  and  $j$ ,  $x$  is the variable of interest,  $\bar{x}$  is the average of the variable of interest  $x$ ,  $w_{ij}$  is a matrix of spatial weights with  $w_{ii} = 0$  and  $W$  is the sum of all  $w_{ij}$ .

The Moran's I and Geary's C are similar measures, which are quite general and flexible due to the use of the custom  $W$  matrix of weights. In case of evenly samples data (in the space), we may typically use the standard approaches using 4, 8 and more adjacent spatial units, or other widely used weight matrices for spatial data - various 2D bell curves. Both the measures of the spatial autocorrelation yield values that are usually in the range from -1 to +1. The values approaching -1 or lower than -1 indicate a strong negative spatial autocorrelation and values nearing +1 or greater than 1 indicate strong positive autocorrelation.

The spatial autocorrelation is practical for several reasons.

In many cases, the spatial autocorrelation can help us to discover spatial relationships, which we can further study in more detail. The spatial autocorrelation can also be beneficial as an overall measure of the smoothness of the spatial data.

In order to work with spatially co-varied data, we have to assume a certain model of the spatial covariance.

Similarly to the measures of the spatial autocorrelation, we have to either directly construct a neighbourhood matrix of weights, that indicates which spatial units are correlated and what is the strength of the relationship, which can be seen as a linear

model of the spatial covariance or utilise another model or method to estimate the spatial relationships. In the case of the linear approach based on the neighbourhood weight matrix, the model of the spatial relationships can be used to explain the spatial variation in the data in a straightforward way. The other approaches usually model the spatial relationships by a set of basis functions. An example of such an approach is kriging. Kriging is a method of spatial statistics that originated in geostatistics. Kriging is a method of interpolation where the model used for the interpolation of the data is a Gaussian process. The Gaussian process is a method of data interpolation or modelling, which estimates the underlying function for the data modelling or interpolation by a function that is governed by a covariance structure. The type and parameters of the covariance structure determine the shape of the function - the function can be constant, periodic, noise, or for example, a linear trend. Apart from the usual modelling techniques, the gaussian processes modelling utilises the local properties of the data to produce the values for interpolation or the prediction. Gaussian processes are used in a wide variety of fields, especially in neurobiology. An excellent resource on the Gaussian processes is [151], and the webpage [152] provides other resources. The kriging based on the Gaussian processes offers a very flexible framework of interpolation methods to use in unevenly sampled spatial data. As is usual in the field of data science, the problem is not how to utilise the kriging, but which of the different varieties to use and how to decide which of the varieties is the best.

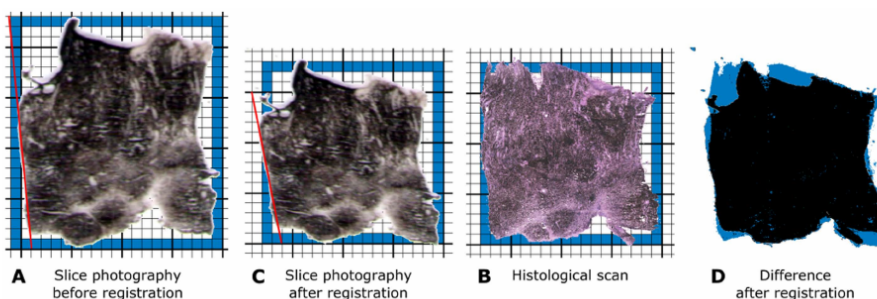
In many cases, the prior knowledge of the problem can provide leads which type of covariance kernel for the Gaussian process to choose; however, the domain knowledge can help only up to some extent. In practice, it is usually better to try several parameters, evaluate the obtained results and choose the best parameters for the given problem.

Another approach to the problem of utilising the spatial covariance in a similar way to kriging are methods based on the spline interpolation are extensively used in the regression, especially the B-splines. [153] The B-splines are defined by simple formulas for different degrees (linear, quadratic, cubic and higher degrees of splines) and by a set of knots, which are the breakpoints between the splines. The B-splines can provide a smooth approximation of non-linear functions. The B-splines are similar to polynomial regression. However, they provide greater variability in modelled relationships.

### ■ 6.4.2 Data integration

As it was mentioned in section 6.3.2, the spatial distribution of each chemical element is a matrix and can be depicted in the form of a heatmap (see Fig. 6.2 D,E). Because we obtain the heatmaps of all elements during one analysis of a single sample, the maps have an identical shape, orientation and resolution and are thus ideal for addressing questions such as the relationship between the presence of Zn and Cu in selected ablation pixels of the studied sample. However, the task becomes much more complex when we take into account additional information, such as the histological properties of the considered zone. Such information has to be determined from another tissue section because the standard procedure of the biological sample staining with dyes severely affects the distribution of analyzed metals. The dyes used for the staining of the biological sample contain metallic ions, and therefore, the addition of a dye changes the metals' content in the biological sample. Two neighbouring serial cryosections of the original tissue sample must be available, one of which is subjected to ablation, whereas we subject the other to standard staining for histological analysis. Both treatments produce digital images providing complementary information about the tissue sample. We must pair the data corresponding to the selected zone from both treatments. Fig. 6.4A and B

shows photographs of two neighbouring cryosections acquired under the same technical conditions; in an ideal case, the two samples should have identical contours. We use the two samples for different types of treatment: the sample in Fig. 6.4A is ready to undergo the ablation procedure, whereas we will subject the sample in Fig. 6.4B to histological analysis. Both samples are so tender that handling them may change both their shape and orientation slightly; the tissue can stretch, or some of its parts may fall off. Comparison of the two images reveals stretching in both dimensions, with a larger change in the vertical axis: the image in Fig. 6.4B fits into the blue rectangle with a size of  $17.5 \times 18$ , whereas an area of  $20 \times 23$  is necessary for the image in Fig. 6.4A. Linear transformation of one of the images is suitable to solve this problem. Linear transformation of the image in Fig. 6.4A is followed by registration to match the histological scan of the sample in Fig. 6.4B. The resulting image, Fig. 6.4C (obtained by transformation of Fig. 6.4A), has a size of  $18 \times 18$ , which is very close to that of Fig. 6.4B. The blue colour identifies the area in which the two images do not match in Fig. 6.4D. [J1]



**Figure 6.4.** Photographs of two neighbouring cryosections prepared for laser ablation and histological analysis (A, B). The image C is the result of registering the images (A, B). The blue rectangle indicates in the images (A-C) the minimal rectangle (with sides parallel to axes) the image fits in. The red line in the image (A, C) accentuates orientation of the corresponding borders on both the images. We compare the Images (B, C) in the image (D): while the places appearing in both images have black colour, the space in blue corresponds to the symmetric difference of both images. [J1]

Comparison of the slices before registration of Fig. 6.4A-B and after this registration (Fig. 6.4C) indicates that the slices do not have the same orientation. The comparison of the angles between the red lines defining one of the borders in both pictures and the horizontal line demonstrate the differences in orientation and sample deformation. Let us estimate the corresponding tangent values using the scale underlying both images: while this value is  $23/2 = 11.5$  for the image a, it is  $15/3 = 5$  for the image B. The resulting difference in the orientation of both images is approximately  $0.055\pi$  (or  $10^\circ$ ). Moreover, the size and resolution obtained from the elemental map and the histological image can differ by order of magnitude. The difference depends on the applied magnification. Thus, the absolute size of a pixel in the histological image differs significantly from that of the ablation pixel. We have to identify a homogeneous cluster of cells in the histological image and locate it in all the elemental maps to make full use of the information about the spatial distribution of different metals in the sample. This task can be approximately resolved manually by taking advantage of the human ability to match similar objects, as demonstrated in a breakthrough study of nine samples of invasive breast carcinoma [154]. However, manual matching is not a viable solution for frequent analysis of large sample sets. We, therefore, utilised a method that



automates the process of combining and matching spatially resolved results from diverse imaging techniques, namely histological and spectroscopic descriptions. Our method, schematised in Fig. 2.7, first registers the digital images to project the contours specified in one image to the corresponding pixels of the other images. Consequently, if we outline a zone of interest  $Z$  in one of the images (e.g., histological image), the matching zone  $Z'$  is identified automatically in any other image. Thus, it is possible to combine available complementary data about both matching zones  $Z$  and  $Z'$  (e.g., histology description of  $Z$  and Zn content obtained from  $Z'$ ). The matching is a necessary step towards a modern methodology of analysis, interpretation and integration of biochemical data from diverse sources. [J1]

Image registration is used to create layered multidisciplinary description; in other words, the integrated data set. Each tissue sample was submitted to analysis by two fully independent methods, namely histological scanning and the LA-ICP-MS measurement. We compared the results for the presence of Zn and Cu in specified histologically uniform locations of the sample. Each of the applied methods processes (and destroys) one of the two bordering serial tissue sections from the same biological sample (Fig. 2.7A), whereas each delivers its results in the form of a digital image. [J1]

Morphology of the studied tissue suggests that the corresponding zones in these sections can be assumed to represent identical histological structures. This proposition holds provided the selected zones are placed inside of a histologically homogenous tissue, and their diameter is several times bigger than the thickness of the used slices. We respected these conditions during data collection. We applied the image registration [30], an extensively studied method of digital image processing widely used in computer vision. Image registration is a standard method for overlaying two different images. Particularly, the methodology of image registration is well developed and offers ample approaches for overlaying two or more images of the same section obtained from different viewpoints or by different sensors [30]. We chose the affine transformation [155] for the registration of the studied images after considering other relevant methods. The main benefit of the affine transformation is its simplicity and understandability due to the linear transformation it applies to map the new image on the reference image. Let us assume that an image is a function of two variables  $I(x, y)$  that assign an intensity value to the pixel with specific coordinates  $x$  and  $y$ . The affine transformation of the 2D image is a simple linear mapping in the form of:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (6.3)$$

Where  $x$  and  $y$  are the coordinates of pixels in the original image and  $x'$  and  $y'$  are the coordinates of the corresponding pixels in the transformed image. The constant parameters  $a-f$  of the  $3 \times 3$  transformation matrix  $T$  in the middle of the equation fully characterise the treatment of the image. The affine transformation can accomplish translation, rotation, and scaling as well as shear deformation of pixels. Symmetric difference of both images characterises the quality of the match between the reference image and the transformed image, as depicted in Fig. 6.5. This difference should be zero in the ideal case. Multiresolution image registration [156] that applies an iterative gradient algorithm is one of the basic procedures for estimation of the parameters  $a-f$  of the transformation. It is robust and ensures good results. The registered images were reduced to silhouettes to simplify the parameter estimation and to avoid problems of different modalities of the registered images. The use of silhouettes allows for the

definition of the brightness function  $I(x, y)$ , defined as follows:

$$\begin{aligned} I(x, y) &= 1 && \text{for a pixel belonging to the silhouette} \\ I(x, y) &= 0 && \text{elsewhere} \end{aligned} \quad (6.4)$$

