Norwegian University
of Life Sciences

Master’s Thesis 2016 30ECTS
Department of Mathematical Sciences and Technology

Assessment of Sparse Multi-Block Partial Least Squares Regression Model Performance in Analysis of High-Dimensional Phenotypic Data

Tor Einar Møller
Lecturer programme

Acknowledgements

The semester spent working with this thesis has been the most rewarding of all my years studying. Finally being able to spend ample time on an academic discipline where mathematics, physics and biology are combined has kept me happily occupied in the study halls for the entire semester. I am very grateful to NMBU for allowing such a thesis to be the final product of a lecturer student.

I extend my first thanks to Karoline, who has been understanding of my workload and hours spent on the thesis throughout, and still denied me from talking about it too much at home.

I would also like to thank Neil Davey of the NMBU Writing Centre for guiding me in how to word and phrase my thesis in a concise and deliberate manner, ShareLATEXfor their excellent online writing program, my fellow students at the study hall for their good mood and coffee brewing skills, and Peter Lasch of Robert Koch-Institut, Berlin, Germany, for providing the data sets used in the thesis.

A very special thanks goes to researcher Valeria Tafintseva at NMBU, who, although not formally my supervisor, has kindly furnished me with MATLAB codes and helped me utilising and tailoring them. Without her help, this thesis would not have been possible.

My final thanks goes to my supervisors, Associate Professor Volha (Olga) Shapaval and Professor Achim Kohler. Olga for providing comments and feedback on the thesis' biological aspects. Achim for his always constructive and encouraging feedback and for all the time he spent helping me write a thesis I can be happy with. He has exceeded any expectations I had for the role of a supervisor beforehand.

NMBU, Ås, 12.12.2016

Tor Einar Møller

Abstract

FTIR and Raman spectroscopy, and MALDI-TOF mass spectrometry are emerging technologies for multidimensional phenotyping of microorganisms. While FTIR and Raman both represent a full metabolic fingerprint, MALDI spectra mainly represent the microbe's ribosomal protein composition.

All methods are used for microbial identification, both by the food industry and in the clinical laboratory, but direct comparison of them by integration into the same statistical model is lacking in scientific literature. To compare the three methods, we applied a Sparse MultiBlock PLSR (SMBPLSR) routine capable of analysing all data types simultaneously.

We present results indicating that this SMBPLSR method can be used to establish connections between the metabolic fingerprint of FTIR and Raman spectra, and ribosomal protein expression in MALDI-TOF data, and that the method to a large extent enables identification of samples on the strain level. Furthermore, we show that the SMBPLSR method can be used to indicate how phenotypic response to varied growth temperature is ascribed to certain types of biomolecules. Finally, we present results showing that different types of phenotypic data are treated differently by the SMBPLSR method. Grouping among variables or samples in FTIR and Raman data is achieved by a different set of latent variables than in grouping in MALDI data. The sensitivity and wealth of information obtainable from the SMBPLSR method makes it a viable complement to the already existing multivariate analysis methods.

Samandrag

FTIR- og Raman-spektroskopi, og MALDI-TOF massespektrometri, er alle framvaksande teknologiar brukt til multidimensjonal fenotyping av mikroorganismar. Medan FTIR or Raman gjev eit fullt metabolsk fingeravtrykk, er det ribosomal proteinkomposisjon som kjem til uttrykk i MALDI.

Alle desse metodane brukast for å identifisera mikrober, både i matvareindustrien og i kliniske laboratorier, men ei direkte statistisk samanlikning av metodane manglar i den vitskaplege litteraturen. For å bøte på mangelen, brukte me ei Sparsomleg MultiBlokk PLSR-metode (SMBPLSR) som kunne analysera alle datatypane samstundes.

Me synar fram resultat som indikerer at SMBPLSR-metoden kan nyttast til å etablera koplingar mellom metabolsk fingeravtrykk i FTIR- og Raman-spektra på den eine sida, og ribosomalt proteinuttrykk i MALDI-TOF data på den annan. SMBPLSR-metoden kan i utstrekt grad identifisera prøver på stammenivå. Vidare syner me at SMBPLSR-metoden kan brukast til å indikera korleis fenotypisk respons på ulike veksttemperaturar kan tilskrivast spesifikke typar biomolekyl. Til slutt presenterast resultat som syner at dei ulike slaga fenotypiske data handsamast svært ulikt av SMBPLSR-metoden. Grupperingar av variablar eller prøver i FTIR- og Raman-data tilskrivast heilt andre latente variablar enn tilsvarande grupperingar i MALDI-data. Følsemda til og vellet av informasjon som kan framskaffast frå SMBPLSR-metoden gjer han til eit levedyktig tilskot til allereie eksisterande multivariate analysemetodar.

Contents

Chapter 1

Introduction

1.1 Motivation

Technology and methods for classification, identification and characterisation of microorganisms has seen tremendous development in recent decades. The main modern way of microbial classification (phylogenetic placement of an organism new to science) is by comparing its genotype - genetic code - with that of already known genera, species or strains. The various techniques employed to do so are called genotyping techniques [1]. The numerous scientific and commercial techniques currently in use, include single nucleotide polymorphism whole-genome sequencing (SNP WGS) and pulsed-field gel electrophoresis (PFGE) [2]. Genotyping techniques make it possible to accurately classify novel microorganisms and place them in the phylogenetic tree, or identify already known ones correctly [2].

The demand for such techniques has never been higher. As occurrences of bacterial multiresistance to antibiotics are on the rise worldwide [3], methods that ensure rapid and adequate classification, identification and/or characterisation (the measure of a microbe's biological constituents) is becoming increasingly important to detect and avoid nosocomial infections (attracted in hospitals or, literally, 'under care' [2]) and epidemics. Similar needs are expressed by the food industry for the prevention of food spoilage [4, 5, 6, 7]. For this purpose, the current genotyping techniques tend to be considered too slow and cumbersome [2, 8].

Generally working at lower cost, offering higher throughput, and requiring less elaborate sample preparation than genotyping, phenotyping is a viable and widely used alternative [2, 9, 10, 11]. The phenotype is the expressed character traits of an organism [8]. A few examples of phenotypic expression in bacteria are growth and replication rates, and production levels of certain chemical compounds [8], often proteins or measures of protein levels whose values are known to be species-specific [2, 12]. The use of such phenotypic techniques, or phenotyping, is to identify microorganisms by examining these traits, also called biomarkers [2].

In this study, we employ three phenotyping techniques, namely Fourier-Transform InfraRed (FTIR) spectroscopy, Fourier-Transform Raman (Raman) spectroscopy and Matrix Assisted Laser Desorption Ionisation Time-of-Flight (MALDI or MALDI-TOF) mass spectrometry, to study these biomarkers. FTIR and Raman both, though in different ways, utilise the specificity of resonance frequencies in the covalent chemical bonds between the atoms of a molecule [13]. MALDI makes use of the fact that an ionised molecule subjected to a magnetic field will, depending on its mass-to-charge ratio, require a certain time-of-flight to reach a detector [14].

Being emerging technologies, the three techniques all have their individual properties, advantages and drawbacks. The data gathered from FTIR and Raman correspond to a metabolic fingerprint (not to be confused with the spectral region), indicating levels of both proteins, lipids and carbohydrates in the sample, their interrelated and absolute levels being the predictive factor in identification or characterisation [9]. In contrast, MALDI measures the prevalence of specific ribosomal biomarker proteins that can be used to characterise, identify, or even characterise microorganisms [12].

As discussed by Dieckmann et al. [2], MALDI possesses a twofold major advantage as a characterisation and identification technique compared to FTIR and Raman. Because of its less strict standardisation requirements in sample preparation, for instance with growth conditions and type of sample preparation, it has become the technique of choice when creating spectral databases of microorganisms, as well as for identifying pathogenic microbes that must be neutralised prior to analysis [15].

On the contrary, strict standardisation requirements for both FTIR and Raman discourage the creation of similar inter-institutional, comprehensive data bases due to the perils of methodological inaccuracy and variation. As of today, most suppliers of spectroscopic equipment include access to their own spectroscopic databases - for MALDI. Notwithstanding, this dominance mostly restricts itself to macromolecules. There are voluminous spectral databases for small organic and inorganic molecules, such as the NIST Chemistry Webbook [16], or as a compiled list of such, at Internetchemistry.com [17].

The paramount distinction between the techniques is their predictive ability; how well do they predict a new set of data and what is their mode of prediction? Phenotypic methods require training before they can classify new data [18]; genotypic techniques do not. This training entails calibration of the model parameters employed for the actual analysis of the data set.

Dieckmann et al. [2] argue that while MALDI shows better discriminative ability at the species level, the intraspecies discrimination of FTIR is superior because of the full fingerprint it provides. Specifically, MALDI mainly detects ribosomal proteins that display less strain diversity [11]. On the other hand, FTIR gathers information from cell surface lipid and carbohydrate levels, both of which show higher intraspecific variation [2].

A common way to analyse MALDI, FTIR and Raman data is by using multivariate methods [19, 20, 21, 22, 23]. In this study, we exploit these methods' ability to handle collinearity within the data set. This enables us to assess the differences and similarities in predictive ability and discrimination by the various phenotypic methods discussed so far. Principal Component Analysis (PCA) is the foundation upon which most of the modern multivariate methods are built, and indeed the foundation for our model extensions as well. Both PCA and our extended version base the prediction on latent variables [24].

The inherent spectral composition of FTIR and Raman spectra us to extract data blocks that only encompass certain specific types of biomolecules, such as lipids or proteins. Separating the total spectrum of these signals enable us to examine in-depth how each type of biomolecule contributes to the overall model prediction [24].

In particular, we want to explore and examine how multiblock methods can be used in combination with sparse variable selection to enhance the predictive ability of phenotypic data sets, and how they can be used to detect phenotypic responses to growth conditions, specifically variation in growth temperature. Arrondo and Goñi [25] showed that FTIR possessed such discriminative capability, but this has to our knowledge never before been done with this kind of data.

In a sparse model, only a subset of the variables are selected. Sparsity is implemented by soft thresholding as proposed by Lê Cao et al. [26]. This implementation allows selecting, to a higher extent, biologically meaningful variables, easing the interpretation of the results [20, 26].

Multiblock (MB) models are effective in several different ways. MB gives the opportunity to 'zoom' by dividing a data set into blocks and then examining the results of each individual block [27, 28]. Furthermore, several types of data can be included in the same model [29], improving predictive ability and enhancing interpretation.

The scope of this study *is to evaluate the ability of sparse multiblock methods to establish a connection between metabolic fingerprints by FTIR and Raman, and protein expression by MALDI, and to compare phylogenetic differences and similarities assessed by different phenotypic methods.*

The multivariate methods used in this study were first proposed by Karaman et al. [29], who implemented them for a purpose similar to ours, but for different data types. The statistical methods have mainly been implemented into MATLAB$^{\circledR}$ R2016a (The MathWorks, Inc., Natick, Massachusetts, US) code by Professor Achim Kohler and researcher Valeria Tafintseva, both of NMBU. The data sets have been kindly provided by

Peter Lasch of Robert Koch-Institut, Berlin, Germany.

1.2 Structure of thesis

The remaining chapters of this thesis are structured in the following way: In Chapter 2, the bacterial strains examined in this study are presented. In Chapter 3, we give a short presentation of the physical principles behind the three phenotyping techniques employed in this study. In Chapter 4, the methods of preprocessing, pretreatment and statistical analysis employed in this study are explained and briefly discussed. Chapter 5 gives a short walk-through on how the statistical models are validated, and some information about the code and statistical script used. In Chapters 6 and 7, we present the main results, one chapter per data set. A discussion of the results is included in each chapter. A general discussion of the statistical methods used and other relevant topics is given in Chapter 8. Finally, the main conclusions and outlook are summarised in Chapter 9.

In the electronic version of this thesis, all links to chapters, sections, references, figures and tables are clickable at their numbers.

Chapter 2

Materials: Two data sets of bacterial strains

In this thesis, two different data sets were examined. The first data set was used to test SMBPLSR methods for identification purposes. It consists of 17 bacterial strains, 16 of which belong to the *Klebsiella oxytoca* species, and one belonging to the *K. pneumoniae* species, totalling 51 samples. This data set is referred to as the *Klebsiella*, 'identification' or 'identificatory' data set. Further description is given in Section 2.1.

The second data set was used for assessment of the discriminative abilities of SMBPLSR methods for varying growth temperatures. This data set consists of two bacterial strains, of the *Bacillus subtilis* and *Escherichia coli* species, cultivated at four different temperatures. Given three samples per species-per temperature, it totals 24 samples. It is referred to as the *Bacillus/Escherichia*, 'experimental' or 'experimental design' data set because it concerns a variable of experimental design. For further description, see Section 2.2.

Both data sets were kindly provided by Peter Lasch of Robert Koch-Institut, Berlin, Germany.

Both data sets include data from FTIR and Raman spectroscopy, and MALDI-TOF-MS. Furthermore, the experimental data set includes low-mass and medium-mass MALDI data, whereas the identificatory data set only includes medium-mass MALDI data. An explanation of the column label abbreviations in the MALDI data sets can be found at [30].

2.1 *Klebsiella oxytoca* and *K. pneumoniae*

The bacteria *K. oxytoca* is an opportunistic, pathogenic, facultative anaerobic species of Gram-negative, rod-shaped and non-motile bacteria of the *Enterobacteriaceae* family, and a close relative to *K. pneumoniae*, responsible for pneumonia in humans [3, 31]. Under natural circumstances, both *K. oxytoca* and *K. pneumoniae* are found on human mucosal surfaces, such as the throat, and in our surroundings [2].

2.1.1 Occurrence in the clinical environment

The bacterial species *K. oxytoca* and *K. pneumoniae* have long been associated with so-called nosocomial (attracted in hospital, or, literally: 'under care') infections, particularly attacking immunocompromised patients [2]. Nordmann, Cuzon and Naas [3] reported that during the first decade of the 21^{st} century, resistance to antibiotics in several strains of both species has been found, first in the USA, then around the world. These strains showed reduced susceptibility to antibiotics (carbapenems or other β-lactams) frequently used in the treatment of Gram-negative bacteria in hospitals.

Both *K. oxytoca* and *K. pneumoniae* are known to contaminate soap and hand sanitiser dispensers in clinical and community environments [32]. In fact, the majority of *K. oxytoca* strains in this study are from two lots (I and II) of hand soap produced in Greece in 2013, see Table 2.1. In particular, sanitation products with high water content are prone to bacterial contamination [2]. Contamination can be divided into two categories: Intrinsic contamination during manufacturing and/or shipping, and extrinsic contamination during use at a health care facility [33]. Correspondingly, increased awareness of the risk of infection stemming from water-containing sanitation products has sparked an increased popularity for waterless, alcohol-based hand sanitisers [32].

Thorough screening for and identification of microbial pollutants during manufacturing and shipping could potentially eliminate intrinsic contamination altogether. Rapid and precise strain identification and characterisation, and subsequent isolation of patients infected by multiresistant strains of *K.* spp. is critical in order to prevent or minimise hospital outbreaks.

2.1.2 Strains and specifications

All samples in this data set, including their background, preparation and measurement, are the same as the ones used and described by Dieckmann et al. [2]. An overview of the strains is shown in Table 2.1. For FTIR and Raman, all strains consist of three biological replicates (samples), each of which is the average of three technical replicates. For MALDI, each sample is the average of two technical replicates.

Table 2.1: Overview of strains of the *Klebsiella* genus included in this data set. The year of isolation is not known for all strains.

Strain	Country of origin	Year of isolation	Source
PHS-890[a]	Greece	2013	Hand soap, lot 1
PHS-891[a]	Greece	2013	Hand soap, lot 1
PHS-892[a]	Greece	2013	Hand soap, lot 1
PHS-893[a]	Greece	2013	Hand soap, lot 1
PHS-894[a]	Greece	2013	Hand soap, lot 1
PHS-895[a]	Greece	2013	Hand soap, lot 2
PHS-896[a]	Greece	2013	Hand soap, lot 2
PHS-897[a]	Greece	2013	Hand soap, lot 2
PHS-898[a]	Greece	2013	Hand soap, lot 2
PHS-899[a]	Greece	2013	Hand soap, lot 2
CB4063[a]	Germany	1995	Child, enteritis
CB4074[a]	Germany	1995	Child, enteritis
CB4072[a]	Germany	1995	Child, enteritis
CCUG 15788[a]	Sweden		Environmental
Oman 61[a]	Oman	2011	Clinical isolate
ATCC 13182[a]	USA		Pharyngeal tonsil
ATCC 25926[b]	Belgium		Human blood

[a] *K. oxytoca*, [b] *K. pneumoniae*, subsp. Ozaenae

2.2 *Bacillus subtilis* and *Escherichia coli*

B. subtilis is an endospore-forming, motile and gram positive bacteria commonly found in soil, but also in the intestinal tract of humans [34]. The species is harmless, non-toxic and not dangerous to humans [34]. *E. coli* is a gram-negative bacterium commonly found in the intestines of warm-blooded animals [35]. Most strains of *E. coli*, including the one we examine in this study [36] are harmless, but some are pathogenic and may infect humans who consume contaminated food or liquids [35].