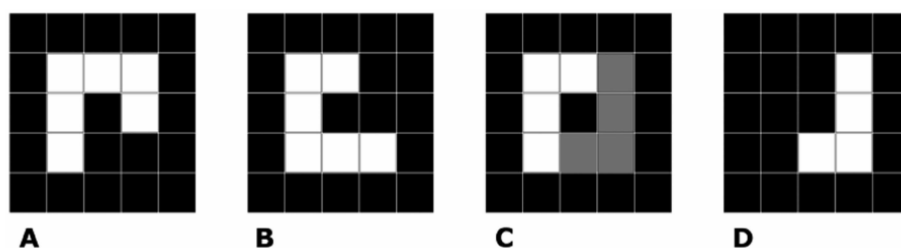
First, the registered image must be in the same coordinate system as the reference image. We chose The sum of squared differences between the reference image and the registered image (SSD, Fig. 6.4) as the criterion to be minimised during the registration procedure. For the silhouettes of the reference image  $I_{ref}(x, y)$  and the image to be registered  $I_{reg}(x, y)$ , the SSD may be defined as:

$$SSD(I_{ref}, I_{reg}) = \sum_x \sum_y (I_{ref}(x, y) - I_{reg}(x, y))^2 \quad (6.5)$$

We can determine the parameters of the mapping between both images by minimising the SSD with the gradient algorithm. We used the iterative (gradual) estimation of the parameters on a dyadic decomposition of the images to ensure convergence of the gradient algorithm [156]. This multiresolution image registration approach decomposes both considered images into a sequence of images with decreasing resolution (the resolution of each successive image is half that of the preceding image). The maximal length  $l_{max}$  of the sequence of these decompositions depends on the integer part of the smallest dimension  $d_{min}$  (width and height of the considered image) of both silhouettes. The maximal length gives the following expression as:

$$l_{max} = \log_2(d_{min}) - 1 \quad (6.6)$$

This upper limit for  $l_{max}$  ensures that any of the images in the sequence will have at least 2 pixels in its smallest dimension. The procedure starts with a pair of images (reference and registered) with the lowest level of resolution. For the given resolution, the gradient algorithm estimates the parameters of the transformation. Their values are used as the initial choice of parameters for the estimation of transformation in the next step, which treats the pair of images with resolution two times higher than the last (the iterative step). This process continues until the original resolution of both images is reached, and we obtain the final parameter estimates. There are even more powerful types of transformations, but affine transformation proved to be sufficient for our purposes. We can identify all operations performed by the affine transformation on the image of the tissue sample with the actual sample treatment. The compressing or stretching relates to the cutting of the cryosections. The shifting and rotating correspond to the placement of the cryosection on the slide. The described transformation of coordinates must be followed by the interpolation of the original brightness function to obtain detailed information about the transformed image in the new coordinate system. We used linear interpolation. We aim to provide complex information about individual areas of the tissue samples as provided by the considered methods for their analysis. The first step toward this goal is to determine the match between the histological scan and tissue slice photography and also between the tissue slice photography and the laser ablation measurements. We estimated the parameters of both transformations by MATLAB's universal optimiser for unconstrained optimisation supplied with the optimisation toolbox. [J1]



**Figure 6.5.** The illustration of the SSD criterion for image registration. The panels (A,B) show the reference image and the image to be registered. The panel (C) shows overlaying of both the images, the gray part corresponds to the difference between the images. The panel (D) shows the difference. As the patterns are  $3 \times 3$  pixels, the difference part corresponds to SSD of 4. [J1]

### 6.4.3 Spatial properties of the melanoma tissue

The assumption that the histological zones coincide in two neighbouring tissue sections seems reasonable. The sections are relatively thin, and in case of neighbouring tissue sections, the tissue structure is actually split and therefore, we should see a mirror image in the two sections. We can easily apply the presented method for creation of multilayered data representation to tissue sections that are not direct neighbours. The question is then, can we reliably register images of more distant tissue sections? In order to examine this proposition, we decided to design a separate experiment which would enable us to assess the similarity of images of tissue sections after image registration. We scanned a few dozens of serial tissue sections and obtained their annotated histological zones. The scanned images of the dozens of serial sections showed that the variability in the deformations applied to the physical sections is much greater than the deformations of the images that can be dealt with by the affine transformation. We can choose a pair of neighbouring sections so that the quality of the sections is fairly high, however, when producing a long sequence of serial sections, the higher quality cannot be maintained. Unfortunately, a not-insignificant amount of the tissue section preparation relies on manual processing. A special device cuts the tissue sections automatically, however afterwards they have to be transferred manually to a slide for further processing and scanning by the microscope. During the transfer, the sections cannot only be stretched, but also torn, or parts of the section can be folded over. We cannot deal with these new types of deformations by the simple use of silhouettes. There are many solutions. First, we can try to get rid of the folded parts and try to perform the image registration with affine transformation. The folding of the section introduces non-linearities to the sought-after transformation. The transformation is not the same for all the pixels in the images. In order to obtain a good matching between the images, we have to deform different parts of the image differently. Second, we can use a more general type of transformation. A type of transformation that can transform differently various parts of the image. Such a transformation is possible with elastic registration [30]. Simple review of the elastic registration methods shows that in many areas there have been developed fast and accurate algorithms to perform the elastic registration. However in general, the accuracy of the overlaying of the registered images is not guaranteed. Certain landmark points (points that can be identified in both the images reliably and matched as corresponding pair) can be utilised and improve the resulting transformation [157].

To assess the spatial properties of the tissue samples with respect to the histological

zones annotations, the images of adjacent serial sections were registered by the elastic registration. The transformation between images of tissue sections was produced as a combination of the transformations between the adjacent images of tissue sections. The estimated transformation was used to transform manual histological annotations of several distinct objects in the images. The annotations consisted mainly of the same histological zones as in the main analysis of the spatial distribution of bioactive metals - namely the growing melanoma tissue zone, the early and late spontaneous regression and the fibrous tissue. However, we also used the bristles because they are easily identified in the histological scans even by an untrained person. The bristles look like several concentric rings and because of their size go through several serial sections. The size of a bristle is several times greater than the size of other objects present in the histology images (cells, capillaries, etc.). The good match of the bristles in adjacent samples can be a sign of good image registration. The histological annotation of the more difficult to identify objects - the tissue zones - can be however considered a decisive sign for the appropriateness of the image registration. A common measure of the overlap between two sets of points is the Dice similarity coefficient [158] which is in other fields known as the F1 score - for example in data mining and machine learning. Other similar measures are the Jaccard similarity and Tanimoto similarity [158]. We computed the values of the Dice similarity coefficient between an original annotation for a given tissue section and a transformed annotation from the registered tissue section. We evaluated each annotated type separately as well as all annotations together.

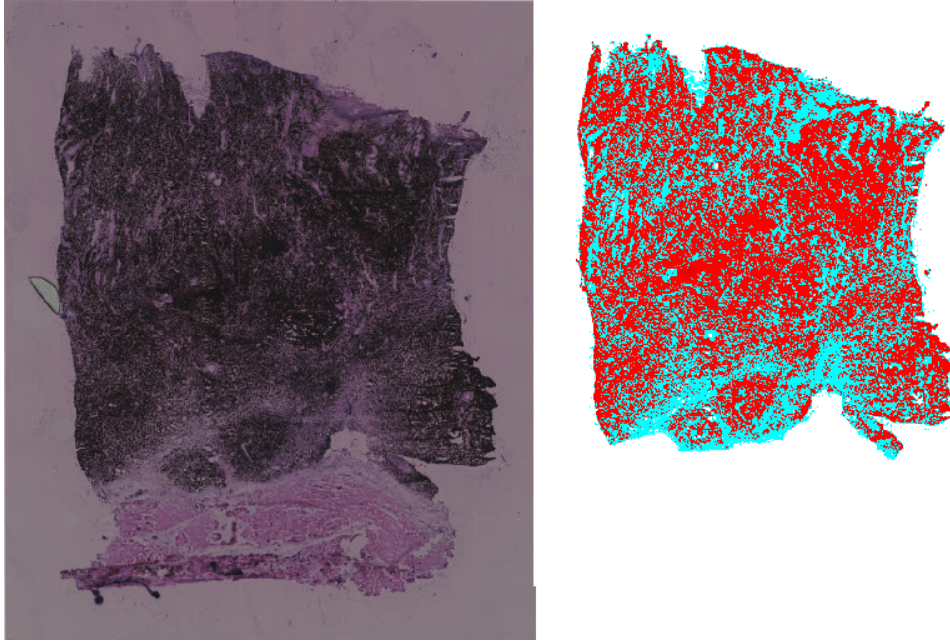
#### ■ 6.4.4 Unsupervised analysis of histological zones and elemental maps

To examine the structure of the tissue sections, we decided to use standard clustering algorithms used in image processing. We performed clustering with the state-of-the-art spectral clustering [159]. The clustering was carried out using the R software [160] and the package Kernlab [161]. In a series of experiments, we applied the clustering algorithm first to the histological images. Second, we analysed the elemental matrices. Third, we used the combination of the histological images and the elemental matrices. We used the 2D maximum overlap wavelet transform (2D MODWT) [162] to extract local features of the image.

We used the spectral clustering with kernel distance matrix. The used kernel was the radial basis function and the parameter  $\sigma$  of the kernel optimised by the heuristic function supplied by the Kernlab package [161]. We chose the number of clusters as the drop in values of the distance matrix eigenvalues. We compared the results of the clustering by tables of the coincidence between cluster assignment and the histological annotation. The tables were tested by  $\chi^2$  test to assess any relationships between the different clustering experiments and the histological zones. Consequently, the logistic regression model was used to test each category of the histological zone and each cluster. We considered the relationship significant if the  $p$ -value was lower than a threshold value of 0.05 corrected for the actual number of all test by Bonferroni correction. [C1]

#### ■ 6.4.5 Statistical analysis of the differences in the metals spatial distribution

The exploratory analysis of the metal distribution in the samples showed a variation in the values of the metal distribution in different individual animals. Another important variable that seems to affect the distribution of the metals is the age of the animal. Furthermore, not all the histology zones could be annotated in all the animals reliably, and therefore, some histology zones were not available for all the biological samples.



**Figure 6.6.** Simple visualisation of image segmentation of a histology scan of tissue section based on the 2D MODWT features. There does not seem to be a relationship between any finer structure in the histological scan than a rough classification of the tissues in melanoma cells and the fibrous tissue. [C1]

Such a problem is difficult to model by the standard statistical modelling techniques. However, the problem definition fits perfectly into the framework of mixed effect models (see chapter 4). We used the following model structure to assess the differences among the annotated areas:

$$C_M = \beta_{Tissue\ type} \cdot Tissue\ type + \beta_0 + b_{0,Animal} + \epsilon \quad (6.7)$$

Where the  $C_M$  indicates the content of a metal  $M$  in a specific annotated histological zone,  $Tissue\ type$  is the indicator of tissue type (GMT, ESR, LSR and FT),  $Age$  is the age of the animal, the  $\beta_{Tissue\ type}$  and  $\beta_0$  are the model coefficient in the fixed effects part of the model. The  $b_{0,Animal}$ , the coefficients in the random effect part of the model are specific parameters for each animal, the  $b_{0,Animal}$  is the random intercept. More complicated model structures - for example, including random slopes are not possible in this case. The variable  $Tissue\ type$  is categorical variable, and therefore, the model coefficients are sample means (the intercept) and the differences between the sample means between the categories. There is no unexplained variation left for the estimation of the random slopes - the inclusion of the random slopes leads to singular model fits. The  $\epsilon$  denotes the residuals. In order to assess the statistical significance of the model coefficients, namely the coefficients of the fixed effect part of the model, we utilised the cases bootstrap, see Algorithm 4.3. We sampled the data on both levels (the measurements for each animal, the animals).

## 6.5 Results

### 6.5.1 Data integration

The simple approach to the data integration as presented in [J1] and reported here is quite simple and robust. The drawback of this approach is its roughness. The resulting

transformation is global, and therefore, it does not allow for any local deformations. The objects' boundaries in the images govern the transformation due to the use of the image silhouettes. In the case of the data integration for mapping of metals in melanoma tissue, the collected data did allow for more refined approaches. The reason was the vast difference in the resolution and the level of detail in the images. The following section presents a more accurate approach using the elastic transformation for image registration, but on a different data set. The data integration resulted in a layered representation of the histological scans and the elemental maps of Zn and Cu. The layered representation allowed identification of specific areas – tissue zones from the histology scans – and their projection on the elemental maps and further estimation of average metal contents for these areas.

### ■ 6.5.2 Spatial properties of the melanoma tissue

An important property of the examined tissue is its spatial consistency. We denote tissue sections, whose structural appearance is similar or even identical not only in adjacent sections but also in more distant tissue sections, as spatially consistent. The spatial consistency is crucial because of the chosen methodology of the examination of the spatial distribution of the bioactive metals. Without the spatial consistency, we cannot assume that the resulting layered description of the histological data and elemental maps enables us to assume the relationships among the distribution of bioactive metals and developmental stages of the porcine melanoma. In order to examine the spatial properties, a series of 12 sections from a pig melanoma was stained and scanned. An expert annotated the relevant structures (tissue zones corresponding to melanoma development, fibrous tissue and bristles) in each section. The scanning process does not preserve the physical dimensions of the tissue section in the resulting image. The manipulation with sections also resulted in some deformation of the tissue sections. Due to these issues with the images of the tissue sections, the images have to be registered, and the affine image transformation does not offer enough distortion to register the images successively. That is why the elastic registration is necessary. The estimation of the parameters of the elastic transformation can be problematic, and generally, the optimisation of the parameters does not guarantee to find the global optimum. We used a general optimisation from the ANTsR package [163] with no special initiation of the parameters.

This approach to the elastic registration, which is to apply the elastic registration to the images, without any specific initialisation or complex image similarity function incorporating information about the landmark and image intensity, is not guaranteed to provide optimal results. The unconstrained optimisation of the elastic transformation parameters can lead to highly variable results.

When applying this procedure to our data set of serial tissue sections, we may observe that there are undoubtedly well-registered pairs of images. However, there are also many not-so-well-registered pairs of images. The extension of the transformations estimated on the pairs of adjacent tissue sections leads to highly distorted results for images that are more than one tissue section apart with badly registered images in the chain of the transformations. The results of the comparison of the matching of the annotated areas reflect the presence of badly registered image pairs. Tab. 6.1 presents the resulting DICE indexes.

#### 6.5.2.1 Concluding remarks

The elastic image registration provides a useful tool for the matching of images of severely distorted tissue sections. However the process so far is not good enough to

	Annotated tissue zones					
	GMT	ESR	LSR	FT	BS	ALL
$D = 8\mu m$	0.78	0.64	0.72	0.12	0.79	0.73
$D = 16\mu m$	0.70	0.55	0.56	0.07	0.72	0.64
$D = 24\mu m$	0.65	0.51	0.41	0.04	0.70	0.55
$D = 32\mu m$	0.56	0.38	0.15	0.03	0.61	0.43
$D = 40\mu m$	0.50	0.31	0.01	0.01	0.52	0.34
$D = 48\mu m$	0.46	0.28	0.01	0.00	0.49	0.27
$D = 56\mu m$	0.40	0.25	0.02	0.00	0.47	0.20
$D = 64\mu m$	0.38	0.25	0.03	0.00	0.42	0.19
$D = 72\mu m$	0.37	0.26	0.02	0.00	0.38	0.18
$D = 80\mu m$	0.37	0.31	0.03	0.00	0.31	0.18
$D = 88\mu m$	0.36	0.33	0.04	0.00	0.25	0.19

**Table 6.1.** The relationship between the average Dice coefficient and the distance of the tissue sections indicated by the value of  $D$  for images registered by the elastic transformation. The annotated states are GMT - growing melanoma tissue, ESR - early spontaneous regression, LSR - late spontaneous regression, FT - fibrous tissue, BS - bristle sheath and ALL which are all the previous tissue zones together. The average values of Dice coefficients indicate that there is a good match between the adjacent ( $D = 8\mu m$ ) tissue sections (except the FT). The Dice coefficients for the tissue sections as far as three sections apart ( $D = 24\mu m$ ) are still showing a good match. However, the sections that are more than three sections apart ( $D = 24\mu m$ ) do not seem to be well-matched against each other.

provide good elastic transformation for any pair of images in the set, and therefore we cannot show, that the spatial properties of the melanoma tissue extend further than  $24\mu m$ . This finding supports the use of adjacent tissue sections for separate chemical, biochemical and histological analyses, which we have performed with the LA-ICP-MS and histology in our main analysis. The finding also means that in planning further analyses it is better to be on the safe side and when performing several chemical and biochemical analyses, tissue sections for the histological analysis should be in between the tissue sections for the biochemical analyses. It is important to note that there are many possibilities for improvements in this task. As in any optimisation, a good set of initial values results in faster convergence and attaining better values of the optimisation criteria - the bristles or other landmarks in the tissue sections may be used for this purpose. These landmarks can be utilised not only for the initiation of the optimisation procedure, but they also may be incorporated into the optimisation criterion, as shown in [157] and [164].

### 6.5.3 Unsupervised analysis histological zones and elemental maps

#### 6.5.3.1 Clustering of elemental maps

The clustering of the elemental maps showed that there are no distinctive clusters in the elemental maps. In all the samples, one cluster dominates the cluster assignment, which contains the majority of the observations, and the remaining clusters represent only a small fraction of the data. Only one sample (N113) indicated a possible relationship between the clustering and the histological annotation. However, none of the following tests by the logistic regression showed any significant result. Therefore, we may assume there is no simple relationship between the elemental maps and the histological annotation.

### 6.5.3.2 Clustering of histology

The histological images contain areas which may be divided into clusters easily. Based on the original histological image and local features, the clustering can assign the pixel correctly into these categories. The clustering seems to be consistent with the histological areas (except samples L619, N129). In the majority of samples, we can detect a relationship and identify the clusters which correspond to the histological zones. Unfortunately, the results do not allow to distinguish all the histological zones, but only the zones in broader categories of melanoma tissue or fibrous tissue. The visualisation of the result of clustering is in Fig. 6.6

### 6.5.3.3 Clustering of the combination of histology and elemental maps

We utilised the layered data description for this experiment. The results of the clustering were similar to those of the clustering of elemental maps. The addition of the elemental maps features to the promising results on histological data did not result in better clustering results. Only two samples passed the  $\chi^2$  test (L619, N129). In the case of these two samples, the following test by logistic regression model did not show any specific relationships among the histological zones and the clustering.

### 6.5.3.4 Concluding remarks

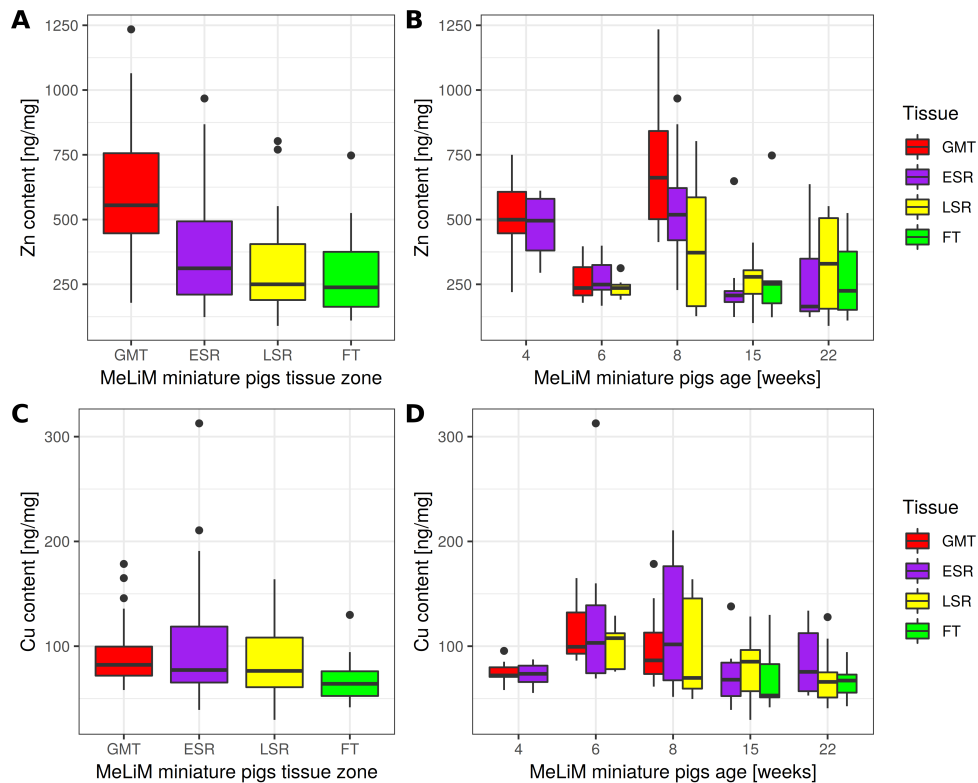
In conclusion by the application of the spectral clustering algorithm on data from pigs melanoma samples, we were able to detect structure in the histological images, but not in the distribution of metals (Cu, Zn) or the combination of the histology and distribution of metals. We tested the relationship between the histological zones and the cluster assignment. The procedure consisted of testing by  $\chi^2$  test for any relationship followed by logistic regression. We tried the follow-up logistic regression in case of a positive result from  $\chi^2$  test to test histological zones against clusters. This procedure showed that the clusters data might be related to the histological zones in broader terms. There may exist a relationship between the melanoma cells and other types of tissue. The exploratory analyses presented in this paper show us that there is a structure in the data, but even though we used the state-of-the-art methods for image processing, we were not able to relate the clusters and histological information completely.

## 6.5.4 Statistical analysis

The data integration procedures provided data for ten tissue samples each from 10 individual animals of five postnatal ages. We observed three histologically different zones in each sample obtained from minipigs at six weeks of age or older (GMT, ESR and LSR at the age of six weeks; ESR, LSR and FT at the age of 15 and 22 weeks), two zones only (GMT and ESR) were detected in the samples from the 4-week-old minipigs. In each sample, 10 to 15 spots (3-5 per each zone) were annotated, resulting in 125 annotated spots. The exploration of the integrated data shows that there are several relationships. We can see a clear decreasing trend in the content of the Zn in the stages of melanoma development.