2.2.1 Strains and specifications

All samples in this data set were cultivated aerobically for 24 hours. Two bacterial strains were measured: *E. coli* K12 DSM 3871 [36] and *B. subtilis* DSM347 [37]. Four growth temperatures were selected: 25, 30, 37 and 43^{o}C. Three biological replicates were produced for each temperature, each biological replicate being the average of three technical replicates. All strains and samples were measured by FTIR, Raman and MALDI-TOF ICMS (Intact Cell Mass Spectrometry [38]). MALDI measurements were taken both in the low mass ($\frac{m}{z}$ between 500 and 3200) and medium mass ($\frac{m}{z}$ between 3200 and 20000) ranges.

All measurements were carried out by Maren Stämmler and the data set was kindly provided by Peter Lasch, both of Robert Koch-Institut, Berlin, Germany. The strains were prepared essentially equally as those described in the previous section (i.e. by Dieckmann et al. [2]), but with certain differences in growth conditions. Further enquiries regarding the details of how the strains were prepared should be addressed to Peter Lasch. An overview of the strains and their growth temperature is shown in table 2.2.

Table 2.2: Overview of *B. subtilis* and *E. coli* strains included in this data set.

Strain	Growth temperature [°C]
DSM 347[a]	25
DSM 347[a]	30
DSM 347[a]	37
DSM 347[a]	43
K12 DSM 3871[b]	25
K12 DSM 3871[b]	30
K12 DSM 3871[b]	37
K12 DSM 3871[b]	43

[a] *B. subtilis*, [b] *E. coli*

Chapter 3

Spectroscopic methods

FTIR, Raman and MALDI are widely used phenotyping techniques, providing rapid and accurate identification of microorganisms [2, 11]. This is done through cultivation and subsequent measurement of prepared samples. The measured samples provide a data set which can be visualised as a continuous spectrum.

The scope of this chapter is to outline the working principles behind FTIR, Raman and MALDI. In Section 3.1, we give a brief motivation for the usage of FTIR, a very short introduction to the physical principles behind the technique, and finally a discussion of advantages and drawbacks of the technique. The procedure is then repeated for Raman in Section 3.2 and MALDI in Section 3.3.

3.1 Fourier Transform InfraRed (FTIR) spectroscopy

Whenever electromagnetic radiation interacts with matter, one of three events are likely to occur: transmittance, absorbance or scattering. The working principle behind FTIR is detecting differences and variations in these three [13]. The recorded absorbance spectrum of a molecule or microorganism can be read as a fingerprint and compared to already existing spectra for identification. This routine is both quick and cheap, and efforts are ongoing to fully automate the cultivation and analysis of samples [5]. FTIR is routinely being used in both the food industry [4, 5, 6, 11] and clinically [2, 39].

3.1.1 Infrared Radiation

The infrared spectrum succeeds the spectrum of visible light at wavelengths around 700 nm and continues up to around 1 mm [40]. However, in FTIR-spectroscopy, a unit called the wavenumber is commonly preferred to wavelength because of the more convenient scale it uses. The wavenumber $\tilde{\nu}$ is the reciprocal of wavelength λ such that

$$\tilde{\nu} = \frac{1}{\lambda}$$

Its unit is given in cm^{-1} according to customs. Corresponding IR wavenumbers range from 12 820 cm^{-1} to 33 cm^{-1}. The set of wavenumbers is further divided into three categories: near, mid and far IR [40]. Most of the fundamental modes of molecular vibration are localised in the mid IR region [13], see subsections 3.1.2 and 3.1.3 for further elaboration. Consequently, this is the most interesting part of the IR spectrum to examine. The mid IR region is the set of wavenumbers ranging from 4000 cm^{-1} to around 500 cm^{-1}.

The relationship between wave energy and wavelength is given as $E = \frac{hc}{\lambda}$ where λ is the wavelength, h is Planck's constant and c the speed of light. This amounts to IR-photons possessing energy from 1.7 eV to 1.24 meV; energy and wavelength are inversely proportional. Mid IR photon energy varies from 0.50 eV to 0.062 eV, or around 8 to 40 kJ/mol.

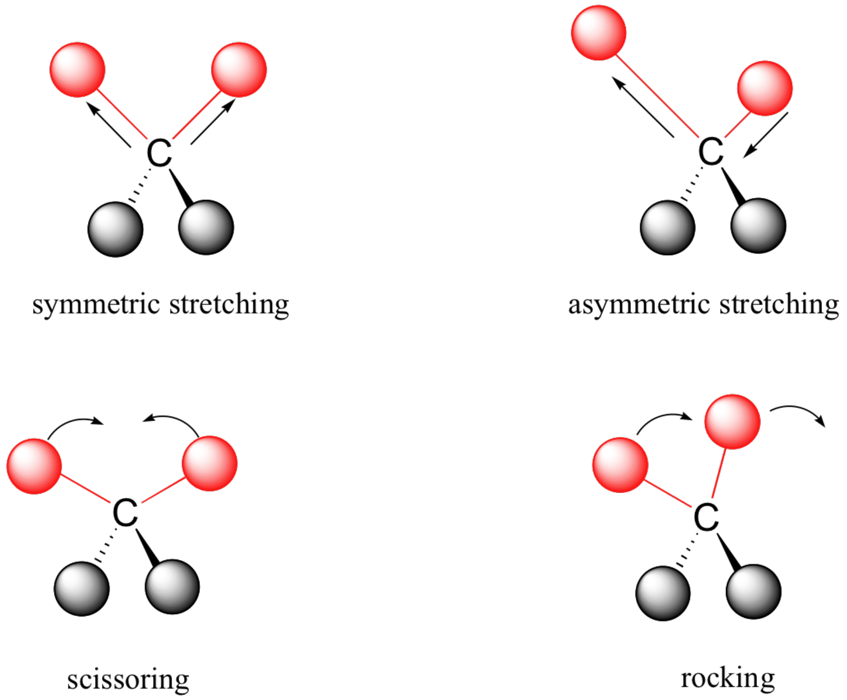
3.1.2 Molecular response to IR-radiation

The mid IR photon energy of 0.062 eV (8 kJ/mol) to 0.50 eV (40 kJ/mol) corresponds to vibration in covalent bonds in organic molecules [13]. Typical modes of vibration are stretching, scissoring and rocking, shown in Figure 3.1 below. Whenever a photon in the IR energy range interacts with a molecule and that molecule has a bond with a resonance frequency (or equivalently an energy gap between two states) that matches the photon frequency (energy), absorption occurs that induces vibration in one of the molecule's covalent bonds [13].

Both the mode of movement, the hybridisation of the bond (single, double or triple) and the species of atoms connected contribute to determine the specific frequency of the vibration [13]. For instance, the bending of a conjugated carbon-carbon double bond is always found at 1615 cm^{-1} and, conversely, a spectrum displaying a peak at that wavenumber indicates the presence of such a bond.

Because of the direct mode of excitation, FTIR spectroscopy is only concerned with asymmetric movement that causes a net displacement or change in dipole moment within the molecule [13]. This is complementary to Raman spectroscopy, which registers changes in molecular polarisation, i.e. changes in the shape, size or orientation of the atom's electron cloud. These are only detectable when the movement is symmetric [40]. More on this in Section 3.2.

The specificity of the resonance frequencies enable the identification of molecules [13] and even microorganisms [4, 6, 7, 9] based on inspection, interpretation and analysis of



symmetric stretching

asymmetric stretching

scissoring

rocking

Figure 3.1: Illustration of bending and stretching modes of covalent bonds, from [41].

their spectra. Figure 3.2 below shows the raw FTIR spectra of the *Klebsiella* data set analysed in this study. The various high-absorbance peaks can be attributed to specific covalent bonds that correspond to common constituents of biomolecules.

Figure 3.2 uses absorbance as y-axis unit. Transmittance may also be used. Absorbance and transmittance are values obtained during spectroscopy from comparing the intensity of emitted and detected radiation for various wavenumbers, denoted $I_0(\tilde{\nu})$ and $I_0(\tilde{\nu})$, respectively. Transmittance is defined simply as the coefficient of detected light, or

$$T(\tilde{\nu}) \equiv \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{3.1}$$

while absorbance is defined as the negative logarithm of T, or

$$A(\tilde{\nu}) \equiv -log(T) = -log\frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{3.2}$$

for some wavenumber $\tilde{\nu}$. These spectral entities in equations (3.1) and (3.2) constitute the fundamentals of signal preprocessing, which is explained in Section 4.1.

The various peaks of Figure 3.2 are often divided into two regions or categories: the functional region and the fingerprint region [13]. The functional region encompasses the spectrum from 4000 cm^{-1} to around 1500 cm^{-1} and indicates the presence of organic functional groups. For instance, the rather narrow peak around 2950 cm^{-1} indicates

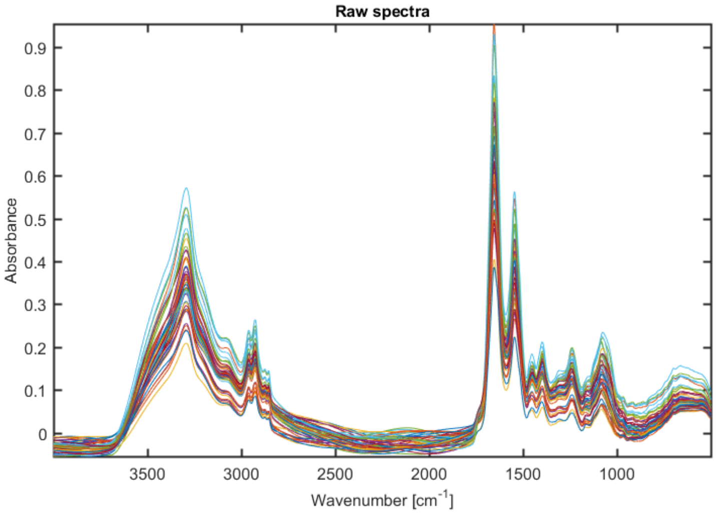


Figure 3.2: Raw spectra of FTIR data from the *Klebsiella* data set analysed in this study.

CH-stretching [13].

The fingerprint region, from 1500 cm^{-1} to 500 cm^{-1}, is densely populated with peaks, and it is difficult to meaningfully interpret this region visually [13]. Still, it plays a crucial role in the identification of both molecules and microorganisms [6] because its highly specific composition of peaks serves as a fingerprint for the molecular or cell composition. However, because absorbance levels usually vary among samples and replicates, visual identification becomes slightly more auspicious for larger data sets than for single samples. Nevertheless, decisive identification and characterisation must be done through statistical analysis, for instance by comparison with spectral libraries [13, 40].

3.1.3 Absorbance regions of biomolecules

The main constituents of any microbe are proteins (amino acids), carbohydrates and lipids (fatty acids). The most abundant chemical bonds usually belong to molecules associated with these three categories, making it relatively easy to statistically identify the dominant biomolecular constituents of a sample.

Proteins are characterised chiefly by their amide I bonds (C=0) at around 1700 - 1550 cm^{-1} [25, 42].

The strongest identifiers of carbohydrates are found in the fingerprint region, between 1400 and 900 cm^{-1} [19, 43]. This region can be further subdivided into the regions 1150-900 cm^{-1} (C-O and C-C stretching) and 1400-1150 cm^{-1} (O-C-H, C-C-H and C-O-H bending).

Fatty acids occupy several spectral subregions, including the primary domains of carbohydrates and proteins [20]. However, the tails of most lipids are full of CH=CH and CH_2-CH_2 bonds. These bonds show characteristic stretching vibration around 3200 - 2800 cm^{-1} [6].

A second inspection of Figure 3.2 indicates that signals localised in the aforementioned spectral regions indeed dominate the FTIR spectrum of microorganisms.

3.2 Fourier-Transform Raman (Raman) spectroscopy

Both FTIR and Raman are vibrational spectroscopic methods, that is, they detect vibrations in the covalent bonds between atoms [40]. But while FTIR spectroscopy uses infrared radiation to induce vibration directly, Raman spectroscopy employs monochromatic laser radiation in the regions of near infrared (NIR), visible light and the near ultraviolet regions to excite electrons to a virtual energy state. The electrons then relax back into an intermediate vibrational energy state before complete relaxation [2, 10].

3.2.1 Scattering in Raman spectroscopy

Electronic scattering, the interactive process between radiation and matter causing the photon to change direction, comes in two forms: Elastic and inelastic [44]. Elastic scattering is also known as Rayleigh scattering and inelastic scattering is also known as Compton scattering. In elastic scattering, the energy of the photon remains unaltered. In inelastic scattering, some energy is either lost or gained in the process. In the case where the net change of photon energy is negative, the excitation/relaxation process is called Stokes or Stokes Raman scattering. If the net change is positive, the process is called Anti-Stokes or Anti-Stokes Raman scattering. Scattering properties of Raman spectroscopy and comparison to direct IR absorption is shown in Figure 3.3.

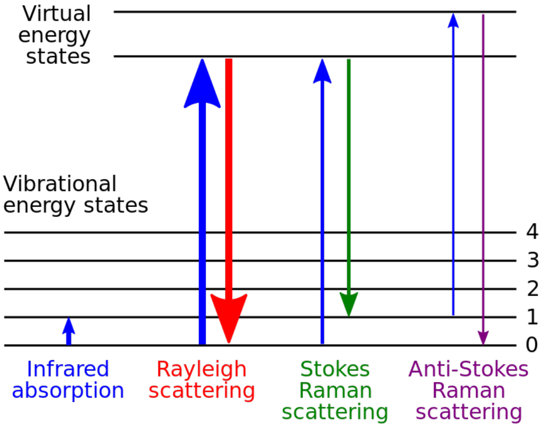


Figure 3.3: Schematic depiction of elastic and inelastic scattering associated with Raman spectroscopy, compared with IR absorption. From [45].

3.2.2 Detecting Raman scattering

The photon energy lost during Stokes scattering is temporarily retained in the covalent bonds between the atoms of the irradiated molecule. However, the now-excited covalent bond will, after a short retention time, emit its surplus energy as an infrared photon [40]. The retention time must be so long that no fluorescence interferes with the Raman signal, or the experiment must be designed so that there is no fluorescence at all. This is because only a tiny fraction of photons, about one in a million, are excited/relaxed via a Stokes/Anti-Stokes pathway. In fact, sample fluorescence is 10^7 times stronger than Raman scattering [40].

One way of ensuring adequate retention time is to use a NIR pulse laser as the source of radiation. NIR energy pulses create sufficiently large energy gaps between the virtual energy states and the vibrational energy states [10, 44]. Also, the number of transitions in the NIR region are fairly few, helping to further reduce or diminish the fluorescence.

There exists a trade-off between strength of Raman effect ($P_{scattered} \propto \frac{I_0}{\lambda^4}$) and fluorescence from an incoming laser beam. Among the common laser frequencies are 532 nm (red/green light), 785 nm (NIR) and 1064 nm (NIR). There is less fluorescence at 1064 nm, but the Raman signal is 16 times weaker than for a 532 nm laser, all other conditions remaining the same [46].

3.2.3 Measuring a Raman spectrum

Like FTIR, Raman spectra are reported in wavenumbers $\tilde{\nu}$ with cm^{-1} as units. However, the x-axis unit is not the wavenumber, but rather the Raman shift, denoted $\Delta\tilde{\nu}$ and derived using the formula

$$\Delta\tilde{\nu} = \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_{vib}}\right) \times \frac{10^7 nm}{cm}$$

where λ_0 is the incident wavelength (from ground state to virtual state) and λ_{vib} is the vibrational Raman spectrum wavelength. For example: Given an incident radiation wavelength of 1064 nm ($\tilde{\nu} = 9399$ cm^{-1}), we detect an IR photon with a wavelength of 1292 nm ($\tilde{\nu} = 7740$ cm^{-1}). Using the above formula, we can determine that the Raman shift for the given photon was $\Delta\tilde{\nu} = 1659$ cm^{-1}, a typical wavenumber corresponding to the amide I bond in proteins. The Raman shift is not dependent on the incident laser beam, only on the absorbing chemical bond [47]. The more photons of this exact wavelength detected, the more intense the signal. Figure 3.4 shows the raw Raman spectra of the *Klebsiella* data set examined in this study. Note the use of arbitrary units on the y-axis, as opposed to absorbance/transmittance for FTIR.

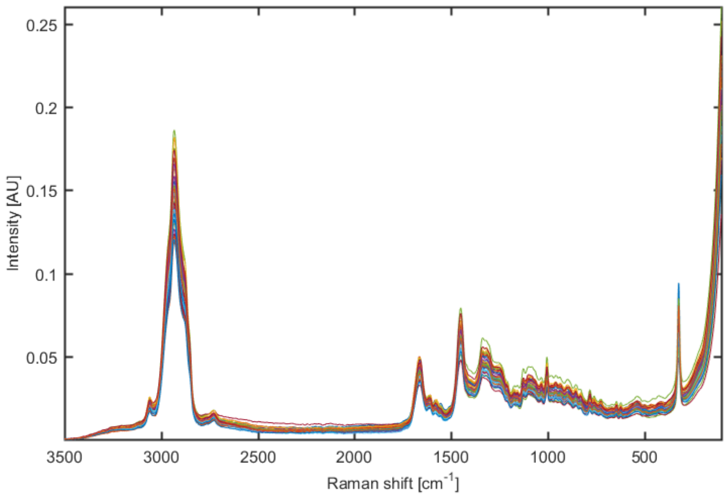


Figure 3.4: Raw spectrum of Raman spectrum for the *Klebsiella* data set under study.

3.3 Matrix Assisted Laser Desorption Ionisation Time-Of-Flight (MALDI-TOF) mass spectrometry

The scope of mass spectrometry is to produce and subsequently weigh ions of compounds for information about molecular structure. The mass spectrometer is a device constructed to achieve this goal [13]. There are several different types of MS that can be used for both qualitative and quantitative analysis [48]. Once such way of qualitative analysis is to classify microorganisms by detecting certain species or even strain specific signal molecules present in the sample [23]. This type of analysis has enjoyed widespread application in both clinical practice [12, 49, 50, 51, 52] and the food industry [11, 38, 53].

One of the major advantages of MS is that, given a reference database of spectra, it can accurately classify microorganisms, sometimes down to strain level [11, 49] in a matter of minutes [48]. What's more, Wenning et al. [11] showed that MALDI compares to and for some classes of bacteria even outperforms FTIR spectroscopy in terms of identification rate.

3.3.1 Principles of mass spectrometry

The core principle of MS is to ionise the molecules or molecular fragments in a sample and then measure the time it takes for them to reach a detector [13, 39]. This time-of-flight will depend on its mass-to-charge $\left(\frac{m}{z}\right)$ ratio. The net charge of most compounds is usually $+1$ because ionisation is achieved by bombardment of protons, so the correct mass will be $m - 1$. Overall, calculating $\frac{m}{z}$ is a straight-forward operation.

One of the main distinctions in MS is the one between so-called hard and soft ionisation

[13]. Hard ionisation involves extensive molecular fragmentation into ions, often to separate the components or functional groups of the molecule under examination [13]. Needless to say, for large biomolecules weighing up to hundreds of kilodaltons (kDa) [39], such fragmentation is undesirable, if not to say completely pointless. MALDI is a soft ionisation technique, causing little molecular fragmentation [11] and preserving the original molecular ion M^+ or fragmenting it into a few, large compounds.

The macro-ions are then subjected to an electric field and hit the detector after a certain time-of-flight, depending on $\frac{m}{z}$. Ions of similar $\frac{m}{z}$ ratio hit the detector simultaneously and their number relative to the total number of detections determine their relative intensity in the recorded spectrum [13]. The results are then plotted in a coordinate system with $\frac{m}{z}$ on the x-axis and relative intensity on the y-axis. As an example, the raw spectra of MALDI data from the *Klebsiella* data set is shown in Figure 3.5.

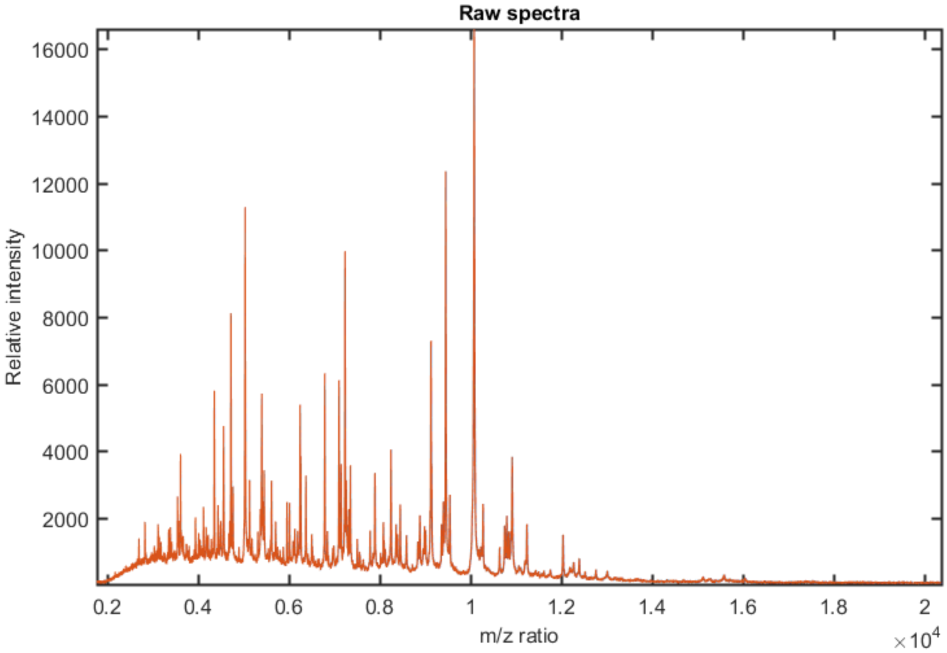


Figure 3.5: Raw spectra of MALDI data from the *Klebsiella* data set.

Each peak in Figure 3.5 corresponds to a specific ion, for example a protein fragment [11]. Knowledge of the protein's likely fragmentation pattern, combined with fragmentation data from a spectrum, serves to identify the protein or its abundance relative to other significant proteins; such levels or occurrences are genus, species of even strain specific - a fingerprint [11, 39].

3.3.2 Biomolecules detected by MALDI

Whereas FTIR and Raman both address the entire metabolic fingerprint of the biomolecule, MALDI-TOF delimits itself to primarily measure protein levels [12, 53], predominantly the abundance of ribosomal proteins in the $\frac{m}{z}$ range between 4 and 13 kDa. In complete cell MS [2, 54, 55]. These are known to display clear inter-species differences, but lower strain-specific variation than the aggregate of biomolecules measured by FTIR and Raman [2], and has previously achieved a lower percentage of correct identification at the strain level [11]. Nevertheless, Sandrin et al. [54] concluded in their review of quantitative profiling by MALDI on the strain level, that while results between studies and for different techniques varied significantly, accurate identification is possible.

Chapter 4

Multivariate methods

In spectroscopy, one often finds oneself with data sets where the number of variables far exceed the number of samples [19, 56]. It is not uncommon to have a sample size of less than one hundred, while the number of variables reaches several thousand. The data are modelled as a matrix $\mathbf{X}$, with samples as rows and variables as columns. Yet, because the variance within each column vector $\mathbf{x}_n$ is rather small due to the similarity of each sample, it is possible to explain most of the variance of the entire data set in only a few dimensions. This variance is described through principal components (PCs), and determining these PCs is the purpose of PCA [57].

But before analysis, the data set needs to be both preprocessed to remove unwanted mechanical, physical or biological noise from the data, and pretreated to emphasise its important information [58].

In Section 4.1, a motivation for data preprocessing is given and the Extended Multiplicative Signal Correction (EMSC) algorithm for data preprocessing of FTIR and Raman is presented. In Section 4.2, we give a short description of the modes of data pretreatment used in this study. Finally, in Section 4.3, the data analysis methods used in this thesis are explained.

4.1 Data preprocessing

The raw spectra shown in Figures 3.2 and 3.4 above provide a lot of information, some of which is important and some which can be disregarded [59]. Sifting out the relevant information is essential to be able to statistically analyse and identify, classify or characterise the molecule or microorganism at hand. Proper preprocessing of the sample spectrum facilitates successful further analysis, even though the measurement process might be burdened by inaccuracy.

Possible sources of error in measurement are numerous in spectroscopic analysis.

Zimmermann and Kohler [60] list a few: Chemical pollution of the sample in analysis or simply a poor sample; a multitude of atmospheric pollutants such as CO_2 or water; unwanted or unaccounted-for refraction or scattering effects from the spectroscopy itself may lower the quality of data acquired from the measurement; instrumental errors and anomalies. Naturally, one seeks to find some way of identifying and removing or at least reducing these impairments before analysing the data, while at the same time preserving or even accentuating important biochemical information [61]. In short: The ultimate purpose of preprocessing is to identify, separate and attenuate or remove unwanted non-biochemical information from the data set.

4.1.1 The Savitzky-Golay filter

First proposed in 1964, the Savitzky-Golay [62], or convolution [63] filter, is a numerical tool for preprocessing spectral data. The procedure relies on two conditions: That the data points be evenly distributed along some axis, and that the function be continuous. Both conditions are met by FTIR spectroscopy [60].

Each data point on the spectrum represents the intersection of some absorbance value along the y-axis and some wavenumber along the x-axis. Centred in this point, a number (denoted m) of neighbouring points are selected, the aggregate of which $(2m + 1)$ is called the window size. A polynomial of predetermined order (often quadratic or quartic [60]) is then fit to these points by least squares regression. The neighbouring points are given weighting coefficients based on proximity to the centre point, facilitating more thorough smoothing of the spectrum. Finally, the spectral derivative of this polynomial is taken. This can be done any number of times, but is usually done twice to preserve and accentuate peaks, see Figure 4.1 below. Then the window moves to the next data point and repeats the process.

Other important effects of taking the derivative are to remove additive variations in the data set and altering the signal-to-noise-ratio [56].

For the SG algorithm to work properly, a full set of neighbouring points must exist for all data points. This means that unless some preventive measure is taken, the m first and last border points of the spectrum will be truncated. A shortcoming of the original model, this was improved upon by Gorry [63]. However, this is not often a problem with FTIR as the spectrum consists of several thousand data points, the optimal window size is but a tiny percentage of this [60], and the most valuable information is rarely found on the edges [13].

Selecting an appropriate window size is still important in order best to obtain the data desired. Zimmermann and Kohler [60] show that the optimal window size depends on both the nature of the spectrum under analysis and the order of polynomial chosen:

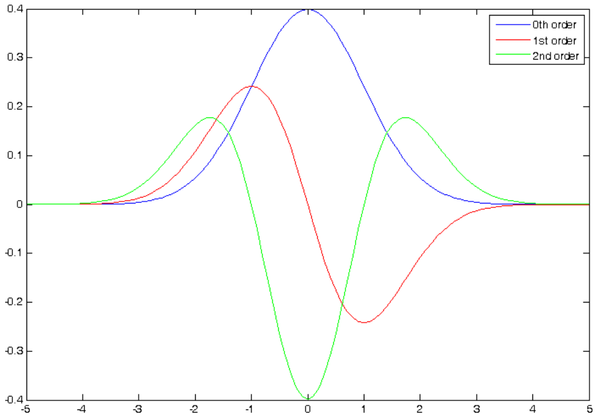


Figure 4.1: The 2^{nd} order derivative, unlike the 1^{st}, preserves the peak x-coordinate. From [64]. Arbitrary axis units.

A higher-order polynomial, say quartic, is more prone to overfitting, i.e. also including random error or noise in the model than for a lower-order polynomial, say quadratic. A large window size will smooth high-frequency noise better than a small, but may at the same time also remove important information contained in narrow peaks.

4.1.2 Deriving the EMSC model

EMSC, as described in [56, 59, 65], is an extension to the Multiplicative Signal Correction (MSC) model developed in the 1980s [66]. The extension was developed because MSC sometimes failed to adequately correct spectra with higher-order baseline effects [65, 67]. EMSC takes this problem into account and provides a model for correcting spectra using a higher-order polynomial. Rinnan, van den Berg and Ellingsen [61] elaborate further on the similarities and differences between MSC and EMSC. EMSC also normalises the spectrum with respect to some reference spectrum.

Absorbance and transmittance

When radiation of intensity $I_0(\tilde{\nu})$ hits a cell or other kind of sample, not all of the radiation is transmitted through the sample; some might be scattered or absorbed. The definitions of transmittance and absorbance are shown in equations (3.1) and (3.2) above.

$A(\tilde{\nu})$ is also given by Beer-Lamberts law [13], stating that

$$A(\tilde{\nu}) \approx k(\tilde{\nu})cb \tag{4.1}$$

where, in a transparent sample with a single absorbing component, $k(\tilde{\nu})$ is the characteristic absorbtivity of said component for a certain wavenumber, c is the component concentration in the sample and b the thickness of the sample, commonly referred to as optical pathlength (unit: cm). Generalised for N absorbing components, (4.1) becomes

$$A(\tilde{\nu}) \approx b \sum_{n=1}^{N} c_n k_n(\tilde{\nu}) \tag{4.2}$$

where c_n is the concentration and $k_n(\tilde{\nu})$ the absorptivity of sample species n, respectively.

Scattering

Absorption is, as previously mentioned, not the only physical effect affecting the total transmittance; there is also scattering. A fraction of I_0 might be scattered by the sample to avoid detection. This effect is denoted by σ and inserted in (3.2) such that

$$A(\tilde{\nu}) \approx -log \frac{\sigma I(\tilde{\nu})}{I_0(\tilde{\nu})} = -log(\sigma) - log \frac{I(\tilde{\nu})}{I_0(\tilde{\nu})} \tag{4.3}$$

The $-\log(\sigma)$ term in (4.3) corresponds to providing (4.2) with an additive term s so that

$$A(\tilde{\nu}) \approx b \left[\sum_{n=1}^{N} c_n k_n(\tilde{\nu}) \right] + s \tag{4.4}$$

describing the superposition of absorbances for N chemical components. The optical path length b is assumed to be equal for all components. c_n is the concentration and $k_n(\tilde{\nu})$ characteristic, wavelength-dependent absorptivity, both of component n. s is the additive baseline effect.

Averaging spectra

Because of the assumed collinearity between samples, the average $\bar{x}(\tilde{\nu})$ of all spectra is a good approximation to each sample. As such, every sample spectrum can be expressed by its deviation $\Delta k_n(\tilde{\nu})$ from the mean:

$$k_n(\tilde{\nu}) = \bar{x}(\tilde{\nu}) + \Delta k_n(\tilde{\nu}) \tag{4.5}$$

Summing up for all components in the sample, (4.5) and (4.4) combine to become

$$A(\tilde{\nu}) = b \left[\sum_{n=1}^{N} c_n \bar{x}(\tilde{\nu}) + \sum_{n=1}^{N} c_n \Delta k_n(\tilde{\nu}) \right] + s \tag{4.6}$$

Assuming that $\sum_{n=1}^{N} c_n$ constitutes all possible spectra, the sum of these must add up to one:

$$\sum_{n=1}^{N} c_n = 1 \tag{4.7}$$

Finally, applying (4.7) to (4.6), produces

$$A(\tilde{\nu}) = b\left[\bar{x}(\tilde{\nu}) + \sum_{n=1}^{N} c_n \Delta k_n(\tilde{\nu})\right] + s \tag{4.8}$$

which may also be stated as the the statistical model

$$A(\tilde{\nu}) = b\bar{x}(\tilde{\nu}) + s + e(\tilde{\nu}) \tag{4.9}$$

where s is the additive scattering effect, b the multiplicative optical path length and the residue $e(\tilde{\nu}) \equiv b\sum_{n=1}^{N} c_n \Delta k_n(\tilde{\nu})$. The unknown parameters s and b are estimated by least squares regression.

Correcting the spectra based on (4.9), we get

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - s}{b} \tag{4.10}$$

also known as the Multiplicative Signal Correction (MSC). It accounts for additive and multiplicative effects, but only in the case that the baseline is linear. This does not always hold true and calls for an extension of (4.9) to include baseline polynomials of higher order:

$$A(\tilde{\nu}) = b\bar{x}(\tilde{\nu}) + s + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + \ldots + d_n\tilde{\nu}^n + e(\tilde{\nu}) \tag{4.11}$$

Here, as with MSC, the coefficients d_1, d_2, $\ldots$, d_n are all calculated by least squares regression. The corresponding extension of (4.10) is as follows:

$$A_{corr}(\tilde{\nu}) = \frac{A(\tilde{\nu}) - s - d_1\tilde{\nu} - d_2\tilde{\nu}^2 - \ldots - d_n\tilde{\nu}^n}{b} \tag{4.12}$$

where $A(\tilde{\nu})$ is the wavenumber-dependent absorption, s is the additive scattering effect, d_n are coefficients, $\tilde{\nu}^n$ are terms of a higher-order baseline-correcting polynomial, and b is the multiplicative optical path length. This is the final version of the EMSC model. In the case when the model is extended to a quadratic polynomial, it is often referred to as the basic or standard EMSC model [65].