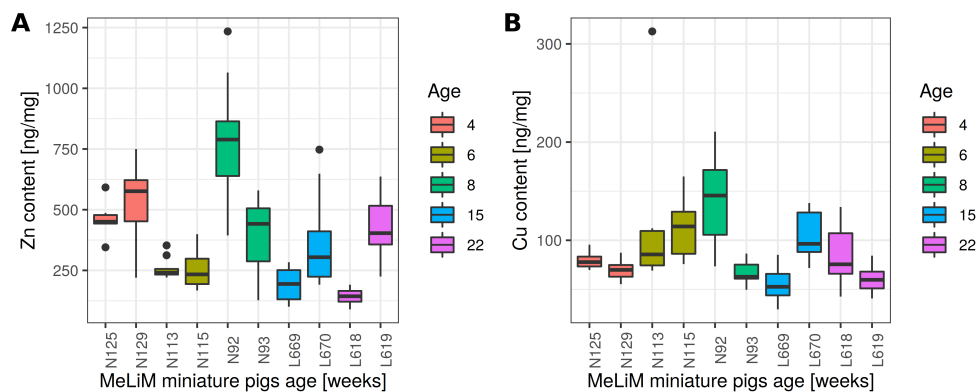
In Fig. 6.7A we can see that the content of the Zn is the highest in the GMT tissue zone and that the contents follow a trend GMT > ESR > LSR > FT. We can observe even more pronounced decrease in the content of Cu with relationship to the melanoma tissue development in Fig. 6.7C. However, other relationships are affecting the data. For example, the age of the animal can be a source of substantial differences in the content of the bioactive metals, see Fig. 6.7BD. Another important





**Figure 6.7.** Content of Zn and Cu in all samples (A, C) and in samples stratified according to age of minipigs (B, D). [J1]

variability in the data is related to the inter-individual differences. Fig. 6.8 shows that the differences in the average content of Zn and Cu vary greatly among the animals.



**Figure 6.8.** The content of Zn (A) and Cu (B) in different samples. There are distinct differences among the animals in the Zn and Cu content. These differences have to be taken into account when modelling the relationships among the melanoma development stages and bioactive metal content. [J1]

All this evidence indicates that simple linear models are not sufficiently complex to capture all the relevant variability in the data. Without using the mixed effect model, we could not integrate information about the inter-individual differences into our reasoning about the relationships among the developmental stages of the porcine melanoma and the content of bioactive metals.

We modelled the differences among the tissue zones as fixed effects in the mixed effect model, and we used the indicator of an animal as a random effect. We corrected the resulting  $p$ -values estimated by non-parametric bootstrap test utilising 9999 random samples from the data by the Bonferroni correction for multiple comparison [93]. Tab. 6.2 and 6.3 presents the estimated fixed effect coefficients.

	GMT	ESR	LSR	FT
GMT	<b>497.05</b>	<b>-138.15</b>	<b>-162.66</b>	<b>-142.32</b>
ESR	-	358.90	-24.51	-4.17
LSR	-	-	334.39	20.34
FT	-	-	-	354.73

**Table 6.2.** Estimated fixed effect coefficients of a model of differences in Zn content among tissue zones. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Zn content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Zn content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

	GMT	ESR	LSR	FT
GMT	154.35	-41.39	-77.43	-70.80
ESR	-	112.96	-36.04	-29.41
LSR	-	-	76.92	6.63
FT	-	-	-	83.55

**Table 6.3.** Estimated fixed effect coefficients of a model of differences in Cu content among tissue zones. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Cu content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Cu content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

The values in Tab. 6.2 and 6.3 shows that the differences among the tissue zones when accounting for the inter-individual differences are less distinctive, than it may seem from the visualisations. The statistical significance from the bootstrap test corrected for multiple comparisons and indicated in the Tab. 6.2 and 6.3 indicates that for the content of Zn, there is a clear difference between the GMT and the other tissue zones. The content of Zn differs significantly from the content of Zn in any other tissue zone.

On the other hand, the three remaining tissue zones (ESR, LSR, FT) are not distinctive enough to be considered different from each other.

Similarly, in the case of the content of Cu in the various tissue zones, there are no significant differences after the correction for multiple comparisons. However, the biggest difference is between the ESR and the LSR stages. These observations can be used to revise our understanding of the problem and reconsider the structure of the model. We can attempt to simplify the classification of the tissue zones into the new description, which could be classification into GMT and non-GMT tissues (more accurately into GMT tissue zone against ESR & LSR & FT), which seem to be the case for the content of ZN in the melanoma tissue. The other simplification of the classification of tissue zones can be GMT & ESR classes versus LSR & FT. This classification seems to reflect the observed differences in the Cu content in the melanoma tissue. Further tests with

Coefficients		
	GMT	SR&FT
GMT	<b>495.90</b>	<b>-146.38</b>
SR+FT	-	349.52
P-values		
	GMT	SR&FT
GMT	<b>0.0018</b>	<b>0.0011</b>
SR&FT	-	0.0362

**Table 6.4.** Estimated fixed effect coefficients of a model of differences in Zn content among revised tissue zones – the tissue zones are now classified as GMT and SR&FT. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Zn content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Zn content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

Coefficients		
	GMT&ESR	LSR&FT
GMT&ESR	401.84	-48.49
LSR&FT	-	353.35
P-values		
	GMT&ESR	LSR&FT
GMT&ESR	0.2488	0.0695
LSR&FT	-	0.1712

**Table 6.5.** Estimated fixed effect coefficients of a model of differences in Zn content among revised tissue zones – the tissue zones are now classified as GMT&ESR and LSR&FT. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Zn content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Zn content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

the same methodology and the revised classification of the tissue zones are presented in Tab. 6.4, 6.5, 6.6 and 6.7

The presented *p*-values indicate that the null hypothesis can be rejected only in the case of Zn. Our data confirm that the Zn content in the zone of growing melanoma tissue (GMT) was significantly higher than in all remaining zones, which represent consecutive stages of the tumour tissue that arise as a result of the spontaneous regression of melanoma (ESR, LSR) and its final rebuilding into fibrous tissue.

## 6.6 Discussion

The presented results show that it is possible to utilise data integration techniques - the image registration - to combine the histological description of a tissue section with the spatial distribution of bioactive metals into an integrated data set, the so-called layered data representation. With such data, many questions about the spatial relationships between the bioactive metals and the histological description can be explored and tested. In doing so, many other problems have arisen. Firstly, it is necessary to judge the appropriateness of the image registration and the type of utilised transform.

Coefficients		
	GMT	SR&FT
GMT	151.70	-55.80
SR&FT	-	95.89
P-values		
	GMT	SR&FT
GMT	0.1137	0.0650
SR&FT	-	0.4127

**Table 6.6.** Estimated fixed effect coefficients of a model of differences in Cu content among revised tissue zones – the tissue zones are now classified as GMT and SR&FT. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Cu content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Cu content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

Coefficients		
	GMT&ESR	LSR&FT
GMT&ESR	126.26	-44.80
LSR&FT	-	81.46
P-values		
	GMT&ESR	LSR&FT
GMT&ESR	0.2491	0.1399
LSR&FT	-	0.2328

**Table 6.7.** Estimated fixed effect coefficients of a model of differences in Cu content among revised tissue zones – the tissue zones are now classified as GMT&ESR and LSR&FT. The values on the diagonal correspond to averages of the respective tissue zone. The off-diagonal values represent the differences. The Cu content in the tissue zone denoted by the row name plus the difference in the appropriate table cell gives the Cu content in the tissue zone denoted by the column name. Statistically significant values are in boldface.

In our work, we showed, that for tissue sections of reasonable quality, the affine transformation is necessary to compensate for the distortion of the tissue sections due to the manipulation and other treatment. In the case of serial sections, which may be treated more roughly, which results in more substantial deformation of the tissue sections such as folding over, we have to use a more powerful transform. Such a transform is the elastic transform in its various forms. The elastic transform brings a whole new lot of problems to the data integration procedure. However, we can overcome these problems through the use of sophisticated image registration procedures. In any case, the data integration - image registration - is necessary for reliable and reproducible research in the spatial distribution of chemical elements and compounds in tissue sections. Secondly, the whole procedure relies on the similarity between the tissue sections submitted for histology and laser ablation measurement. Without the assumed similarity, we cannot use the whole process for any further analysis. In case of adjacent tissue sections, there does not seem to be a problem - the device splits the tissue structure (for example clusters of tumour cells), and therefore the tissue structures should be present in both the adjacent sections in the same places. However, in cases of tissue sections, that are more distant, the similarity may not be guaranteed. We tried to address this problem

by integrating a large set of tissue sections from one piece of tissue. In these sections, the prominent tissue structures were annotated and afterwards compared between the sections. The serial sections from one piece of tissue called for a different treatment when integrating. However, the resulting comparison of the matching of these structures suggests that not only the adjacent but also more distant sections are reasonably similar and can be used for measurement by different analyses with consequent integration by image registration. We did not succeed in our endeavour to perform an unsupervised analysis of the tissue section. To recognise the different tissue zones by analysing the histological scans. We were only able to differentiate between the most distinct tissue types, which are the melanoma cells and the fibrous tissue. Serum levels of Zn and Cu in melanoma patients have been suggested as valuable diagnostic and prognostic parameters but have yielded conflicting results. The Cu level (but not the Zn level) was generally elevated in melanoma patients, reflecting the degree and extent of tumour activity [165]. By contrast, serum Cu concentrations were identical in melanoma patients and healthy individuals, whereas the serum Zn concentration was significantly increased in melanoma patients [166]. In tissue sections, the Zn level was elevated in the majority of melanomas in comparison with the skin of healthy controls. However, the Cu level was increased in some melanoma patients [167].

The suggested technique for Zn and Cu mapping permits not only simultaneous quantification in the same tissue cryosection but also detection of both metal ions in very small, histologically characterised zones of tissue. This is an important advantage compared to their determination in tissue homogenates or whole tissue sections. We used skin melanoma samples from MeLiM animals of various ages to develop and validate this technique. Our findings (Fig. 6.7A–C) indicate that the Zn content of a given zone is approximately 3 or 4 times higher than that of Cu (Fig. 6.7D–F). Moreover, the content of both metals declines as a result of advancing spontaneous regression (due to destruction of melanoma cells by anti-tumour immune reaction).

## Chapter 7


### Conclusion

The thesis provides several results contributing the biochemical data analysis. The biochemical data analysis is specific due to three main problems of the biochemical data. First, the biochemical data are high-dimensional; the data are in the form of vectors, but more often in the form of matrices and multidimensional arrays. We have to utilise feature extraction methods to obtain a practical data representation. Second, the data are very diverse and we usually combine several different measurements into one data set. We need to perform the data integration in order to create a data set that we are able to model machine learning and statistical methods. Third, the biochemical data are usually collinear. The collinearity indicates that there are relationships among the variables in the data. The collinearity poses problems for the standard estimation procedures. For example, the collinear data may not be invertible when estimating with least sum of squares method. Specific methods for collinear data have to be used in the biochemical data analysis.

To solve these problems with biochemical data, we reviewed the available methods to deal with feature extraction, data integration and modelling, we devised a modular data processing pipeline. The inputs to the pipeline are the raw biochemical data. The following steps are the data preprocessing and feature extraction, specific for each input data type. The data integration combines the extracted features from the preprocessed biochemical data into a standard data matrix whose rows represent the observations and the columns represent the variables (extracted features). In this step, we also add the additional information common in the biochemical analyses. The additional information can be the classification of the measured subjects (patients vs controls). The next step is the modelling. We use the data to model the response, such as the patient and control classification. The last step is the statistical assessment and inference of the modelled relationships. This pipeline was widely applied in the author's publications [J1].

The modelling of the biochemical data is difficult due to the many problems of the data. Choosing the right method for the modelling is often not straightforward procedure. Many similar methods exist and their advantages and disadvantages are not often clear enough to decide which model to use.