4.1.3 Notation in pretreatment and analysis

In the following, matrices are denoted in capital bold-face (e.g. $\mathbf{X}$), vectors in lower case bold-face (e.g. $\mathbf{t}$) and scalars in lower case italics (e.g. b). Furthermore, principal components are written as subscripts, so that for A principal components, $\mathbf{T}_A = [\mathbf{t}_1\mathbf{t}_2 \ldots \mathbf{t}_a \ldots \mathbf{t}_A]$. Block numbers are written as superscripts: The concatenated matrix of B blocks, $\mathbf{X}$, may also be written as $\mathbf{X} = [\mathbf{X}^1\mathbf{X}^2 \ldots \mathbf{X}^b \ldots \mathbf{X}^B]$. A superscript t denotes the transposed of the matrix or vector in question.

4.2 Data pretreatment

After preprocessing, it is necessary to adjust the data set so that it is optimally prepared for the subsequent statistical analysis. This step is also called pretreatment [58].

Van den Berg et al. [58] divide data pretreatment into three classes, which are applied in the order they are listed here: Centring, scaling and transformation. Centring is simply subtraction of the mean from each column of the data set. Scaling is an important part of pretreatment in a number of phenotyping techniques, but is rarely performed on any of the data types we use in this study and hence only mentioned. Transformation is nonlinear conversion of the data to correct for heteroskedasticity, or fluctuations in variance between data blocks.

In this study, (S)PLSR and (S)MBPLSR models are subject to different types of pretreatment. Whereas both the descriptor and response data of single-block PLSR routines in this study are only mean-centred, MBPLSR routines employ a more elaborate pretreatment. The descriptor data is pretreated block-wise. Each block $1 \leq b \leq B$ is first centred according to

$$\mathbf{X}^b = \mathbf{X}_0^b - \mathbf{1}\bar{\mathbf{x}}^b$$

where $\mathbf{X}^b$ is the mean-centred descriptor data block b, $\mathbf{X}_0^b$ is the non-centred descriptor data, $\mathbf{1}$ is a matrix whose dimensions are equal to $\mathbf{X}^b$ but whose elements are all 1, and $\bar{\mathbf{x}}^b$ is the mean value of each column of $\mathbf{X}^b$. Then each block is divided by its Frobenius norm

$$\hat{\mathbf{X}}^b = \frac{\mathbf{X}^b}{\|\mathbf{X}^b\|_F}$$

where $\hat{\mathbf{X}}^b$ is the pretreated block and $\mathbf{X}^b$ is the mean-centred block. The response data undergoes centring in the MBPLSR routines.

4.3 Data analysis

Preprocessed and pretreated data usually show a high degree of collinearity - they are intended to do so. Despite the number of samples, most of the total variance in the set - both in $\mathbf{X}$ and $\mathbf{Y}$ - can often be explained adequately in only a few dimensions by methods based on latent variables [19]. These methods function by computing the latent variables, or principal components (PCs), that explain the most of this variance. Analysis of the latent variables is commonly referred to as Principal Component Analysis (PCA).

Several algorithms may be used to perform PCA, of which Singular Value Decomposition (SVD) and Non-Iterative PArtial Least Squares (NIPALS) are perhaps the most common. One key difference between the two is that SVD extracts all the PCs at once, while NIPALS does so sequentially; one at the time [68]. Wu, Massart and de Jong [68] also showed that for a sufficiently large amount of variables, i.e. several hundred, the SVD algorithm required far less computation (in Megaflops) than sequential algorithms to obtain the preselected number of PCs. In this study, SVD is used as a basis to extract starting vectors for the various PLSR algorithms employed. Scores and loadings are then calculated by NIPALS algorithms.

A ubiquitous prerequisite in the upcoming section is exact row-to-row correspondence between the descriptor data $\mathbf{X}$ and response data $\mathbf{Y}$; columnar correspondence is not required.

4.3.1 SVD

Shlens [57] provides an excellent explanation of the SVD algorithm. The general formula for SVD is [68, 69]

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^t, \tag{4.13}$$

where $\mathbf{X}$ is an $n \times m$ matrix of rank r, n and m being the number of samples and variables, respectively. $\mathbf{V} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r]$ is the set of orthonormal eigenvectors with corresponding eigenvalues $\lambda_1, \ldots, \lambda_r$ for the symmetric matrix $(\mathbf{X}^t\mathbf{X})\hat{\mathbf{v}}_i = \lambda_i\hat{\mathbf{v}}_i$. $\sigma_i \equiv \sqrt{\lambda_i}$ are called the singular values of the decomposition. These are the r first diagonal entries of Σ. $\mathbf{U} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_r]$ is the set of orthonormal vectors defined by $\hat{\mathbf{u}}_i = \frac{1}{\sigma_i}\mathbf{X}\hat{\mathbf{v}}_i$.

In this study, SVD is used to extract start weights to be used in the NIPALS algorithm for PLSR and MBPLSR (described below), guaranteeing a unique solution. Equation (4.13) above shows the basic version of SVD for PCA. For PLSR, with both descriptor data $\mathbf{X}$ and response data $\mathbf{Y}$, one can perform SVD on $\mathbf{M} \equiv \mathbf{X}^t\mathbf{Y}$ to obtain starting weights for both sets ($\mathbf{w}_a$ for $\mathbf{X}$ and $\mathbf{c}_a$ for $\mathbf{Y}$). These vectors are $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{v}}_1$, the first columns of $\mathbf{U}$ and $\mathbf{V}^t$, respectively. The tildes signify normalisation to a length of 1.

4.3.2 Partial Least Squares Regression (PLSR)

PLSR is the basic statistical model employed to study the relationship between a model based on a set of descriptor data $\mathbf{X}$, and a set of unknowns, or response data $\mathbf{Y}$ [28, 70]. This prediction is achieved through the model

$$\mathbf{X} = \mathbf{1}_x\bar{\mathbf{x}} + \mathbf{T}_A\mathbf{P}_A^t + \mathbf{E}_{X,A}$$
$$\mathbf{Y} = \mathbf{1}_y\bar{\mathbf{y}} + \mathbf{T}_A\mathbf{Q}_A^t + \mathbf{E}_{Y,A}$$

where $\mathbf{1}_X$ and $\mathbf{1}_Y$ are matrices of dimensions equal to $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean of $\mathbf{Y}$ and $\mathbf{X}$, $\mathbf{T}_A$ is the score matrix, $\mathbf{Q}_A$ are the loadings of $\mathbf{Y}_A$ and $\mathbf{P}_A$ of $\mathbf{X}$, and $\mathbf{E}_{X,A}$ and $\mathbf{E}_{Y,A}$ are the residuals after A principal components have been calculated. Notice how the score vectors $\mathbf{T}_A = [\mathbf{t}_1 \ldots \mathbf{t}_A]$ are common to both $\mathbf{X}$ and $\mathbf{Y}$, assuring mutual relevance. Figure 4.2 provides an illustration of the algorithm, explained below.

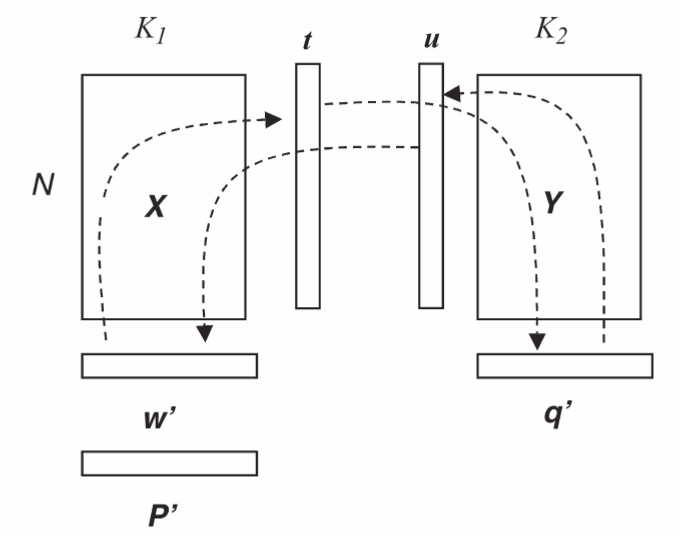


Figure 4.2: Illustration of the PLSR algorithm. From [19], page 343.

The algorithm for normal PLSR in this study goes as follows:
Step 1. Initialisation.

- Select two sets of variables: $\mathbf{X}$ and $\mathbf{Y}$.
- Pretreat data according to single-block PLSR routine presented in Section 4.2.

Step 2. Calculation of PCs. The following procedure is run for each PC to be calculated. Each PC is computed from the residual matrices $\mathbf{E}_{X,a-1} = \mathbf{X}$ and $\mathbf{E}_{Y,a-1} = \mathbf{Y}$ from the previous PC-calculation.

- Decompose $\mathbf{X}^t\mathbf{Y}$ to $\mathbf{U}\Sigma\mathbf{V}^t$ by SVD and extract the first column $\mathbf{w}$ of $\mathbf{U}$ as initial loading weights.

- Calculate **X** scores: $\tilde{\mathbf{t}} = \mathbf{X}\mathbf{w}$
- Normalise the score vector: $\mathbf{t} = \frac{\tilde{\mathbf{t}}}{\|\tilde{\mathbf{t}}\|}$.
- Calculate **Y** loadings: $\mathbf{q} = \mathbf{Y}^t\mathbf{t}$.
- Calculate **X** loadings: $\mathbf{p} = \mathbf{X}^t\mathbf{t}$.

Because SVD (ball point 1) extracts the **w** with the highest variance to begin with, this particular algorithm converges after the first iteration.

Step 3. Deflation. Höskuldsson [70] showed that in PLS routines, deflation on **X** or **Y** alone would suffice and that this would indeed enhance the computational speed of the algorithm. Nevertheless, deflation procedures for both matrices are performed in this study.

Y is deflated according to the formula

- $\mathbf{Y}_{a+1} = \mathbf{Y}_a + \mathbf{t}_a\mathbf{q}_a^t$.

Recall from the PCA algorithm that some loading vector **p** is required to deflate **X**. This first needs to be defined through the projection of **X** onto **t**:

- $\mathbf{p} = \frac{\mathbf{X}^t\mathbf{t}}{\mathbf{t}^t\mathbf{t}}$.

Then, for PC a, **X** is deflated according to the formula

- $\mathbf{X}_{a+1} = \mathbf{X}_a - \mathbf{t}_a\mathbf{p}_a^t$

just as in PCA. The algorithm is repeated on $\mathbf{X}_{a+1}$ until the desired amount of PCs are obtained.

4.3.3 Sparse PLSR (SPLSR)

For large matrices of spectroscopic data where variance between samples is small, it is not always necessary to include all variables in the calculation of PCs. Usually, careful selection of the most informative variables will suffice. Karaman et al. [42] showed that sparse PLSR, even with only a fraction of the original variables selected, still performed well and proved stable in selecting the relevant variables.

Sparsity

Based on SVD where $\mathbf{M} = \mathbf{X}^t\mathbf{Y} = \mathbf{V}_X\Delta\mathbf{V}_Y$, the optimisation problem raised by imposing sparsity is formulated by Zou et al. [71] as

$$\min_{w,c} \|\mathbf{M} - \mathbf{w}\mathbf{c}^t\|_F^2 + 2\lambda|\mathbf{w}| \tag{4.14}$$

where $\mathbf{w}$ and $\mathbf{c}$ are the loading weights of $\mathbf{X}$ and $\mathbf{Y}$, respectively, $\| \|_F$ is the Frobenius norm and $2\lambda|\mathbf{w}|$ is the penalty function for so-called soft thresholding. Note that the threshold value λ of this section is different from the eigenvalue λ_i of Section 4.3.1 on SVD above.

Soft thresholding

Lê Cao et al. [26] set their soft thresholding function to

$$ST_\lambda = sign(z)(|z| - \lambda)_+ \tag{4.15}$$

where $z \in \mathbb{R}$ are the elements of a loading weight vector, λ is a threshold value and the function $(z)_+ \equiv \max(0, z)$. (4.15) implies that any element $z \leq \lambda$ is forced to value 0 and $z > \lambda$ gets λ subtracted from its value. The number of variables deemed 'worthy' decreases as λ is increased; $\lambda = 0$, conversely, corresponds to ordinary PCA/PLSR. In principle, this soft thresholding can be applied to both $\mathbf{X}$ and $\mathbf{Y}$ in (4.14), but because of the often low number of variables in $\mathbf{Y}$, it is usually applied to $\mathbf{w}$ only [29]. A visualisation of how soft thresholding impacts on variables selection for sparse PLSR algorithms is shown in Figure 4.3. The dotted lines represent the threshold, the blue line the original sample spectrum and the orange line the effect of soft thresholding on the selected variables.

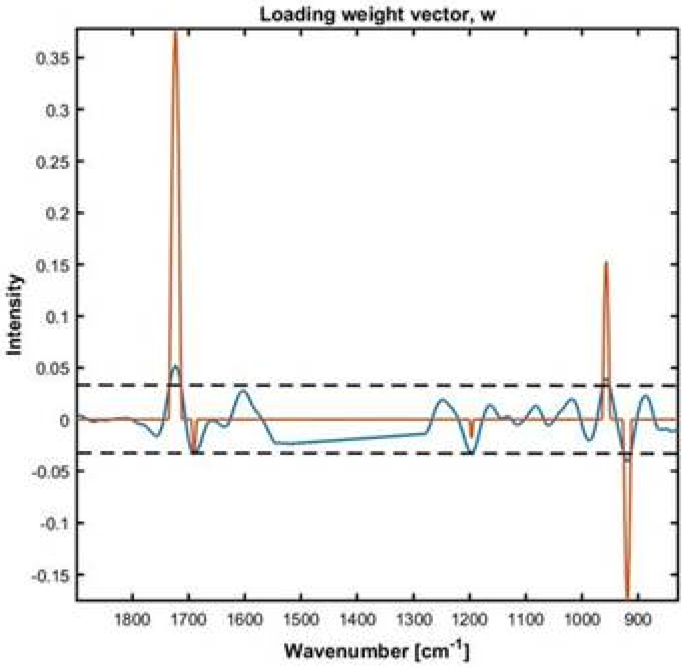


Figure 4.3: Visualisation of soft thresholding. Figure courtesy of Valeria Tafintseva, researcher at NMBU.

The algorithm

In their 2014 article, Karaman et al. [29] present two analogous methods for calculating SPLSR, one by themselves and one by Lê Cao et al. [26]. The MATLAB-functions in this study use the second one.

Step 1: Initialisation.

- Select two sets of preprocessed and pretreated variables: $\mathbf{X}$ and $\mathbf{Y}$.
- Define $\mathbf{M}_{a-1} = \mathbf{X}^t_{a-1}\mathbf{Y}_{a-1}$.
- Decompose $\mathbf{M}_{a-1}$ by SVD to obtain $\mathbf{V}_X \Delta \mathbf{V}^t_Y$.
- Assign the first singular vectors of $\mathbf{V}_X$ and $\mathbf{V}^t_Y$ as start super loading weights of $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{w}_{old} = \mathbf{v}_{a,x}$ and $\mathbf{c}_{old} = \mathbf{v}_{a,y}$

Step 2. Calculation of PCs. Repeat until convergence of $\mathbf{w}_a$ and $\mathbf{c}_a$.

- Calculate new $\mathbf{X}$ loading weights: $\mathbf{w}_{new} = \frac{ST_\lambda(\mathbf{M}_{a-1}\mathbf{c}_{old})}{\|ST_\lambda(\mathbf{M}_{a-1}\mathbf{c}_{old})\|}$.
- Calculate new $\mathbf{Y}$ loading weights: $\mathbf{c}_{new} = \frac{\mathbf{M}^t_{a-1}\mathbf{w}_{old}}{\|\mathbf{M}^t_{a-1}\mathbf{w}_{old}\|}$.
- Check for convergence.
- Update both loading weights: $\mathbf{w}_{old} = \mathbf{w}_{new}$ and $\mathbf{c}_{old} = \mathbf{c}_{new}$.

No scores are calculated directly in this algorithm until convergence.

Step 3. Deflation.

- Calculate $\mathbf{X}$ and $\mathbf{Y}$ scores: $\mathbf{t}_a = \frac{\mathbf{X}^t_{a-1}\mathbf{w}_{new}}{\mathbf{w}^t_{new}\mathbf{w}_{new}}$ and $\mathbf{u}_a = \frac{\mathbf{Y}^t_{a-1}\mathbf{c}_{new}}{\mathbf{c}^t_{new}\mathbf{c}_{new}}$.
- Calculate $\mathbf{X}$ and $\mathbf{Y}$ loadings for deflation: $\mathbf{p}_a = \frac{\mathbf{X}^t_{a-1}\mathbf{t}_a}{\mathbf{t}^t_a\mathbf{t}_a}$ and $\mathbf{q}_a = \frac{\mathbf{Y}^t_{a-1}\mathbf{t}_a}{\mathbf{t}^t_a\mathbf{t}_a}$.
- Deflate $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}^t_a$ and $\mathbf{Y}_a = \mathbf{Y}_{a-1} - \mathbf{t}_a\mathbf{q}^t_a$.