We devised a model comparison method for multivariate methods such as partial least squares regression and discriminant analysis or supervised principal component analysis. The method relies on the simulation of artificial data that closely mimic the properties of the actual data. The multivariate methods are assessed by bootstrap test that accounts for the relationships between the model coefficients (such as the normalisation of the loadings in the principal component analysis by the  $L^2$  norm) by testing the effect of each modelled variable on the model separately. We applied the suggested model comparison method to the comparison of the partial least squares discriminant analysis and the supervised principal component analysis in the metabolomic fingerprinting experiment setting. By considering the various measures of the model assessment, we were able to show that the supervised principal analysis is better in smaller data sets



and for the classification into more than 2 classes. The results indicate that supervised principal component analysis is beneficial in the assessment of metabolomic fingerprinting experiments.

Following the more general results, we moved to the analysis of the elemental maps measured by the laser ablation inductively coupled mass spectrometry in the analysis of the melanoma of melanoma-bearing Libechev minipigs. The analysis by laser ablation destroys the analysed biological sample. It is usually not possible to histologically annotate the biological sample measured by the laser ablation because it is substantially thicker than the tissue sections used in microscopy and the histology requires staining that may affect the content of the analysed elements in the biological sample. Therefore, two different tissue sections have to be used for the biochemical analysis.

Due to the utilisation of more than one biological sample, we have to integrate the elemental maps and the histological annotation for further processing. The results of the two methods are in form of digital images – matrices. The difference is mainly in the resolution, however, the different tissue section can be deformed. We devised a data integration procedure based on the image registration of image silhouettes. This method significantly improves the manual identification of the tissue zones [154]. The author published the method in the work [J1]. The different spatial properties of the data originating from the laser ablation combined histological annotations can be a source of considerable variation. In this and similar experiments, we want to know what is the spatial variability of the data. It is important to assess the relative sizes of the areas of interest represented in the various data projections (elemental maps, histology) in order to safely index the data for further analyses.

In order to examine the spatial properties of the data, we performed a series of experiments. We utilised the spatial covariance and the properties of the data integration procedure – the affine image transform – to estimate the minimal size of the annotated areas of interest. This way we can annotate only sufficiently large areas of consistent tissue type in the histology and be sure that the area of interest would not be severely affected by the spatial covariance in the elemental maps or result in areas of size smaller than a pixel in the elemental map. The safe indexing procedure was utilised in the work [J1]. The important variation in the elemental maps is due to the uneven thickness of the biological material. We alleviated this problem by using the elemental map of carbon as an internal standard. The normalisation of the metal elemental maps by the carbon elemental map eliminated the effect of the uneven thickness and improved the results of the modelling the work [C1] of the author. We tried to model the variation in the spatial data with clustering methods. The state-of-the-art clustering methods were able to distinguish the broad categories of tissue types. These findings were published in the conference contribution [C1] of the author. And last but not least, we examined the properties of tissues in serial tissue sections. The importance of these properties is in the planning of future experiments with more imaging methods of the tissue sections. The elastic image registration estimated the relationships between the serial tissue sections and a manual annotation of tissue structures was utilised to assess the similarity of the tissue structures in distant tissue sections. In conclusion, the tissue sections at least as far as 24  $\mu\text{m}$  apart are similar enough and can be utilised for destructive tissue imaging measurements. The final part of the thesis dealt with the statistical analysis of the differences in bioactive metals in various development stages of the melanoma in melanoma-bearing Libechev minipigs. The data were provided by the previously described data integration procedure and the annotation of the developmental stages of the melanoma respected the spatial properties of the tissue. By utilising the linear

mixed effect model assessed by non-parametric simulation-based bootstrap, we showed that the content of zinc in the growing melanoma stage of the melanoma development significantly differs from the other development stages. This finding alongside its biological and biochemical interpretation is published in the article [J1] of the author.

## 7.1 Achieved goals of the thesis

### The first goal of the thesis:

- To devise a pipeline of the processing of the integrated data with statistical and machine learning methods that will lead to results that are reliable, stable and reproducible

The building blocks of the pipeline were discussed in Chapter 3 and 4. Specifically, for the problem of the analysis of porcine melanoma by the combination of laser ablation inductively coupled plasma mass spectrometry and histology. The pipeline consists of a data integration step, which creates a layered data description. The layered data description enables indexing, subsetting and propagation of labels and similar annotations. Therefore a histology annotation can be used for selecting specific areas of interest in the elemental maps. We can then compare the selected areas – such as in the case of the comparison of the content of the bioactive metals in porcine melanoma. The pipeline is in described in Chapter 6.

### The second goal of the thesis:

- To develop a method for model comparison for models for high-dimensional biochemical data

The comparison method for multivariate data is the topic of the Chapter 5. The description of the method is in the Section 5.1. The method was applied to the problem of metabolomic fingerprinting and the comparison of the partial least squares discriminant analysis and the supervised principal component analysis. In a simulation utilising artificial data we were able to show that the supervised principal component analysis is better than the partial least squares discriminant analysis in smaller data sets and problems of classification into more than two classes (see Section 5.2).

### The third goal of the thesis:

- To develop a method for the integration of data in biochemical experiments combining spatial measurements (especially the digital microscopy and laser ablation inductively coupled plasma mass spectrometry analysis) and external sources (specialists annotations) that will generate a dataset enabling the use of standard statistical and machine learning methods

The methods for data integration are discussed in general in Section 3.2 of Chapter 3. Section 6.4.2 of Chapter 6 provides description of the method for integration of biochemical data from LA-ICP-MS imaging and histology scan. The method relies on the image registration of image silhouettes. The resulting layered data description enables indexing among the different data types and is suitable for further use by machine learning and statistical data analysis methods.



**The fourth goal of the thesis:**

- To explore the spatial structure of the tissue in serial tissue sections and consequently suggest a strategy for planning further follow-up experiments on tissue sections that can bring a better understanding to oncological processes.

The analysis of the spatial properties of the melanoma tissue is in Sections 6.5.2 and 6.5.4 in Chapter 6. The structure of the melanoma was performed by estimating elastic image transformation between a histological scans of serial tissue sections. We obtained annotations for each tissue section from the series for the specific tissue zones and the matching between the original and transformed annotations from other serial sections. The results indicate that the tissue structure is reasonably consistent for at least three consecutive tissue sections (the distance of 24  $\mu\text{m}$ ).

**The fifth and last goal of the thesis:**

- To examine the relationships between the spatial distribution of bioactive metals and the developmental stage of melanoma tissue development

The complete assessment of the relationships between the content of bioactive metals and melanoma tissue developmental stages in minipigs is described in Chapter 6. By applying the developed data integration procedure, we obtained the layered description. Through the processing pipeline, we extracted the data set of areas of interest for statistical modelling. Eventually, we were able to show that there is a significant difference in zinc content in the development stages of melanoma. The observed average zinc content in growing melanoma tissue zone and early spontaneous regression tissue zone compared to the late spontaneous regression tissue zone and fibrous tissue is significantly higher. This finding supports the hypothesis of the role of the bioactive metals in the cell cycle.

## **7.2 Future work**

Even though we may think that we performed a thorough analysis of the data, there are many unanswered questions about the data. The open questions are mainly about the possible improvements of the presented methods. For one thing, the elastic registration would benefit from further work. These are mainly the improvements in the transformation estimation by using sophisticated initialisation based on the landmarks in the images such as the bristles, or in development of complex criteria for the optimisation procedure combining the landmarks and the global measures of image similarity.

Another question is the use of the presented procedures on different data types and chemical and biochemical analyses. Many new biochemical analyses can produce spatial data that can be analysed by a similar approach. These are the method matrix-assisted laser desorption ionisation as a mass spectrometry imaging, the spatial immunochemical methods and the high pressure liquid chromatography with mass spectrometry on segments of tissue sections. The analysis of the new spatial data cannot be just a copy-and-paste, most definitely the other methods would require the development of specific approaches for the new data. The important conclusions from the analyses of the current data can be used in the planning of the new experiments, especially in minimising the manually performed tasks (such as the annotations of histological scans) by reducing the number of annotated tissue sections.

## 7.3 List of author's publications

### 7.3.1 Publications related to the topic of the thesis

#### 7.3.1.1 Journal contributions

- [J1] Jiri Anyz, Lenka Vyslouzilova, Tomas Vaculovic, Michaela Tvrdonova, Viktor Kanicky, Hajo Haase, Vratislav Horak, Olga Stepankova, Zbynek Heger and Vojtech Adam. Spatial mapping of metals in tissue-sections using combination of mass-spectrometry and histology through image registration. *Scientific reports*. 2017, 7, 40196. Available at <https://www.nature.com/articles/srep40169.pdf>.  
Cited 8 times (Scopus). IF (2018) = 4.011

The author contributed to this project by developing the data processing procedure for the diverse data in this study. The author developed the data integration procedure for the combination of histological annotations and elemental maps measured by laser ablation inductively coupled plasma mass spectrometry. The author suggested the normalisation procedure for the elemental maps that compensates for the uneven thickness of the biological samples. The author estimated the smoothness of the elemental maps for selection of sufficiently large tissue zones. The author performed the statistical analysis of the differences in the content of bioactive metals with respect to the individual differences among the animals utilising the linear mixed effect model assessed by non-parametric bootstrap.

- [J2] Helena Pelantova, Simona Bartova, Jiri Anyz, Martina Holubova, Blanka Zelezna, Lenka Maletinska, Daniel Novak, Zdena Lacinova, Miroslav Sulc, Martin Haluzik, and Marek Kuzma. Metabolomic profiling of urinary changes in mice with monosodium glutamate-induced obesity. *Analytical and bioanalytical chemistry*. 2016. 408 (2). 567–578. Available at <https://link.springer.com/article/10.1007/s00216-015-9133-0>.  
Cited 12 times (Scopus). IF (2018) = 3.286

The author contributed to this project by performing the fingerprinting analysis of the measured nuclear magnetic resonance spectra with unsupervised principal component analysis method and the supervised partial least squares method.

- [J3] Helena Pelantova, Martina Baganova, Jiri Anyz, Blanka Zelezna, Lenka Maletinska, Daniel Novak, Martin Haluzik, and Marek Kuzma. Strategy for NMR metabolomic analysis of urine in mouse models of obesity – from sample collection to interpretation of acquired data. *Journal of pharmaceutical and biomedical analysis*. 2015. 115. 225–235. Available at <https://www.sciencedirect.com/science/article/abs/pii/S0731708515300534>.  
Cited 10 times (Scopus). IF (2018) = 2.983

The author's contribution in this project was the processing of the nuclear magnetic resonance spectra and the assessment of the effects of different measurement strategies in the monosodium glutamate induced obesity mice model on the results of the fingerprinting experiment.

- [J4] Zbynek Heger, Petr Michalek, Roman Guran, Natalia Cernei, Katerina Duskova, Stepan Vesely, Jiri Anyz, Olga Stepankova, Ondrej Zitka, Vojtech Adam, and Rene Kizek. Differences in urinary proteins related to surgical margin status after radical prostatectomy. *Oncology reports*. 2015. 34 (6). 3247–3255. Available at <https://www.spandidos-publications.com/or/34/6/3247?text=fulltext>.  
Cited 3 times (Scopus). *IF* (2018) = 3.041

The author's contribution in this project was the analysis of the image electrophoreograms by the preprocessing method developed in his diploma thesis. The method extracts the brightness curves from the electrophoreogram images. These curves and the relation to disease outcome in the biological samples was modelled by the partial least squares method.