4.3.4 Sparse Multiblock PLSR (SMBPLSR)

Multiblock PLSR implies that there is more than one descriptor matrix containing the same amount of samples: $\mathbf{X} = [\mathbf{X}^1 \ldots \mathbf{X}^b]$. These are regressed upon by assembling a common score vector $\mathbf{t}$ to deflate $\mathbf{X}$, gathering information from all blocks [29].

The algorithm largely remains the same if no sparsity is imposed; the difference is explained in the relevant ball point below.

The following NIPALS-based algorithm was proposed by Karaman et al. [29], who based their algorithm on the original MBPLSR algorithm developed by Wangen and Kowalski [72].

Step 1: Initialisation. Unlike the algorithms for NIPALS and PLSR, this algorithm employs subscripts for PC number throughout. This is to show that for some of the steps, using information from earlier PC calculations is a necessity. Furthermore, with the exception of multiple descriptor blocks, the initialisation steps of SMBPLSR are identical to those of SPLSR.

- Select preprocessed and pretreated sets of descriptor and response data $\mathbf{X}_0 = \mathbf{X} = [\mathbf{X}^1, ..., \mathbf{X}^b, ..., \mathbf{X}^B]$ and $\mathbf{Y}_0 = \mathbf{Y}$
- Define $\mathbf{M}_{a-1} = \mathbf{X}^t_{a-1}\mathbf{Y}_{a-1}$.
- Decompose $\mathbf{M}_{a-1}$ by SVD to obtain $\mathbf{V}_X\Delta\mathbf{V}^t_Y$.
- Assign the first singular vectors of $\mathbf{V}_X$ and $\mathbf{V}^t_Y$ as start super loading weights of $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{w}_a = \mathbf{v}_{a,x}$ and $\mathbf{c}_a = \mathbf{v}_{a,y}$

Step 2. Calculation of PCs. Estimation of global parameters

- Define super scores of $\mathbf{Y}$: $\mathbf{u}_a = \mathbf{Y}_{a-1}\mathbf{c}_a$.
  - SMBPLSR: Select loading weights based on soft thresholding: $\mathbf{w}_a = \frac{ST_\lambda(\mathbf{X}^t_{a-1}\mathbf{u}_a)}{\|ST_\lambda(\mathbf{X}^t_{a-1}\mathbf{u}_a)\|}$
  - MBPLSR: Calculate loading weights: $\mathbf{w}_a = \frac{\mathbf{X}^t_{a-1}\mathbf{u}_a}{\|\mathbf{X}^t_{a-1}\mathbf{u}_a}$

Estimation of block parameters. First, $\mathbf{w}_a$ is split into non-normalised block loading weights: $\mathbf{w}_a(b)$.

- For each block $1 \ldots b \ldots B$, block loading weights are normalised: $\mathbf{w}^b_a = \frac{\mathbf{w}^b_a}{\|\mathbf{w}^b_a\|}$.
- Calculate block scores: $\mathbf{t}^b_a = \mathbf{X}^b_a\mathbf{w}^b_a$ for each block $\mathbf{X}^b$.
- Append block scores: $\mathbf{T}_a = [\mathbf{t}^1_a \ldots \mathbf{t}^b_a \ldots \mathbf{t}^B_a]$.
- Calculate $\mathbf{X}$ super weights: $\mathbf{ws}_a = [\|\mathbf{w}_a(1)\| \ldots \|\mathbf{w}_a(b)\| \ldots \|\mathbf{w}_a(B)\|]^t$.
- Calculate $\mathbf{X}$ super scores: $\mathbf{t}_a = \mathbf{T}_a\mathbf{ws}_a$.

Back to estimation of global parameters:

- Calculate new super loading weights of $\mathbf{Y}$: $\mathbf{c}_a = \frac{\mathbf{Y}^t_{a-1}\mathbf{t}_a}{\|\mathbf{T}^t_a\mathbf{u}_a\|}$.

Step 2 is repeated until convergence of $\mathbf{w}_a$.

Step 3. Calculating loadings for deflation and deflating. This step is a fusion of steps 3 and 4 in [29].

- Calculate $\mathbf{X}$ and $\mathbf{Y}$ loadings: $\mathbf{p}_a = \frac{\mathbf{X}^t_{a-1}\mathbf{t}_a}{\mathbf{t}^t_a\mathbf{t}_a}$ and $\mathbf{q}_a = \frac{\mathbf{Y}^t_{a-1}\mathbf{t}_a}{\mathbf{t}^t_a\mathbf{t}_a}$.

- Deflate $\mathbf{X}$ and $\mathbf{Y}$: $\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}'_a$ and $\mathbf{Y}_a = \mathbf{Y}_{a-1} - \mathbf{t}_a\mathbf{q}'_a$.

Once an iteration of the algorithm is completed, $\mathbf{X}_a$ and $\mathbf{Y}_a$ are updated and stages 2. and 3. repeated.

According to Karaman et al. [29] and Hassani et al. [73], deflation can be done upon either block or super scores of $\mathbf{X}$. Both refer to Westerhuis and Smilde [27] who showed that deflating upon super scores leads to mixing of information between blocks and recommended deflating only on $\mathbf{Y}$. Yet, because this method causes super loading weights to become non-sparse, it is not a viable deflation method under the constraint of sparsity. Therefore, in this study we chose to deflate using super scores on both $\mathbf{X}$ and $\mathbf{Y}$.

Chapter 5

Model validation, script and data set

Once statistical analysis has been undertaken, we must employ methods both to validate the stability and performance of the model, and to make the results as easily interpretable as possible. Easy interpretation can be achieved by presenting the results in a visually meaningful way.

In Sections 5.1 and 5.2, we present methods for validating the established statistical model. Next, in Section 5.3, we explain how the predictive ability of the analysis is measured, based on the previously presented validation methods. We then explain why score and correlation loading plots were chosen for visualisation in Section 5.4. Finally, in Section 5.5, we briefly describe the code behind the statistical methods and functions used for analysis.

5.1 Cross Validation (CV)

CV is a model validation technique used to check how well some data set at hand will be able to predict some other independent data set [22, 24]. In this context, independence means using keeping all variables, such as spectrometer and other instrumentation, and growth conditions, constant, while varying the type of microbe grown. Variation can be in strain, species or genus.

For some data set $\mathbf{X}$, a subset of samples, the test set $\hat{\mathbf{X}}$, is left out. Then, a model is established based on the remaining data set $\tilde{\mathbf{X}}$. Following the calibration, $\tilde{\mathbf{X}}$ is used to predict $\hat{\mathbf{X}}$ and the degree of concordance between the two determines the success of the prediction. When this stage is done, another test set is extracted and the process is repeated until all samples have been used.

Test sets can be extracted in numerous ways, and many strategies have been given their own names [74]. In this study, we chose to use two of the most common strategies, namely 'Venetian Blinds' and 'Full CV'. For Venetian Blinds, for k subsets ('k-fold'), every

k^{th} sample is extracted. This configuration is mainly employed on the SMBPLSR runs described. In the data sets analysed in this study, samples are listed so that sample 1, 2 and 3 belong to the same strain, 4, 5, 6 to the next, etc. Therefore, we chose to run a three-fold Venetian Blinds configuration so that no strain or group of samples would be completely removed from any iteration of the CV procedure. For certain analyses of the *Klebsiella* data set, we instead used eight-fold Venetian Blinds CV to better ascertain the results.

For Full CV, also known as full leave-one-out CV, one sample is extracted from the data set before analysis. In practice, leave-one-out CV on an n-sized samples set equals an n-fold Venetian Blinds. More on how full leave-one-out CV was in the next section.

5.2 Cross Model Validation (CMV) and frequency plots

Cross Model Validation could also be called a meta or two-layer CV [75]. While CV only runs a single loop to optimise the model, CMV runs two, one internal and one external [22, 42]. The internal loop corresponds to CV, while the outer loop is left out as an 'independent' set. Westad, Afseth and Bro [75] provide a simple yet apt illustration of this in Figure 5.1.

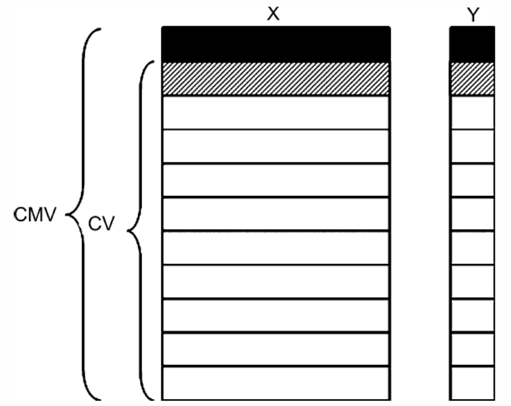


Figure 5.1: Visualisation of the CMV and CV loops. The black row represents the external set for CMV, while the grey row represents the internal CV set. From [75], page 325.

It is worth noting that sparse PLSR methods involve some degree of optimisation, and as such may suffer from overfitting [42]. CMV is employed as a measure to prevent this.

In this study, we used full leave-one-out CV as CMV method on both data sets when running preliminary SPLSR runs on the data. Specifically, for each data type (FTIR/Raman/MALDI), we analysed the entire spectrum and, for each CMV step, recorded the sparse regression coefficients that were selected. The regression coefficients had their values set to one after each step. As the CMV progressed, often-selected coefficients (corresponding to persistent strong signal in variables) accumulated faster than seldom selected ones. After all the CMV steps had been taken, the total accumulated scores for each variable was plotted in a bar plot, displaying how often a variable had been chosen. Blocks to be used in the subsequent SMBPLSR models were selected based on the frequency of selected variables in the frequency plot [42]. A sample frequency plot based on the FTIR data from the *B. subtilis/E. coli* data set is shown in Figure 5.2. Note that this frequency plot was made using Venetian Blinds only for illustrative purposes and is not used later in the thesis.

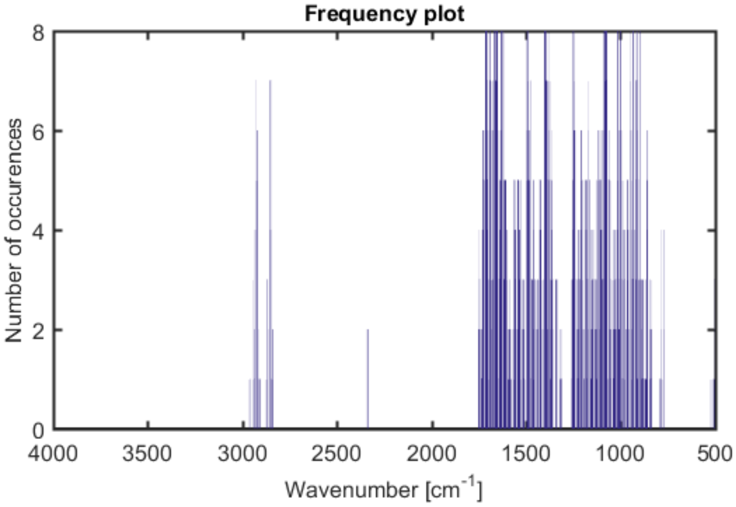


Figure 5.2: Dummy frequency plot based on FTIR data from the *B. subtilis/E. coli* data set.

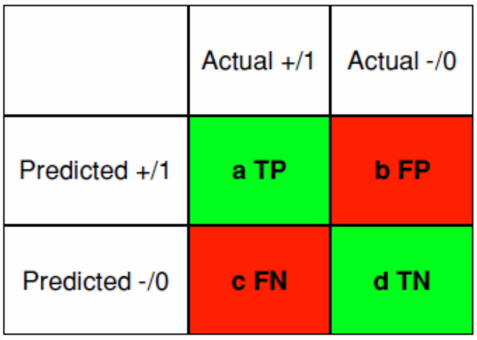
5.3 Misclassification and success rates

In spectroscopy, an often-used method of assessing the predictive ability of a model is to calculate the number of misclassifications (NMC). The way the model works is simply to relate the elements of the predicted $\hat{\mathbf{Y}}$ to some threshold value or classification boundary. Each column $\mathbf{y}_n$ in the indicator matrix $\mathbf{Y}$ produced from the response data contains 0's or 1's based on actual classification: $\mathbf{y}_1$ contains 1's for all samples of group 1 and 0's for all other samples, $\mathbf{y}_2$ 1's for group 2 and 0's for all other groups, and so on. The predicted values $\hat{\mathbf{y}}_n$ are based on the model

$$\hat{\mathbf{Y}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\beta}_0 \qquad (5.1)$$

where $\bar{\mathbf{X}}$ is the pretreated descriptor matrix, $\boldsymbol{\beta}$ are the regression coefficients and $\boldsymbol{\beta}$ the intercept. If, for an element i, j, $|Y_{i,j} - \hat{Y}_{i,j}| < \delta$, δ being a threshold value, the predicted element $\hat{Y}_{i,j}$ in Equation (5.1) is classified according to the corresponding element $y_{n,i}$ and assigned a value of either 0 or 1. All elements of $\hat{\mathbf{Y}}$ are predicted this way. In this study, we used $\delta = 0.5$.

For a two-class system, one often labels the classes in terms such as ‘case’ and ‘control’ [76] and assign values 1 - or positive - to case, and 0 - or negative - to control. For each prediction, the turnout must be either True or False Positive (TP/FP) or True or False Negative (TN/FN). This system, called the confusion matrix or contingency table [76], is displayed in Figure 5.3.



	Actual +/1	Actual -/0
Predicted +/1	**a TP**	**b FP**
Predicted -/0	**c FN**	**d TN**

Figure 5.3: Classification of a two-class system. Predicted values on the y-axis are compared with actual reference values on the x-axis. From [76], page 172.

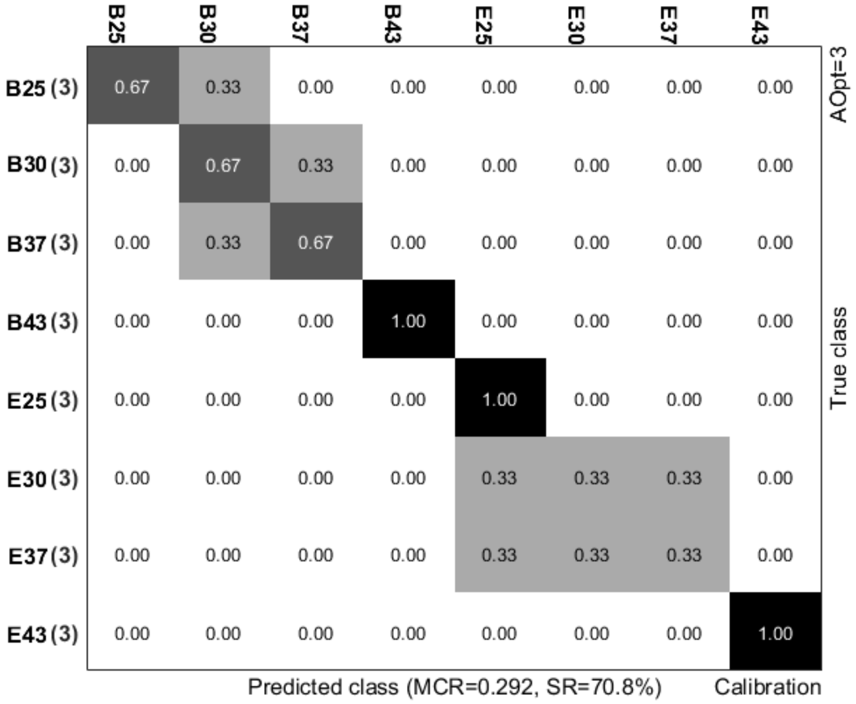
After classification, the number of misclassifications (NMC) can easily be calculated by summing the number of FP’s and FN’s: $NMC = FP + FN$. But this number is in itself of limited usefulness [76]. By dividing it on the number of samples in the data set (N), one can obtain the misclassification rate (MCR) [76, 29]:

$$MCR = \frac{NMC}{N} = \frac{FP + FN}{N} \tag{5.2}$$

Where NMC is the number of misclassifications, N is the number of samples, FP are the false positives and FN are the false negatives.

The success rate (SR) is simply the percentage of correctly predicted samples: $SR = (1 - MCR) \times 100\%$. In the following, this number will be the most often reported when assessing any model’s predictive ability.

The confusion matrix of Broadhurst and Kell [76] is extendable to higher-class systems, in which case (5.2) will show as what class some sample was falsely identified. In addition to this, grey-scaling can further enhance intuitive understanding. Hue saturation is proportional to fraction size. Thus, for a completely correctly identified set, all diagonal entries are black and all the others white. However, this is rarely the case, as the example in Figure 5.4 shows.