- [J5] Marketa Kominkova, Petr Michalek, Roman Guran, Natalia Cernei, Branislav Ruttkay-Nedecky, Jiri Anyz, Ondrej Zitka, Olga Stepankova, Jiri Pikula, Vojtech Adam, Miroslava Beklova, and Rene Kizek. From Amino Acids Profile to Protein Identification: Searching for Differences in Roe Deer Papilloma. *Chromatographia*. 2014. 7 (7-8). 609–617. Available at <https://link.springer.com/article/10.1007/s10337-014-2658-0>.  
Cited 1 time (Scopus). *IF* (2018) = 1.552

The author's contribution in this project was the analysis of the image electrophoreograms by the preprocessing method developed in his diploma thesis. The method extracts the brightness curves from the electrophoreogram images. These curves and the relation to disease outcome in the biological samples was modelled by the partial least squares method.

- [J6] Lenka Vyslouzilova, Sona Krizkova, Jiri Anyz, David Hynek, Jan Hrabeta, Jarmila Kruseova, Tomas Eckschlager, Vojtech Adam, Olga Stepankova, and Rene Kizek. Use of brightness wavelet transformation for automated analysis of serum metallothioneins and zinc-containing proteins by Western blots to subclassify childhood solid tumours. *Electrophoresis*. 2013. 34 (11). 1637–1648. Available at <https://onlinelibrary.wiley.com/doi/abs/10.1002/elps.201200561>.  
Cited 10 times (Scopus). *IF* (2018) = 2.754

The author contributed to this project by developing the preprocessing procedures for the electrophoreograms. This procedure extracted the brightness curves from the electrophoreogram images. These curves were then described by their wavelet decomposition, which was subsequently used for the classification of the tumours. This developed preprocessig methods were also published as the author's diploma thesis.

- [J7] Pavlina Sobrova, Lenka Vyslouzilova, Olga Stepankova, Marketa Ryvolova, Jiri Anyz, Libuse Trnkova, Vojtech Adam, Jaromir Hubalek, and Rene Kizek. Tissue specific electrochemical fingerprinting. *PLoS ONE*. 2012. 7 (11). e49654. Available at <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049654>.  
Cited 19 times (Scopus). *IF* (2018) = 2.776

The author's contribution in this project was the development of the feature extraction procedure for the Brdička curves utilising local extreme. These features were further used in the classification of rats' tissues.

### 7.3.1.2 Conference contributions

[C1] Jiri Anyz, Lenka Vyslouzilova, Vratislav Horak, Olga Stepankova, Tomas Vaculovic, and Vojtech Adam. *Examination of the Spatial Structure of Pigs' Melanoma in Tissue Sections Based on Histology and Mass Spectrometry*. In *World Congress on Medical Physics and Biomedical Engineering 2018*. 2019. 255–259.

The authors contribution is the analysis of the spatial structure on the data collected and processed in [J1]. The author utilised the unsupervised methodology for the description of the data.

[C2] Jiri Anyz and Olga Stepankova. *Using the radial basis function model for the Brdička curve fitting*. In *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*. 2015. 1–5.

The author improved the feature extraction method for the Brdička curve presented in [J7] and [C4] by using the radial basis function to model the peaks in the signal.

[C3] Jiri Anyz, and Olga Stepankova. *Visualization of Individuals Characterized by a Set of Synchronized Signals*. In *17th International Conference on Information Visualisation*. 2013. 511–516.

The author's contribution was the development of visualisation technique based on the data vases visualisation for the sets of Brdička curves with the aim to eliminate the inter-individual differences.

[C4] Lenka Vyslouzilova, Vojtech Adam, Andrea Szaboova, Olga Stepankova, Rene Kizek, and Jiri Anyz. *Brdicka curve – A new source of biomarkers*. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. 2011. 193–198.

The author's contribution in this project was the development of the feature extraction procedure for the Brdička curves utilising local extreme. These features were further used in the classification of rats' tissues.

## 7.3.2 Publications unrelated to the topic of the thesis

### 7.3.2.1 Journal contributions

[J8] Jiri Anyz, Eduard Bakstein, Daniela Dudysova, Karolina Veldova, Monika Klikova, Eva Farkova, Jana Koprivova, and Filip Spaniel. No wink of sleep: Population sleep characteristics in response to the brexit poll and the 2016 US presidential election. *Social Science & Medicine*. 2019. 222. 112–121. Available at <https://www.sciencedirect.com/science/article/abs/pii/S027795361830707X>.

$IF(2018) = 3.087$  The author performed the complete data analysis of the effects of global events on the sleep of Sleep as Android users. The statistical analysis utilised non-parametric bootstrap test controlled for several effects such as the national

differences and the weekly sleep pattern.

- [J9] Filip Spaniel, Eduard Bakstein, Jiri Anyz, Jaroslav Hlinka, Tomas Sieger, Jan Hrdlicka, Natalie Gornerova, and Cyril Hoschl. Relapse in schizophrenia: Definitely not a bolt from the blue. *Neuroscience letters*. 2018. 669 (). 68–74. Available at <https://www.sciencedirect.com/science/article/abs/pii/S0304394016302658>.

*Cited 3 times (Scopus). IF (2018) = 2.173*

The author's contribution in this project was the development and implementation of the bootstrap test for the longitudinal differences in the reported self-assessments in the ITAREPS system. The author also modelled the relationships by the negative binomial mixed effect model of the prodromal timecourse assessed by non-parametric bootstrap.

- [J10] Jan Blaha, Milos Mraz, Petr Kopecky, Martin Stritesky, Michal Lips, Michal Matias, Jan Kunstyr, Michal Porizka, Tomas Kotulák, Ivana Kolnikova, Barbara Simanovska, Mykhaylo Zakharchenko, Jan Rulisek, Robert Sachl, Jiri Anyz, Daniel Novak, Jaroslav Lindner, Roman Hovorka, Stepan Svacina, and Martin Haluzik. Perioperative tight glucose control reduces postoperative adverse events in nondiabetic cardiac surgery patients. *The Journal of Clinical Endocrinology & Metabolism*. 2015. 100 (8). 3081–3089. Available at <https://academic.oup.com/jcem/article/100/8/3081/2830268>.

*Cited 31 times (Scopus). IF (2018) = 5.605*

In this project the author contributed by performing the statistical analysis of the differences in the patients' outcomes in surgical operations in the group with and without perioperative tight glucose control.

- [J11] Zdenek Vojtech, Hana Malikova, Lenka Kramska, Jiri Anyz, Martin Syrucek, Josef Zamecnik, Roman Liscak, and Vilibald Vladyka. Long-term seizure outcome after stereotactic amygdalohippocampectomy. *Acta neurochirurgica*. 2014. 156 (8). 1529–1537. Available at <https://link.springer.com/article/10.1007/s00701-014-2126-5>.

*Cited 11 times (Scopus). IF (2018) = 1.834*

The contribution of the author to this project was the statistical assessment of the observational case study of the outcomes of the lobotomy in patients with severe epilepsy.



## References

- [1] Marco Viceconti, Peter J Hunter, and Rod D Hose. Big data, big knowledge: big data for personalized healthcare.. *IEEE J. Biomedical and Health Informatics*. 2015, 19 (4), 1209–1215.
- [2] James D Watson, and Francis HC Crick. *The structure of DNA*. In: *Cold Spring Harbor symposia on quantitative biology*. Cold Springs, New York: Cold Spring Harbor Laboratory Press, 1953. 123–131. ISBN 9780879690663.
- [3] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, and others. The sequence of the human genome. *science*. 2001, 291 (5507), 1304–1351.
- [4] Jay Shendure, and Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*. 2008, 26 (10), 1135.
- [5] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, and others. Big data: The future of biocuration. *Nature*. 2008, 455 (7209), 47.
- [6] Michael S Waterman. *Introduction to computational biology: maps, sequences and genomes*. Cleveland, OH: Chapman & Hall/CRC, 1995. ISBN 9780412993916.
- [7] Wayne W Daniel. *Biostatistics: A Foundation for Analysis in the Health Sciences (Probability & Mathematical Statistics)*. Hoboken, NJ: John Wiley & Sons, 1987. ISBN 9781118302798.
- [8] Edward H Shortliffe, and James J Cimino. *Biomedical informatics: computer applications in health care and biomedicine*. London: Springer-Verlag, 2013. ISBN 9781447144731.
- [9] Rüdiger Wirth, and Jochen Hipp. *CRISP-DM: Towards a standard process model for data mining*. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. University Park, PA: CiteSeerX, 2000. 29–39. ISBN 9781902426082.
- [10] HV Jagadish. Big data and science: Myths and reality. *Big Data Research*. 2015, 2 (2), 49–52.
- [11] Jun Zhao, Wei Wang, and Chunyang Sheng. *Data-Driven Prediction for Industrial Processes and Their Applications*. New York: Springer International Publishing, 2018. ISBN 9783319940502.
- [12] Agnieszka Smolinska, Lionel Blanchet, Lutgarde MC Buydens, and Sybren S Wijmenga. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica chimica acta*. 2012, 750 82–97.
- [13] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews genetics*. 2006, 7 (1), 55.

- [14] Johan Trygg, and Svante Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics: A Journal of the Chemometrics Society*. 2002, 16 (3), 119–128.
- [15] Svante Wold, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent laboratory systems*. 1998, 44 (1-2), 175–185.
- [16] Pawel Ciborowski, and Jerzy Silberring. *Proteomic profiling and analytical chemistry: the crossroads*. Amsterdam: Elsevier, 2016. ISBN 9780444636881.
- [17] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*. 1990, 36 (5), 961–1005.
- [18] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, and others. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013, 36 (1), 27–46.
- [19] R Brdička. Polarographic studies with the dropping mercury kathode. Part XXXI. A new test for proteins in the presence of cobalt salts in ammoniacal solutions of ammonium chloride. *Collection of Czechoslovak Chemical Communications*. 1933, 5 112–128.
- [20] Fritz Scholz, and others. *Electroanalytical methods*. Berlin, Heidelberg: Springer, 2010. ISBN 9783642029141.
- [21] Lenka Vyslouzilova, Vojtech Adam, Andrea Szaboova, Olga Stepankova, Rene Kizek, and Jiri Anyz. *Brdicka curve—A new source of biomarkers*. In: *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*. Atlanta, GA: IEEE, 2011. 193–198. ISBN 9780769534527.
- [22] I Pizeta. Deconvolution of non-resolved voltammetric signals. *Analytica Chimica Acta-Including Cumulative Indexes*. 1994, 285 (1), 95–102.
- [23] William Reusch. *Nuclear Magnetic Resonance Spectroscopy*. <https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/spectrpy/nmr/nmr1.htm>. Accessed: 2019-08-12.
- [24] Harald Günther. *NMR spectroscopy: basic principles, concepts and applications in chemistry*. Hoboken, NJ: John Wiley & Sons, 2013. ISBN 9783527330003.
- [25] Oliver Fiehn. Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*. 2002, 48 155–171.
- [26] David I Broadhurst, and Douglas B Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*. 2006, 2 (4), 171–196.
- [27] Carla Vogt, and Christopher Latkoczy. Laser Ablation ICP-MS. *Inductively Coupled Plasma Mass Spectrometry Handbook*. 2005, 228–258.
- [28] T Vaculovič, T Warchilová, Z Čadková, J Száková, P Tlustoš, V Otruba, and V Kanický. Influence of laser ablation parameters on trueness of imaging. *Applied Surface Science*. 2015, 351 296–302.
- [29] JFE Techno-Research Corporation. *Analysis of Battery Constituent Materials Based on Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS)*. <https://www.jfe-tec.co.jp/en/battery/analysis/material/la-icp-ms.html>. Accessed: 2019-08-12.