	B25	B30	B37	B43	E25	E30	E37	E43
B25 (3)	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00
B30 (3)	0.00	0.67	0.33	0.00	0.00	0.00	0.00	0.00
B37 (3)	0.00	0.33	0.67	0.00	0.00	0.00	0.00	0.00
B43 (3)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
E25 (3)	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
E30 (3)	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.00
E37 (3)	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.00
E43 (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

AOpt=3
True class
Predicted class (MCR=0.292, SR=70.8%)
Calibration

Figure 5.4: Sample figure of a confusion matrix for eight classes, showing the result of three-fold CV of FTIR data from the *B. subtilis*/*E. coli* data set. *AOpt* is the optimal number of principal components for the model.

5.4 Score and correlation loading plots

Score plots are a common modern technique of visualising sample variation within multivariate data [19, 29, 77]. Plots can be shown both for block and global scores. The purpose of the score plot is to give a two-dimensional visualisation of the variance in the data set [19, 29]. This is done by plotting a set of scores from two different latent variables of the same data set against each other, showing how each sample in the data set deviates from the mean (origin). Assuming collinearity between samples, we can expect similar samples, for instance the biological replicates between strains, to show similar patterns of deviation for each latent variable in the global scores. The different groups of samples in the plot can be named or colour-coded to enhance interpretability even further [19, 29, 77]. In the block scores, we can zoom in on differences and similarities in chemical composition between samples. This is especially useful when examining for instance phenotypic response to varying growth conditions, or what type of biomolecule explains the most of the variation between strains [2]. Both examples just mentioned are essential to the scope of this thesis.

A sample score plot is shown in Figure 5.5, where global scores of latent variables 1 and 4 are plotted against each other for an SMBPLSR run on four blocks of FTIR data.

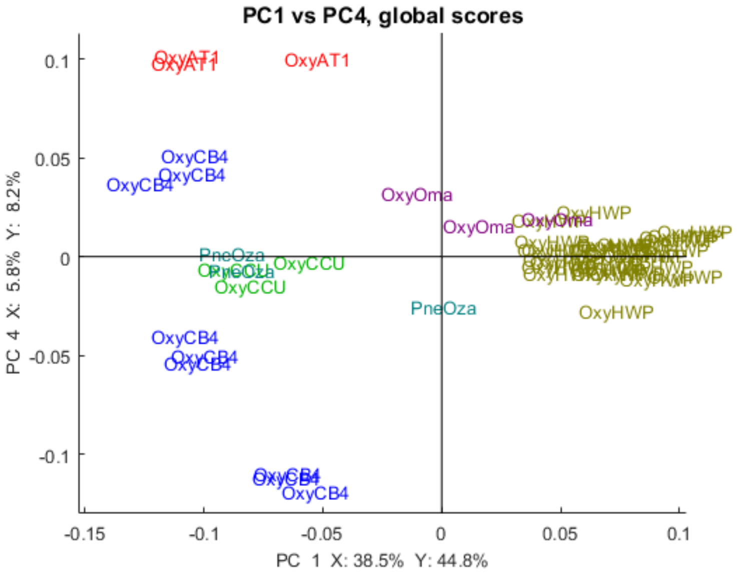


Figure 5.5: Sample figure for a global score plot of PC1 vs PC4 from a four block SMBPLSR model on FTIR data.

Correlation loading plots are used to visualise the correlation between variables for two selected latent variables [19, 29]. A sample plot is shown in Figure 5.6, showing the correlation between variables in the FTIR lipid region for a: All block variables and b: Sparse block variables. Each number in the plot represents the wavenumber of the examined variable.

The outermost circle in the plot denotes 100% correlation between the two components of the variable; the innermost, dotted circle, a correlation of 50%. As seen in Figure 5.6a, these plots have a tendency of becoming cluttered. Normally, we are most interested in the variables explaining the most information, in this case the variables remaining after sparsity has been imposed, shown in Figure 5.6b. For both subplots, there are tendencies of strong correlation or anticorrelation between variables in the lipid region. For further examination not shown in these sample plots, it is possible to add data points corresponding to certain experimental conditions, such as susceptibility to chemicals [19] or diet [29].

5.5 Test scripts

Two main MATLAB scripts were set up for this study, one for each data set. These two scripts share a common main setup originally done by Valeria Tafintseva, reseracher at NMBU, but was modified and extended by the author for the specific purposes of

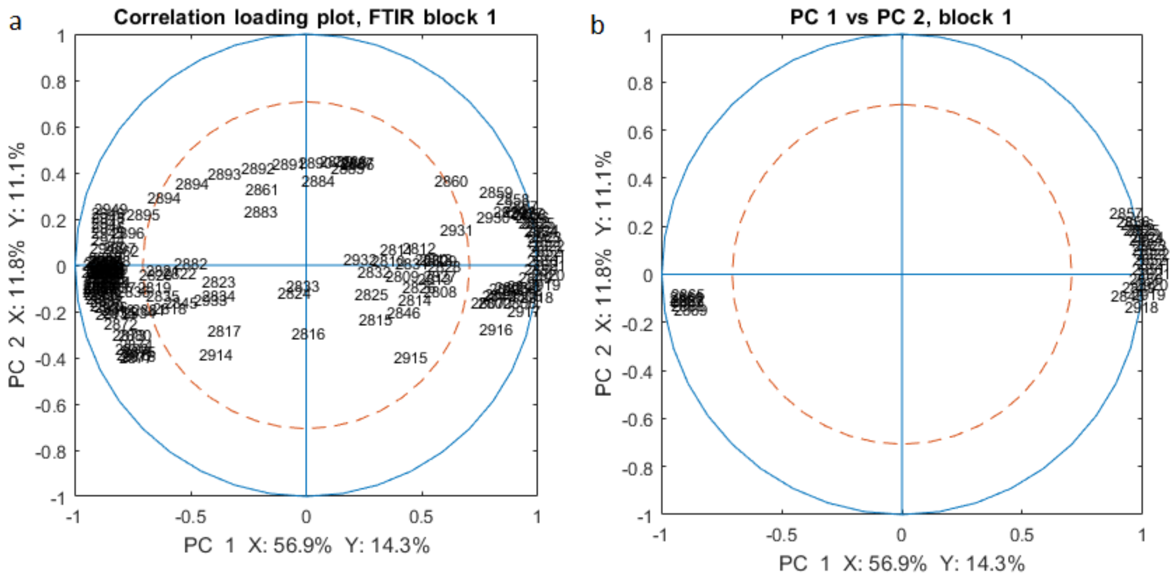


Figure 5.6: Sample figure for a correlation loading plot of block 1 (lipid variables) from a four block SMBPLSR model on FTIR data. a: All block variables, b: Sparse variables.

this study. The script `id_script.m`, presented in Appendix C.1, is used to analyse the identification data set, whose results are presented in Chapter 6. The script `exp_script.m`, presented in Appendix C.2, is used to analyse the experimental design data set, whose results are presented in Chapter 7.

5.5.1 MATLAB functions and other software

The MATLAB software provides packages for both algorithms and allows for the creation of custom packages. The program The Unscrambler X version 10.4 (CAMO, Oslo, Norway) software has been used to prepare spectra for the aforementioned scripts.

Most of the functions used in this study are part of the Saisir (fr: ‘to grasp’, metaphorically: ‘to understand’; Statistics Applied to the Interpretation of Spectra in the InfraRed) package [78]. Saisir is a package of around 200 MATLAB functions, developed by Dr Dominique Bertrand (Institut National de la Recherche Agronomique (INRA), Nantes, France) and Dr Christophe Cordella (INRA and AgroParisTech, Paris, France). The functions are specifically designed for chemometricians. Additional functions based on the Saisir data structure have been coded or assembled in-house (NMBU), mostly by Professor Achim Kohler or researcher Valeria Tafintseva. Some functions were made by the author of this thesis, including the main data scripts used for data processing in this thesis.

Chapter 6

Results and discussion, identification

This chapter deals with the first part of the scope for this thesis, namely how to use sparse multiblock methods to establish a connection between metabolic fingerprint by FTIR and Raman, and protein expression by MALDI. Unless otherwise stated, all data sets and spectra in this chapter refer to the *Klebsiella* data set described in Section 2.1. Data sets are referred to as multiblock sets or simply sets, depending on whether each data type is represented by one block, or their respective number of selected blocks. For example, FTIR multiblock set 1 consists of four blocks, while FTIR set 1 means that these four blocks have been merged into one.

In Sections 6.1 and 6.2, details of the experimental setup are described; how blocks were selected and demarcated to increase interpretability and computational speed. In the following four sections, we present and discuss the results of four different SMBPLSR runs. In Section 6.8, results of the runs are summarised and the most important findings highlighted, before being discussed in Section 6.9.

6.1 Experimental setup

6.1.1 Preprocessed spectra

In this study, we preprocessed the FTIR data set by applying a 2^{nd} derivative, 2^{nd} degree polynomial SG filter, using a window size of 9. The data set was further preprocessed by EMSC with a quadratic extension, also known as standard EMSC [59]. The preprocessed FTIR spectra are shown in Figure 6.1a.

The Raman data set used in this study was preprocessed by applying a 2^{nd} degree polynomial SG filter, using a window size of 9, but no derivation. This is in accordance with the preprocessing techniques of Dieckmann et al. [2] for the same data set. The data set was further preprocessed by the standard EMSC algorithm [59]. The resultant preprocessed spectra are shown in Figure 6.1b.

The MALDI data set was already preprocessed upon receipt, first by using an SG filter with a window size of 21, then by baseline correction and finally by vector-normalisation. In this case, however, the preprocessed spectra only made up an intermediate step and were not used for further analysis. This role was instead taken by a bar code plot [79]. The bar code plot was set up by selecting the 30 highest-intensity peaks in the preprocessed spectra. Their y-values were then set to 1, thus omitting their relative intensity in further pretreatment and analysis. All other values in the spectrum were set to 0. This preprocessing technique is in accordance with the one used by Dieckmann et al. [2] and Lasch et al. [79] for MALDI preprocessing. The MALDI bar code spectra are shown in Figure 6.1c.

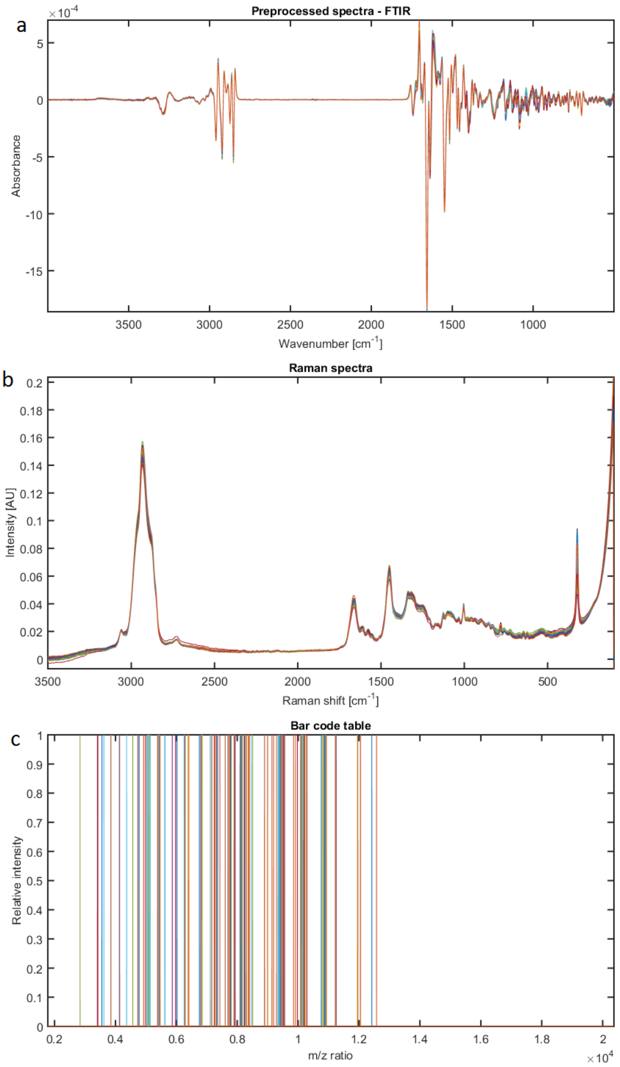


Figure 6.1: Preprocessed spectra (FTIR and Raman), and bar code plot (MALDI) from the *Klebsiella* data set. a: FTIR, b: Raman, c: MALDI.

Figure 6.1c shows that the MALDI spectra still contained a large, redundant tail region (from $\left(\frac{m}{z}\right) \approx 1.3 \times 10^4$). This tail was cut off to increase computational speed.

6.1.2 Significance of appropriate SG filter parameters for optimising Raman spectral analysability

During the data preparation stages, we attempted several different ways of preprocessing the Raman spectra. Dieckmann et al [2] preprocessed the spectra using an SG smoothing filter with a window size of 9 and no derivative, followed by vector-normalisation. At first, we preprocessed the spectra using an SG filter with a 2^{nd} degree polynomial, a window size of 9 and 2^{nd} derivative taken, followed by standard EMSC [19, 65]. The resultant frequency plot showed that variables had been selected across the entire spectrum, including areas known to be of no value to analysis [13]. This frequency plot is shown in Figure 6.2.

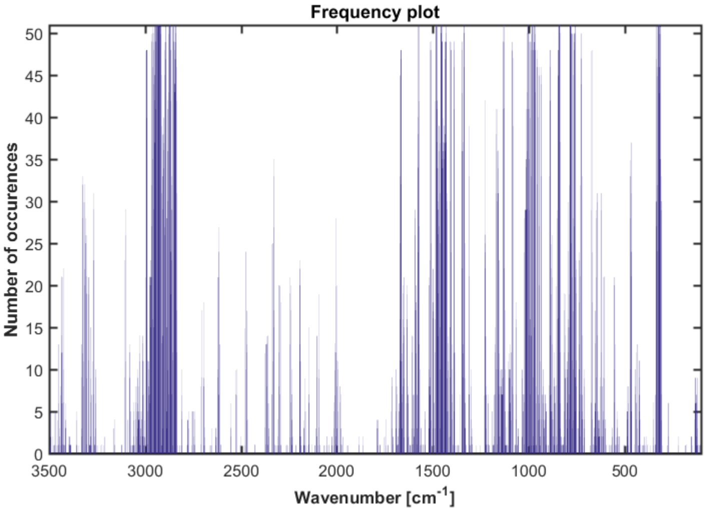


Figure 6.2: Discarded frequency plot obtained from a full leave-one-out CMV (51-fold) of Raman data, using an SPLSR model and discarding 90% of variables. The Raman data were preprocessed by an SG filter that included taking the 2^{nd} derivative, resulting in a high signal-to-noise ratio and subsequently poor selection of sparse regression coefficients.

We attributed this undesired selection to a high signal-to-noise ratio in the samples and attempted to copy the preprocessing procedure used by Dieckmann et al. [2] as described above in order to attenuate the signal-to-noise ratio, but to no avail. Variables were still selected throughout the spectrum (not shown).

A combination of the two preprocessing algorithms (our initial attempt and the one used by Dieckmann et al.) was then attempted: An SG filter with a 2^{nd} degree polynomial and a window size of 9, but no derivative, followed by standard EMSC. Both model prediction and selection of sparse regression coefficients improved as a result: For an eight-fold Venetian Blinds CV, 86% of the samples were correctly identified, compared to 78% for our initial attempt, all other conditions remaining the same. This is in accordance with findings of Karaman et al. [42], who reported that the predictive ability of Sparse PLSR methods was poor for high signal-to-noise ratios.

The result of the non-derivative selection of regression coefficients is shown in Figure 6.3b. Our results show that the choice of the preprocessing method was crucial. The process of taking the 2^{nd} derivative of the Raman spectra did, unlike the FTIR spectra, increase the signal-to-noise ratio to such a degree that it greatly altered the selection of regression coefficients and subsequent data analysis, as is easily seen comparing Figures 6.2 and 6.3b.

6.1.3 Selecting relevant spectral regions

One of the major advantages of multiblock methods is their ability to select only the specific regions containing most of the desired spectral information [27]. To locate the most important regions of each spectrum, full leave-one-out CMV was run on all data types in single descriptor runs in order to create frequency plots as described in Section 5.2. In this thesis, we let the frequency plots form the basis for the blocks we selected in the subsequent sparse multiblock analyses.

Figure 6.3a shows the frequency plot for the SPLSR model using the entire FTIR spectrum as the descriptor data. The sparsity parameter was set to discard 90% of the variables, and 10 PCs were calculated. The vast majority of the most frequently selected regression coefficients were collected from the regions around 1750-1550, around 1400 and 1250-900 cm^{-1}, the former region corresponding to the amide bonds of the protein region and the latter two to the various bonds associated with carbohydrates. Furthermore, a minority of the frequently selected coefficients stemmed from the lipid region between 2950-2800 cm^{-1}. This information prompted the creation of FTIR multiblock set 1, consisting of the following blocks:

- Block 1 from 2950-2800 cm^{-1}.
- Block 2 from 1750-1550 cm^{-1}.
- Block 3 from 1450-1350 cm^{-1}.
- Block 4 from 1250-900 cm^{-1}.