- [30] Barbara Zitova, and Jan Flusser. Image registration methods: a survey. *Image and vision computing*. 2003, 21 (11), 977–1000.
- [31] Oded Maimon, and Lior Rokach. *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2005. ISBN 9780387098234.
- [32] Ian Jolliffe. *Principal component analysis*. New York: Springer-Verlag, 2011. ISBN 9780387224404.
- [33] Herman Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*. 1975, 12 (S1), 117–142.
- [34] Jianchang Mao, and Anil K Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE transactions on neural networks*. 1995, 6 (2), 296–317.
- [35] James W Cooley, and John W Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*. 1965, 19 (90), 297–301.
- [36] David J Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*. 1982, 70 (9), 1055–1096.
- [37] Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*. 1976, 39 (2), 447–462.
- [38] Donald B Percival, and Andrew T Walden. *Spectral analysis for physical applications*. Cambridge: cambridge university press, 1993. ISBN 9780511622762.
- [39] Paul E Anderson, Michael L Raymer, Benjamin J Kelly, Nicholas V Reo, Nicholas J DelRaso, and Travis E Doom. Characterization of <sup>1</sup>H NMR spectroscopic data and the generation of synthetic validation sets. *Bioinformatics*. 2009, 25 (22), 2992–3000.
- [40] Geoffrey McLachlan, and David Peel. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons, 2004. ISBN 9780471006268.
- [41] Biserka Raspor. Elucidation of the mechanism of the Brdička reaction. *Journal of electroanalytical chemistry*. 2001, 503 (1-2), 159–162.
- [42] Reino Laatikainen, Matthias Niemitz, Willy J Malaisse, Monique Biesemans, and Rudolph Willem. A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: analysis of overlapping <sup>13</sup>C-<sup>2</sup>H multiplets of <sup>13</sup>C enriched metabolites from cell suspensions incubated in deuterated media. *Magnetic resonance in medicine*. 1996, 36 (3), 359–365.
- [43] Gareth A Morris, Hervé Barjat, and Timothy J Home. Reference deconvolution methods. *Progress in nuclear magnetic resonance spectroscopy*. 1997, 31 (2-3), 197–257.
- [44] Peter D Wentzell. Measurement errors in multivariate chemical data. *Journal of the Brazilian Chemical Society*. 2014, 25 (2), 183–196.
- [45] Peter Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*. 1967, 15 (2), 70–73.
- [46] Dennis Gabor. Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*. 1946, 93 (26), 429–441.
- [47] Wikipedia contributors. *Wavelet transform — Wikipedia, The Free Encyclopedia*. [https://en.wikipedia.org/wiki/Wavelet\\_transform](https://en.wikipedia.org/wiki/Wavelet_transform). Accessed: 2019-08-14.



- [48] Pan Du, Warren A Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*. 2006, 22 (17), 2059–2065.
- [49] Richard G Brereton. *Chemometrics: data analysis for the laboratory and chemical plant*. Hoboken, NJ: John Wiley & Sons, 2003. ISBN 9780471489788.
- [50] Bradley Efron, and Robert J Tibshirani. *An introduction to the bootstrap*. Cleveland, OH: Chapman & Hall/CRC, 1994. ISBN 9780412042317.
- [51] Mary Thompson. *Theory of sample surveys*. Cleveland, OH: Chapman & Hall/CRC, 1997. ISBN 9780412317804.
- [52] Sergey E Ilyin, Stanley M Belkowski, and Carlos R Plata-Salamán. Biomarker discovery and validation: technologies and integrative approaches. *Trends in biotechnology*. 2004, 22 (8), 411–416.
- [53] Marco D Sorani, Ward A Ortmann, Erik P Bierwagen, and Timothy W Behrens. Clinical and biological data integration for biomarker discovery. *Drug discovery today*. 2010, 15 (17-18), 741–748.
- [54] Joseph Lee Rodgers, and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*. 1988, 42 (1), 59–66.
- [55] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*. 1904, 15 (1), 72–101.
- [56] Maurice G Kendall. A new measure of rank correlation. *Biometrika*. 1938, 30 (1/2), 81–93.
- [57] Thomas M Cover, and Joy A Thomas. *Elements of information theory*. New York: John Wiley & Sons, 1991. ISBN 9780471241959.
- [58] Ronald Newbold Bracewell, and Ronald N Bracewell. *The Fourier transform and its applications*. New York: McGraw-Hill, 1986. ISBN 978-0073039381.
- [59] Kirk Wolter. *Introduction to variance estimation*. New York: Springer-Verlag, 2007. ISBN 9780387329178.
- [60] Donald E Farrar, and Robert R Glauber. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*. 1967, 49 (1), 92–107.
- [61] Wikipedia contributors. *Nuclear magnetic resonance spectroscopy — Wikipedia, The Free Encyclopedia*.  
[https://en.wikipedia.org/wiki/Nuclear\\_magnetic\\_resonance\\_spectroscopy](https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance_spectroscopy). Accessed: 2019-08-12.
- [62] Michèle Basseville. Detecting changes in signals and systems-A survey.. *Automatica*. 1988, 24 (3), 309–326.
- [63] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. Hoboken, NJ: John Wiley & Sons, 2015. ISBN 9781118675021.
- [64] James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*. 1992, 58 (1-4), 77–94.
- [65] Noel Cressie. Statistics for spatial data. *Terra Nova*. 1992, 4 (5), 613–617.
- [66] Imad A Moosa. Blaming suicide on NASA and divorce on margarine: the hazard of using cointegration to derive inference on spurious correlation. *Applied Economics*. 2017, 49 (15), 1483–1490.

- [67] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. New York: Springer-Verlag, 2001. ISBN 9780387848570.
- [68] Leo Breiman. *Classification and regression trees*. Cleveland, OH: Chapman & Hall/CRC, 2017. ISBN 9780412048418.
- [69] Leo Breiman. Bagging predictors. *Machine learning*. 1996, 24 (2), 123–140.
- [70] Leo Breiman. Random forests. *Machine learning*. 2001, 45 (1), 5–32.
- [71] Yoav Freund, Robert E Schapire, and others. *Experiments with a new boosting algorithm*. In: *Machine Learning: Proceedings of the Thirteenth International Conference*. Burlington, MA: Morgan Kaufmann Inc., 1996. 148–156. ISBN 9781558604193.
- [72] Arthur E Hoerl, and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970, 12 (1), 55–67.
- [73] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996, 267–288.
- [74] Hui Zou, and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005, 67 (2), 301–320.
- [75] Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 2000, 30 (4), 451–462.
- [76] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*. 2011, 44 (7), 1357–1371.
- [77] David L Davies, and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*. 1979, (2), 224–227.
- [78] Tilman Lange, Volker Roth, Mikio L Braun, and Joachim M Buhmann. Stability-based validation of clustering solutions. *Neural computation*. 2004, 16 (6), 1299–1323.
- [79] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*. 1999, 10 (5), 988–999.
- [80] Peter McCullagh, and John A Nelder. *Generalized linear models*. Cleveland, OH: Chapman & Hall/CRC press, 1989. ISBN 9780412317606.
- [81] Wikipedia contributors. *Kernel method — Wikipedia, The Free Encyclopedia*. [https://en.wikipedia.org/wiki/Kernel\\_method](https://en.wikipedia.org/wiki/Kernel_method). Accessed: 2019-08-14.
- [82] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain.. *Psychological review*. 1958, 65 (6), 386.
- [83] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. *Multi-column deep neural networks for image classification*. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. Washington, DC: IEEE Computer Society, 2012. 3642–3649. ISBN 9781467316118.
- [84] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*. 2006, 101 (473), 119–137.
- [85] Finbarr O’Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical science*. 1986, 502–518.

- [86] K Ruben Gabriel. Generalised bilinear regression. *Biometrika*. 1998, 85 (3), 689–700.
- [87] Ron Kohavi. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In: *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*. Burlington, MA: Morgan Kaufmann Publishers Inc., 1995. 1137–1143. ISBN 9781558603639.
- [88] Ping Zhang. Model selection via multifold cross validation. *The Annals of Statistics*. 1993, 299–313.
- [89] George Casella, and Roger L Berger. *Statistical inference*. Pacific Grove, CA: Duxbury Press, 2002. ISBN 9780534243128.
- [90] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*. 1945, 1 (6), 80–83.
- [91] NT Gridgeman. The lady tasting tea, and allied topics. *Journal of the American Statistical Association*. 1959, 54 (288), 776–783.
- [92] R Core Team. *R: A Language and Environment for Statistical Computing*. 2018. <https://www.R-project.org/>.
- [93] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*. 1961, 56 (293), 52–64.
- [94] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*. 1967, 62 (318), 626–633.
- [95] Yoav Benjamini, and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*. 1995, 289–300.
- [96] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 1979, 65–70.
- [97] Yoav Benjamini, and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*. 2001, 29 (4), 1165–1188.
- [98] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015, 67 (1), 1–48.
- [99] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. New York: Springer-Verlag, 2013. ISBN 9781461471370.
- [100] Hoai-Thu Thai, France Mentré, Nicholas HG Holford, Christine Veyrat-Follet, and Emmanuelle Comets. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics*. 2013, 12 (3), 129–140.
- [101] FN Gumedze, and TT Dunne. Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*. 2011, 435 (8), 1920–1944.
- [102] Jan De Leeuw, Erik Meijer, and Harvey Goldstein. *Handbook of multilevel analysis*. New York: Springer-Verlag, 2008. ISBN 9780387731834.
- [103] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*. 2009, 24 (3), 127–135.

- [104] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Cleveland, OH: Chapman and Hall/CRC, 2013. ISBN 9781439840955.
- [105] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 1970, 57 97-109.
- [106] Mary Kathryn Cowles, and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*. 1996, 91 (434), 883–904.
- [107] Charles J Geyer. Practical markov chain monte carlo. *Statistical science*. 1992, 473–483.
- [108] Jeremy K Nicholson, John C Lindon, and Elaine Holmes. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*. 1999, 29 (11), 1181–1189.
- [109] John C Lindon, Jeremy K Nicholson, Elaine Holmes, and Jeremy R Everett. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance: An Educational Journal*. 2000, 12 (5), 289–320.
- [110] GA Nagana Gowda, Shucha Zhang, Haiwei Gu, Vincent Asiago, Narasimhamurthy Shanaiah, and Daniel Raftery. Metabolomics-based methods for early disease diagnostics. *Expert review of molecular diagnostics*. 2008, 8 (5), 617–633.
- [111] Abdul-Hamid M Emwas, Reza M Salek, Julian L Griffin, and Jasmeen Merzaban. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*. 2013, 9 (5), 1048–1072.
- [112] John C Lindon, Jeremy K Nicholson, and Elaine Holmes. *The handbook of metabonomics and metabolomics*. Amsterdam: Elsevier, 2011. ISBN 9780444528414.
- [113] Herbert J Bernstein, John A Pople, and WG Schneider. The analysis of nuclear magnetic resonance spectra: I. Systems of two and three nuclei. *Canadian Journal of Chemistry*. 1957, 35 (1), 67–83.
- [114] David I Ellis, Warwick B Dunn, Julian L Griffin, J William Allwood, and Royston Goodacre. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*. 2007, 8 1243–1266.
- [115] Olaf Beckonert, Hector C Keun, Timothy MD Ebbels, Jacob Bundy, Elaine Holmes, John C Lindon, and Jeremy K Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*. 2007, 2 (11), 2692.
- [116] Ulf Indahl. A twist to partial least squares regression. *Journal of Chemometrics: A Journal of the Chemometrics Society*. 2005, 19 (1), 32–44.
- [117] Dongjun Chung, and Sunduz Keles. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*. 2010, 9 (1),
- [118] Maria De Iorio, Timothy MD Ebbels, and David A Stephens. Statistical techniques in metabolic profiling. *Handbook of statistical genetics*. 2008, 1 347.
- [119] LLdiko E Frank, and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*. 1993, 35 (2), 109–135.
- [120] Matthew Barker, and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*. 2003, 17 (3), 166–173.