Figure 6.3b shows the frequency plot for the SPLSR model using the entire Raman spectrum as descriptor data. Blocks were selected following the same principles as those described for FTIR. Raman multiblock set 1 was constructed; it consisted of the following blocks:

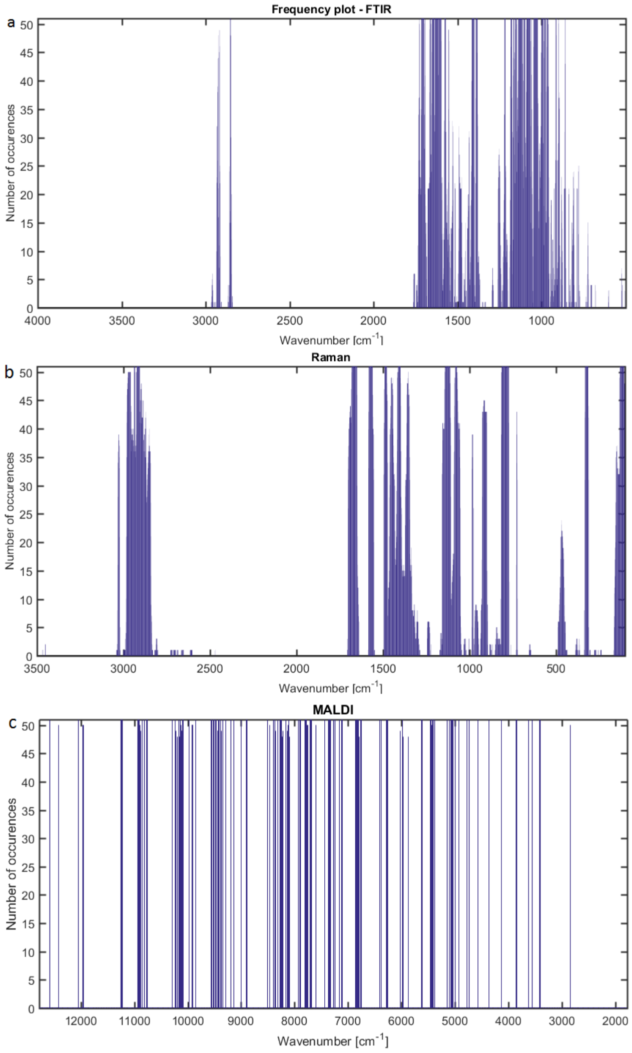


Figure 6.3: Frequency plot from full leave-one-out CMV (51-fold) on all data types. Each data type was analysed by the SPLSR model described in Section 4.3.3. For all analyses, 90% of variables were discarded by soft thresholding. a: FTIR, b: Raman, c: MALDI.

- Block 1 from 3050-2800 cm^{-1}.
- Block 2 from 1700-1550 cm^{-1}.
- Block 3 from 1500-1300 cm^{-1}.
- Block 4 from 1200-750 cm^{-1}.
- Block 5 from 350-100 cm^{-1}.

Figure 6.3c shows the frequency plot for the SPLSR model using the entire MALDI spectrum, minus the redundant tail region mentioned above, as descriptor data. No clear variable clustering is evident. Notwithstanding, some incisions were made to redundant regions of the data set. The spectrum was reduced to include only the $\frac{m}{z}$ region between 3000 and 11000 in accordance with existing literature [2, 38, 79]. This data set is from now on referred to as MALDI set 1.

6.2 Predictive abilities and weighting of regression coefficients

6.2.1 Correspondence between strain and script names

Table A.1 in Appendix A shows the correspondence between the sample strain names shown in Section 2.1.2 and their given names in the data script. The six-letter names in the middle column were used further on in the text and figures, for brevity. The first three letters refer to species and the final three to group.

6.2.2 Description of input data

After blocks had been selected and redundant regions cut away, we proceeded by running sparse multiblock routines on all data types. The input data setups we analysed, were

- Analysis 1: FTIR multiblock set 1 as descriptor data, four blocks as described above.
- Analysis 2: Raman multiblock set 1 as descriptor data, five blocks as described above.
- Analysis 3: FTIR set 1[1], Raman set 1[2] and MALDI set 1, in total three blocks. FTIR and Raman data blocks as described above, were concatenated.
- Analysis 4: FTIR multiblock set 1, Raman multiblock set 1 and MALDI set 1, ten blocks.

As MALDI set 1 was not grouped after SPLSR, we chose to proceed using only the cut data set, as described above in Section 6.1.3. For all analyses, we set the sparsity parameters equal to those for SPLSR, discarding 90% of the variables and calculating 10 PCs. For the single descriptor-data type analyses, three-fold Venetian Blinds CV was performed in order to expose possible overfitting and to assess predictive ability. For the multiple-type descriptor data runs, we increased to eightfold CV. Table B.1 in B lists the global explained variance in analyses 3 and 4.

[1] FTIR set 1 means one data block consisting of the four data blocks in FTIR multiblock set 1.

[2] Raman set 1 means one data block consisting of the five data blocks in Raman multiblock set 1.

To establish a connection between the predictive abilities of the metabolic fingerprint produced by FTIR/Raman sets 1, and the protein expression in MALDI set 1, we wanted to compare both from what regions and what groups the sparse regression coefficients were selected, and see how this selection was expressed in global and block score plots.

6.3 Analysis 1: FTIR multiblock set 1

Figure 6.4 shows the selected regression coefficients from analysis 1. Regression coefficients are selected from all subregions and seemed to correspond well with known resonance frequencies for common biomolecular bonds [13]. The signal strength for each group varied between subregions, an early indication of differences in chemical composition; strong signals indicate high prevalence of certain chemical bonds. The success rate for prediction in this analysis was 96.1%. Only two samples were wrongly identified as OxyHWP.

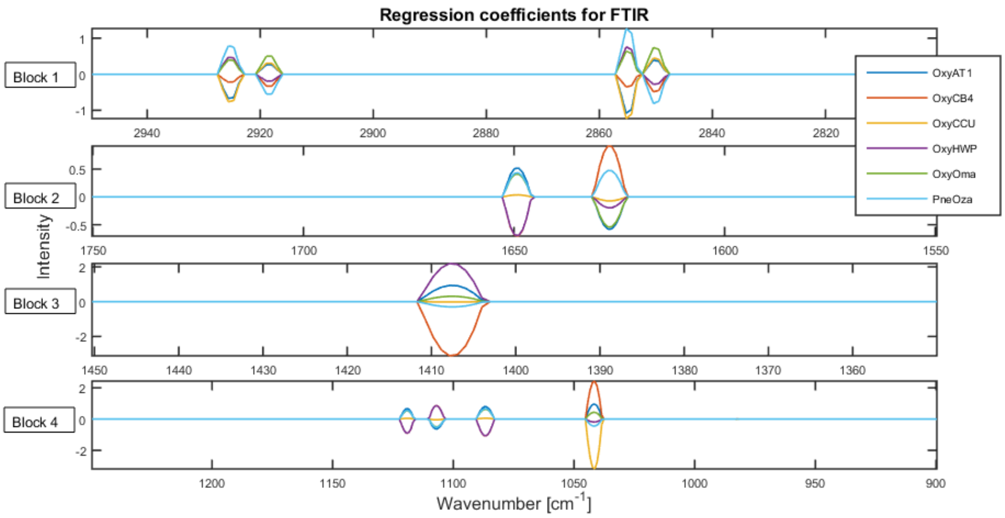


Figure 6.4: Selected sparse regression coefficients for analysis 1, on FTIR multiblock set 1.

6.4 Analysis 2: Raman multiblock set 1

Figure 6.5 shows the selected regression coefficients from the Raman multiblock set 1 analysis. Signal strength in PneOza, for the upper fingerprint region in particular, differed from the regression coefficients in analysis 1. No regression coefficients were selected from block one, covering the lipid region. This was unexpected given the number of variables of that region being selected in Figure 6.1b. Large numbers of variables were selected from the protein and upper carbohydrate regions, especially in the OxyHWP group. Furthermore, the success rate of prediction was 84.3%. Most misclassifications were identified as OxyHWP. Two samples were also misclassified as OxyCB4.

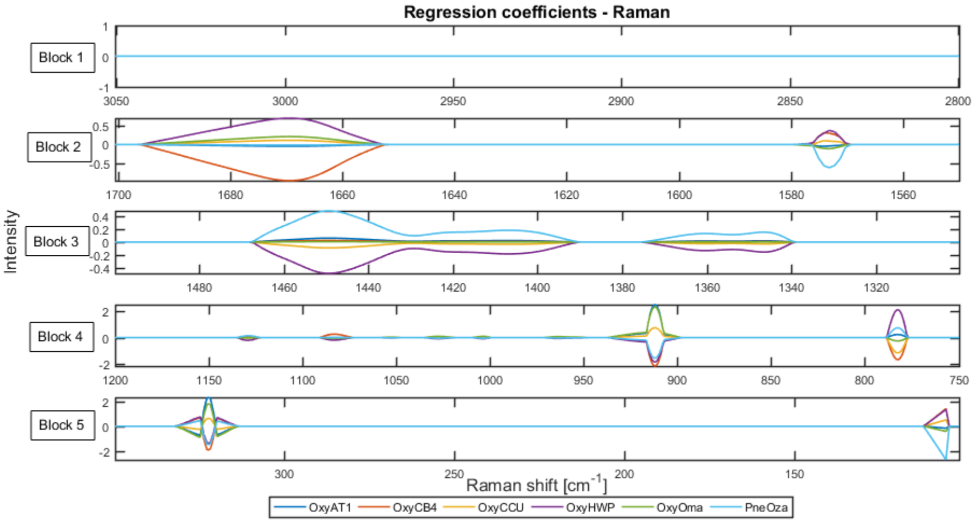


Figure 6.5: Selected sparse regression coefficients for analysis 2, on Raman multiblock set 1.

6.5 Analysis of MALDI set 1

The scattering of regression coefficients in MALDI set 1 (Figure 6.3c) prompted no division into blocks. Instead, the regression coefficients obtained from the SPLSR analysis were used for further comparison and are shown in Figure 6.6. The corresponding SR was 82.4%, mostly misclassifying samples as OxyHWP. Two samples were also misclassified as OxyCB4 (not shown).

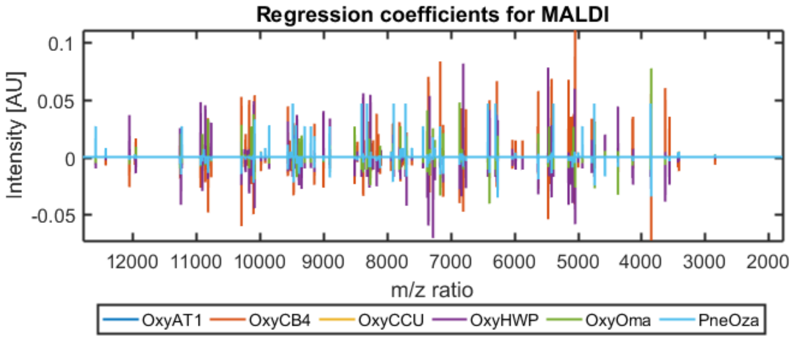


Figure 6.6: Selected sparse regression coefficients for MALDI set 1 analysis with SPLSR. These were obtained before the spectrum was cut as described in subsection 6.1.3.

The tendency of strong signals from OxyCB4 in the protein region was continued; many of the strongest signals shown by recorded regression coefficients in Figure 6.6 are attributed to this group. Almost no coefficients seemed to be selected from OxyCCU, which for analyses 1 and 2 showed the strongest signal in the lower fingerprint region. In accordance with existing literature ([2, 38, 79]), most regression coefficients were selected in the $\frac{m}{z}$ range between 3 and 10 kDa.

6.6 Analysis 3: Concatenated data blocks, all data types

Figure 6.7 shows the selection of regression coefficients from concatenated blocks of each data type analysed by an SMBPLSR routine. Comparing with the regression coefficients of FTIR multiblock set 1 in Figure 6.4, Raman multiblock set 1 in Figure 6.5 and MALDI in Figure 6.6, we see a few important changes: Variables were selected far more sparsely in the lipid and carbohydrate regions in FTIR set 1 than for analysis 1, while variables were more often selected in the protein and lower carbohydrate regions. For Raman set 1, the changes were even more clear: Variable selection was almost inverted - only variables between 350 and 100 cm^{-1} were selected in both analyses. The number of variables selected in MALDI was quite high and scattered across the entire spectrum. Regression coefficients for PneOza and OxyCB4 now seemed to be the strongest, partly in accordance with their strong signal in the protein regions in analyses 1 and 2.

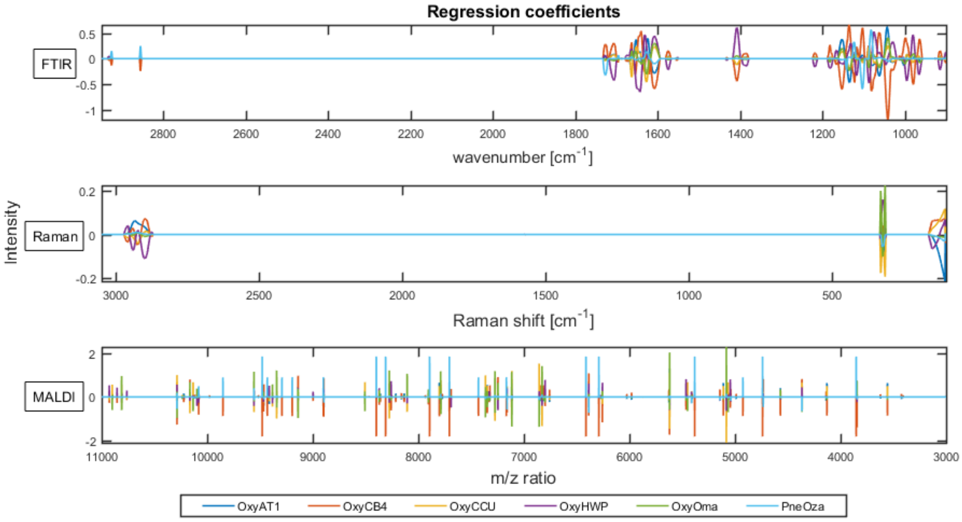


Figure 6.7: Selected sparse regression coefficients for analysis 3: Concatenated blocks, all data types.

The corresponding confusion matrix is shown in Figure 6.8. The success rate of prediction was 84.3%, only slightly better than MALDI set 1 and worse than Raman multiblock set 1. This time, eight-fold Venetian Blinds CV was used to address the model's predictive ability. We also see that this time, fewer samples were misclassified as OxyHWP, possibly owing to its apparent absence from the MALDI regression coefficients.

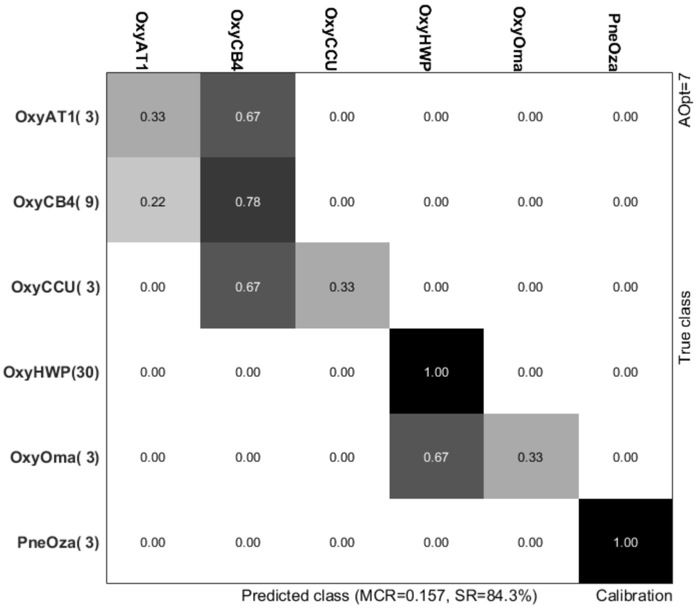


Figure 6.8: Confusion matrix from analysis 3. The CV configuration was eight-fold Venetian Blinds.

The global scores from analysis 3 are shown in Figure 6.9. All combinations of PC1, PC2, PC3 and PC4 were plotted against one another. In any plot involving PC1, OxyHWP was separated from the other groups with only a few borderline overlaps in PC1 vs PC2 and PC1 vs PC4. In any plot involving PC4, we see three distinct tendencies. The first is that PneOza was clearly separated from all other groups. The second is that OxyCCU and OxyOma were clustered. The third is that OxyAT1 and OxyCB4 congregated away from the other groups.