- [121] Ewa Szymańska, Edoardo Saccenti, Age K Smilde, and Johan A Westerhuis. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*. 2012, 8 (1), 3–16.
- [122] Seoung Bum Kim, Victoria CP Chen, Youngja Park, Thomas R Ziegler, and Dean P Jones. Controlling the false discovery rate for feature selection in high-resolution NMR Spectra. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2008, 1 (2), 57–66.
- [123] David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*. 2011, 2 37–63.
- [124] Adolph Buse. Goodness of fit in generalized least squares estimation. *The American Statistician*. 1973, 27 (3), 106–108.
- [125] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*. 2012, 118 62–69.
- [126] Mariëlle Linting, Bart Jan Van Os, and Jacqueline J Meulman. Statistical significance of the contribution of variables to the PCA solution: an alternative permutation strategy. *Psychometrika*. 2011, 76 (3), 440–460.
- [127] Reza M Salek, Mahon L Maguire, Elizabeth Bentley, Denis V Rubtsov, Tertius Hough, Michael Cheeseman, Derek J Nunez, Brian C Sweatman, John N Haselden, RD Cox, and others. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat and man. *Physiological genomics*. 2007,
- [128] Christian Ludwig, John M Easton, Alessia Lodi, Stefano Tiziani, Susan E Manzoor, Andrew D Southam, Jonathan J Byrne, Lisa M Bishop, Shan He, Theodoros N Arvanitis, and others. Birmingham Metabolite Library: a publicly accessible database of 1-D  $^1\text{H}$  and 2-D  $^1\text{H}$  J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics*. 2012, 8 (1), 8–18.
- [129] Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in  $^1\text{H}$  NMR metabonomics. *Analytical chemistry*. 2006, 78 (13), 4281–4290.
- [130] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, and others. HMDB: the human metabolome database. *Nucleic acids research*. 2007, 35 (suppl.1), D521–D526.
- [131] Robert A van den Berg, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*. 2006, 7 (1), 142.
- [132] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*. 1987, 82 (397), 171–185.
- [133] Manual MatLab. The language of technical computing. *The MathWorks, Inc.* <http://www.mathworks.com>. 2012,
- [134] Animesh Acharjee, Richard Finkers, Richard GF Visser, and Chris Maliepaard. Comparison of regularized regression methods for omics data. *Metabolomics*. 2013, 3 (3), 1.

- [135] V Horak, K Fortÿn, V Hruban, and J Klaudy. Hereditary melanoblastoma in miniature pigs and its successful therapy by devitalization technique.. *Cellular and molecular biology (Noisy-le-Grand, France)*. 1999, 45 (7), 1119–1129.
- [136] C Geffrotin, V Horak, F Crechet, Y Tricaud, C Lethias, S Vincent-Naulleau, and P Vielh. Opposite regulation of tenascin-C and tenascin-X in MeLiM swine heritable cutaneous malignant melanoma. *Biochimica et Biophysica Acta (BBA)-General Subjects*. 2000, 1524 (2-3), 196–202.
- [137] Jan Borovanskÿ, Vratislav Horák, Milan Elleeder, Karel Fortÿn, Nico PM Smit, and Adriana M Kolb. Biochemical characterization of a new melanoma model—the minipig MeLiM strain. *Melanoma research*. 2003, 13 (6), 543–548.
- [138] Silvia Vincent-Naulleau, Catherine Le Chalony, Jean-Jacques Leplat, Stéphan Bouet, Christiane Bailly, Alain Spatz, Philippe Vielh, Marie-Françoise Avril, Yves Tricaud, Joseph Gruand, and others. Clinical and histopathological characterization of cutaneous melanomas in the melanoblastoma-bearing Libechov minipig model. *Pigment cell research*. 2004, 17 (1), 24–35.
- [139] Daniela Planska, Monika Burocziova, Jan Strnadel, and Vratislav Horak. Immunohistochemical analysis of collagen IV and laminin expression in spontaneous melanoma regression in the melanoma-bearing Libechov minipig. *Acta histochemica et cytochemica*. 2015, 48 (1), 15–26.
- [140] J Sa Becker, MV Zoriy, C Pickhardt, N Palomero-Gallagher, and K Zilles. Imaging of copper, zinc, and other elements in thin section of human brain samples (hippocampus) by laser ablation inductively coupled plasma mass spectrometry. *Analytical chemistry*. 2005, 77 (10), 3208–3216.
- [141] J Sabine Becker, Miroslav Zoriy, Andreas Matusch, Bei Wu, Dagmar Salber, Christoph Palm, and J Susanne Becker. Bioimaging of metals by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). *Mass spectrometry reviews*. 2010, 29 (1), 156–175.
- [142] Christopher Latkoczy, Yves Müller, Patrik Schmutz, and Detlef Günther. Quantitative element mapping of Mg alloys by laser ablation ICP-MS and EPMA. *Applied surface science*. 2005, 252 (1), 127–132.
- [143] Ludwik Halicz, and Detlef Günther. Quantitative analysis of silicates using LA-ICP-MS with liquid calibration. *Journal of Analytical Atomic Spectrometry*. 2004, 19 (12), 1539–1545.
- [144] Johannes T van Elteren, Norman H Tennent, and Vid S Šelih. Multi-element quantification of ancient/historic glasses by laser ablation inductively coupled plasma mass spectrometry using sum normalization calibration. *Analytica Chimica Acta*. 2009, 644 (1-2), 1–9.
- [145] Jeffrey S Crain, and David L Gallimore. Inductively coupled plasma-mass spectrometry of synthetic elements: <sup>99</sup>Tc. *Applied spectroscopy*. 1992, 46 (3), 547–549.
- [146] Daniel A Frick, and Detlef Günther. Fundamental studies on the ablation behaviour of carbon in LA-ICP-MS with respect to the suitability as internal standard. *Journal of Analytical Atomic Spectrometry*. 2012, 27 (8), 1294–1303.
- [147] Curtis T Rueden, Johannes Schindelin, Mark C Hiner, Barry E DeZonia, Alison E Walter, Ellen T Arena, and Kevin W Eliceiri. ImageJ2: ImageJ for the next generation of scientific image data. *BMC bioinformatics*. 2017, 18 (1), 529.
- [148] Melissa Linkert, Curtis T Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, Josh Moore, Carlos Neves, Donald MacDonald,

- and others. Metadata matters: access to image data in the real world. *The Journal of cell biology*. 2010, 189 (5), 777–782.
- [149] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*. 1950, 37 (1/2), 17–23.
- [150] Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*. 1954, 5 (3), 115–146.
- [151] Christopher KI Williams, and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Cambridge, MA: MIT Press, 2006. ISBN 026218253X.
- [152] *The Gaussian Processes Web Site*.  
<http://www.gaussianprocess.org/>. Accessed: 2019-02-25.
- [153] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*. New York: Springer-verlag, 1978. ISBN 9780387953663.
- [154] David Riesop, Alfred V Hirner, Peter Rusch, and Agnes Bankfalvi. Zinc distribution within breast cancer tissue: A possible marker for histological grading?. *Journal of cancer research and clinical oncology*. 2015, 141 (7), 1321–1331.
- [155] Franco P Preparata, and Michael I Shamos. *Computational geometry: an introduction*. New York: Springer-Verlag, 2012. ISBN 9780387961316.
- [156] Marco Corvi, and Gianluca Nicchiotti. Multiresolution image registration. *International Conference on Image Processing - Proceedings*. 1995, 3 224–227.
- [157] Jan Kybic, and Michael Unser. Fast parametric elastic image registration. *IEEE transactions on image processing*. 2003, 12 (11), 1427–1442.
- [158] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*. 2006, 25 (11), 1451–1461.
- [159] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*. 2002, 14 849–856.
- [160] R Core Team. *R: A Language and Environment for Statistical Computing*. 2019.  
<https://www.R-project.org/>.
- [161] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*. 2004, 11 (9), 1–20.
- [162] Tai Sing Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence*. 1996, 18 (10), 959–971.
- [163] Brian B Avants. *ANTsR: ANTs in R: Quantification Tools for Biomedical Images*. 2019. R package version 0.4.9.
- [164] Karl Rohr, Mike Fornefett, and H Siegfried Stiehl. Spline-based elastic image registration: integration of landmark errors and orientation attributes. *Computer Vision and Image Understanding*. 2003, 90 (2), 153–168.
- [165] GL Fisher, KL McNeill, LE Spitler, and LS Rosenblatt. Serum copper and zinc levels in melanoma patients. *Cancer*. 1981, 47 (7), 1838–1844.
- [166] MR Ros-Bullon, P Sanchez-Pedreno, and JH Martinez-Liarte. Serum zinc levels are increased in melanoma patients.. *Melanoma research*. 1998, 8 (3), 273–277.

- [167] Raphael Gorodetsky, Jacob Sheskin, and Arye Weinreb. Iron, copper, and zinc concentrations in normal skin and in various nonmalignant and malignant lesions. *International journal of dermatology*. 1986, 25 (7), 440–445.



# Appendix A

## List of abbreviations

### A.1 Biochemical and other measurement methods

CPMG	Carr-Purcell-Meiboom-Gill sequence
CT	Computational tomography
LA-ICP-MS	Laser ablation - Inductively coupled plasma - Mass spectrometry
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
PET	Positron emission tomography

### A.2 Biological and biochemical terminology

DNA	Deoxyribonucleic acid
ESR	Early spontaneous regression
FT	Fibrous tissue
GMT	Growing melanoma tissue
LOD	Limit of detection
LSR	Late spontaneous regression
MELiM	Melanoma-bearing Libechev Minipig
SR	Spontaneous regression

### A.3 Machine learning and statistical methods

ANN	Artificial neural network
AR	Autoregressive model
ARMA	Autoregressive moving average model
AUC	Area under the ROC
EM	Expectation maximisation
GLM	Generalised linear model
GLMM	Generalised linear mixed effect model
LASSO	Least absolute shrinkage and selection operator
LM	Linear model
LMM	Linear mixed effect model
MA	Moving average model
MCMC	Markov chain Monte Carlo
MLRT	Maximum likelihood ratio test
MODWT	Maximum overlap discrete wavelet transform
PCA	Principal component analysis
PLS	Partial least squares
PLSDA	Partial least squares discriminant analysis

PLSR	Partial least squares regression
REMLRT	Restricted maximum likelihood test
ROC	Receiver operating characteristic
STFT	Short-time Fourier transform
SSD	Sum of squared differences
SVD	Singular value decomposition
WT	Wavelet transform