Figure 6.10 shows a selection of recorded block scores plots of the FTIR set 1 from analysis 3. In PC1 vs PC3, the grouping of OxyHWP was very similar to that in the corresponding global score plot in Figure 6.9. The block scores showed a slightly better separation between OxyHWP and any other group, and even managed to split OxyCCU and OxyOma. Comparing PC1 vs PC4 for the two plots, we see that the separation of PneOza was still clear in the block scores, although not as pronounced as in the global scores. In the block scores, we also see that OxyCCU has been properly separated from the rest and is identifiable as a group.

Figure 6.11 shows the most clearly grouped result from the Raman set 1 block scores. There is some separation between OxyHWP and the remaining groups, but it is not nearly as clear as in the FTIR set 1 block scores. Apart from this, the Raman set 1 block scores bear no clear resemblance to the global scores.

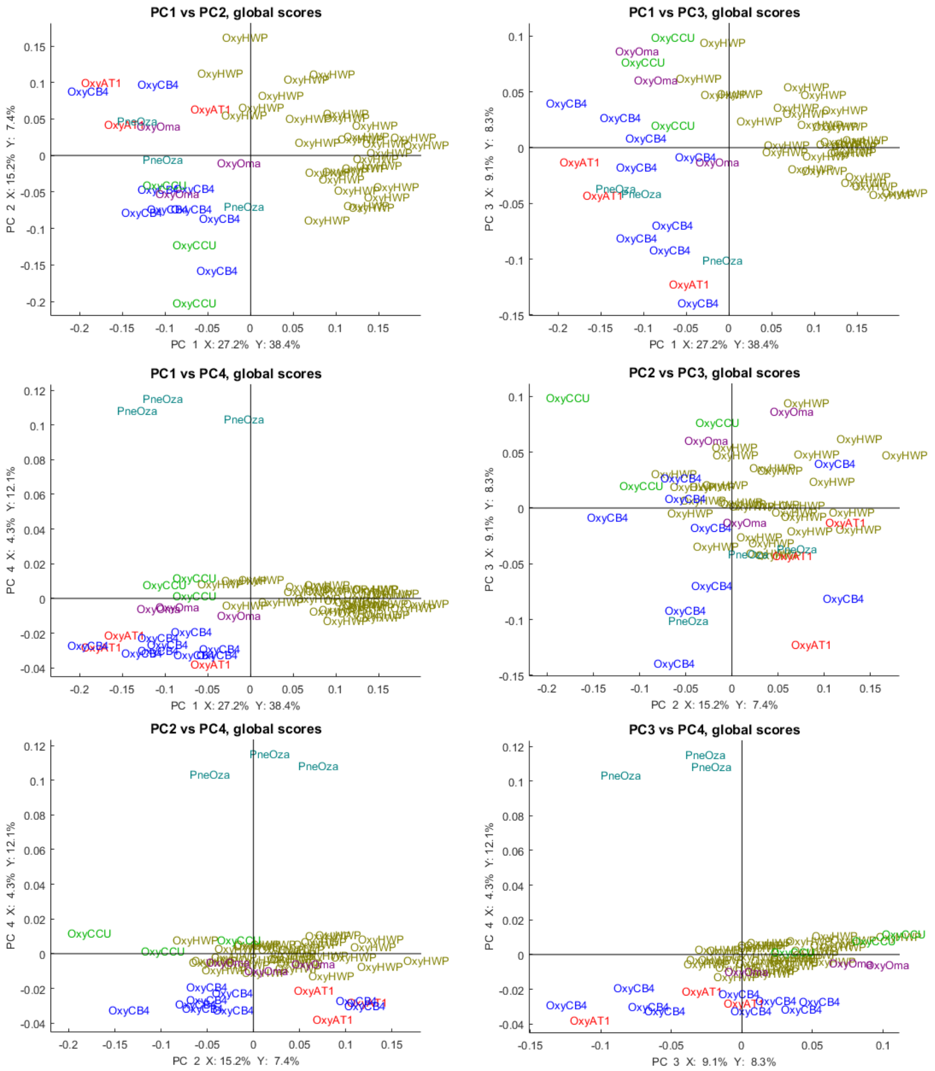


Figure 6.9: Global scores from analysis 3: Concatenated blocks, all data types.

Figure 6.12 shows the selected set of block score plots from MALDI set 1 in analysis 3.

MALDI represented a rather curious case. Unlike FTIR and Raman, the variance in MALDI data was not explained in any primary dimension; there was no dominant principal component (no component even exceeded 10% explained variance) due to the lower degree of collinearity between samples. This resulted in very little of the total variance being explained by MALDI in the first one to PCs. First when we plotted PCs from 3 and up against each other, we saw trends and patterns also evident in the corresponding global score plots. For example, the MALDI block score plot of PC3 vs PC4 and its global peer showed striking overall resemblance. A rather clear separation and grouping of OxyCCU was also present in PC1 vs PC3 in the block scores, but not present in the global scores. In the MALDI block scores, PneOza was also more consistently clustered than seen in the global scores. This effect did, as discussed above, not become fully apparent until most of

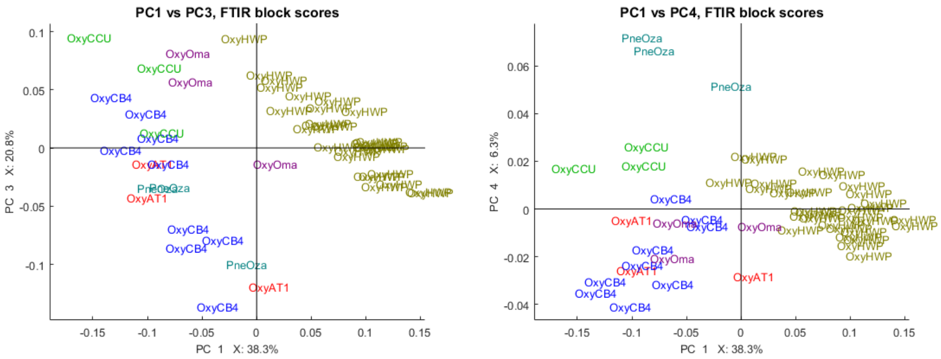


Figure 6.10: FTIR set 1 block scores from analysis 3.

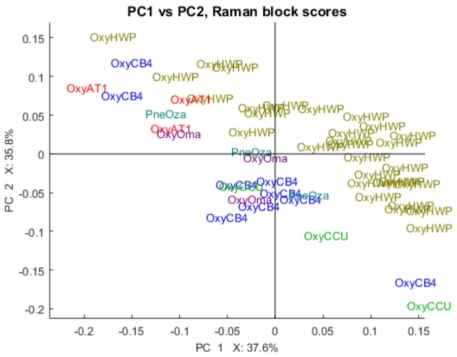


Figure 6.11: Raman set 1 block scores from analysis 3.

the global variance in a PC was explained by MALDI.

6.7 Analysis 4: Split data blocks, all data types

Figure 6.13 shows the regression coefficients selected when the analysed data consisted of FTIR multiblock set 1, Raman multiblock set 1, and MALDI set 1, totalling 10 data blocks. Comparing with 6.4, 6.5 and 6.6 above, we see that the that selection of variables in Raman multiblock set 1 to a much higher degree resembled the one from analysis 2. MALDI set 1 variables were selected far less frequently than for analysis 3, and this time only variables pertaining to groups PneOza and OxyHWP seemed to be selected.

The biological interpretation of this result is that in analysis 4, biological information was gathered from a larger variety of chemical bonds than in analysis 3. That is, in analysis 4, biological information was also gathered from the protein and fingerprint regions of Raman multiblock set 1, potentially providing a more detailed correspondence between metabolic fingerprint in FTIR/Raman and protein expression in MALDI set 1 in the score plots.

The confusion matrix produced from analysis 4 is shown in Figure 6.14. The success

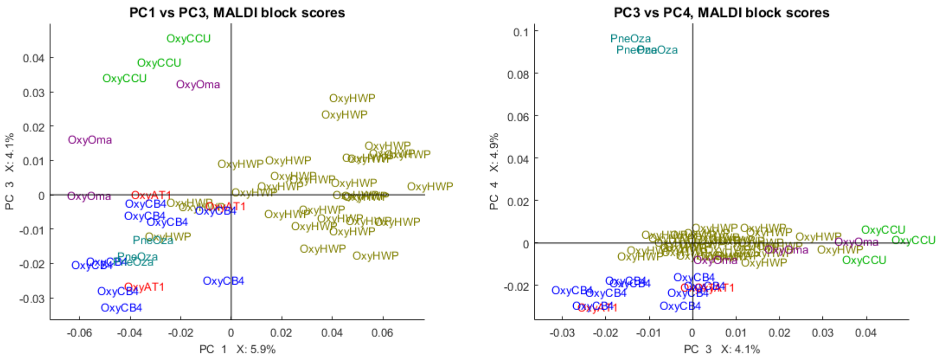


Figure 6.12: MALDI set 1 block scores from analysis 3.

rate was even higher than for the concatenated block model: 86.3%. Although misclassified samples without exception were identified either as OxyHWP or OxyCB4, the two largest groups, most of the groups were either all right or all wrong, a result divergent from the more partial group misclassifications shown in Figure 6.8 and also seen in single data type models (not shown). The abundance of OxyHWP and relative absence of other groups of variables selected from the MALDI set 1 block, might explain this ascription. This model used an eight-fold Venetian Blinds CV configuration for prediction.

Figure 6.15 shows the global scores from analysis 4. The proximity of OxyOma and one sample of PneOza to OxyHWP shown in all subplots involving PC1, and the clustered grouping of OxyAT1 and OxyCB4 seen in PC1 vs PC2, were both expected from the misclassifications shown in Figure 6.14.

6.7.1 Resemblance between global scores and block scores in the lipid region

In this subsection, score plots from blocks 1 from FTIR multiblock set 1, and 1 from Raman multiblock set 1, are compared with the global scores in Figure 6.15. The weak signal seen in Raman block 1 in Figure 6.13 was also present in the block scores: Variables were only selected from PC4 and PC6. In FTIR block 1, no variables were selected in the block loadings of PCs 4 and 5.

Figure 6.16 shows a selection of block score plots from the lipid blocks. Some of the trends present in the global scores were also seen here: OxyAT1, OxyCB4 and OxyHWP showed a similar interrelated pattern for subplots of PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3. The PneOza outlier was present in both block and global scores. OxyOma was separated better in the global scores than in the present block scores.

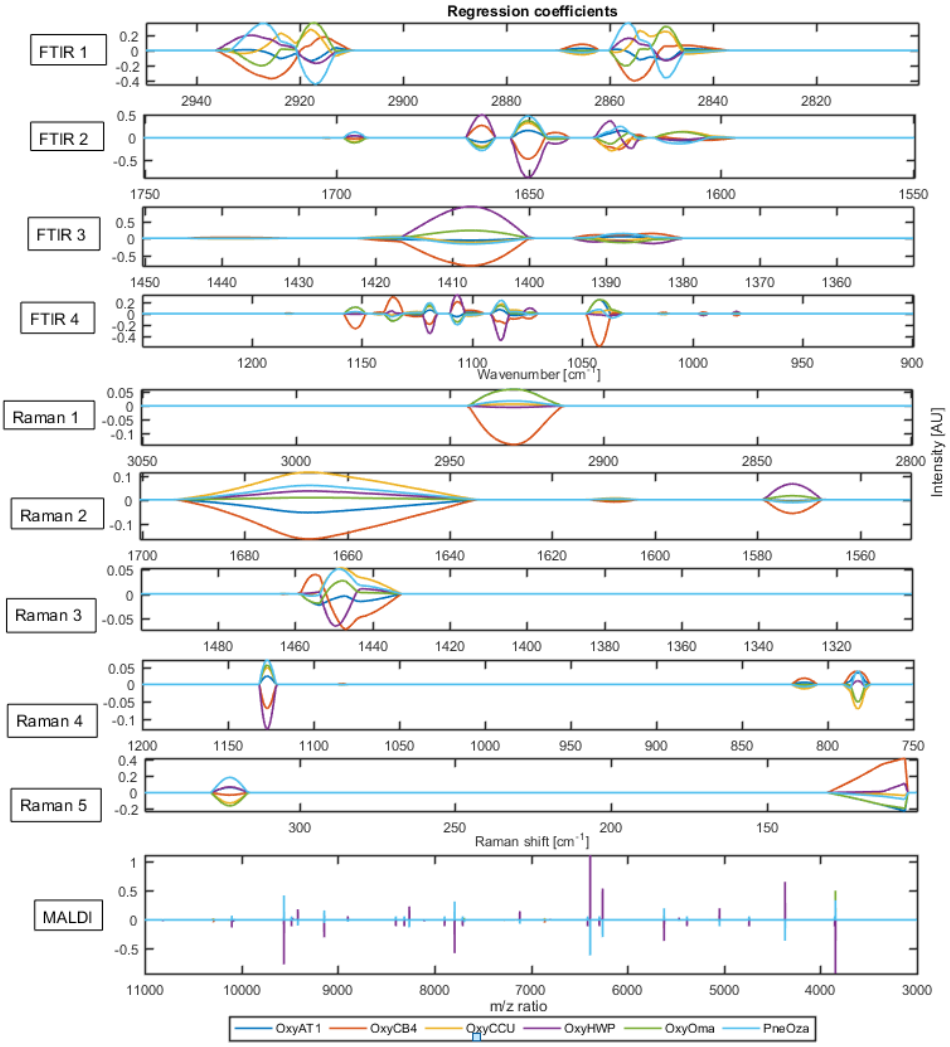


Figure 6.13: Selected sparse regression coefficients for analysis 4: Split blocks, all data types.

6.7.2 Resemblance between global scores and block scores in the amide region

In this subsection, score plots from blocks 2 from FTIR multiblock set 1, and 2 from Raman multiblock 1 one are compared with the global scores in Figure 6.15. Variables were only selected for the block scores of PCs 1, 2 and 7 for the Raman block, and from all PCs but 4 and 6 for the FTIR block.

Figure 6.17 shows a selection of block scores from the amide blocks. No detailed resemblance seemed to exist between the amide region block scores and the corresponding global scores. We did, however, see a clearer separation of PneOza in plots involving PC3 than in the global scores. We also saw a much denser grouping of OxyOma in PC1 vs PC2 of Raman block 2 in the block scores.

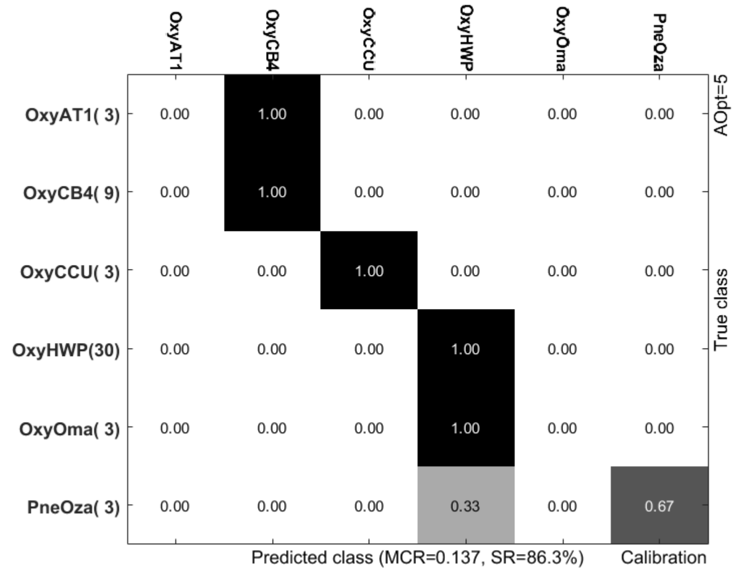


Figure 6.14: Confusion matrix showing the predictive ability of analysis 4. The CV configuration was eight-fold Venetian Blinds.

6.7.3 Resemblance between global scores and block scores in the fingerprint/carbohydrate and far IR region

In this subsection, score plots from blocks 3 and 4 from FTIR multiblock set 1 and 3, and 3 and 4 from Raman multiblock set 1, are compared with the global scores in Figure 6.15. In FTIR block 3, no variables were selected in the block loadings for PCs 4-6; variables were selected for all PCs in block 4. In Raman block 3, variables were selected for PCs 2 and 4-6; in block 4, from PCs 2, 3 and 5.

Figure 6.18 shows a selection of block score plots from the Raman multiblock set 1 fingerprint/carbohydrate region. In block 3 (left), we see that the model managed to separate OxyOma and OxyAT1 from the remaining strains, although clustering them together. OxyAT1 was also in block 4 (right), but not separate from the main cluster of strains. No other clear grouping was apparent.

Despite a strong signal in the far IR region, the block scores of Raman multiblock set 1, block 5 did not produce any grouping of samples or strains resembling that of the global scores for corresponding latent variables.

Figure 6.19 shows a selection of block score plots from blocks 3 and 4 of FTIR multiblock set 1, corresponding to the fingerprint/carbohydrate region. Overall, the fingerprint region proved much better in discriminating between groups than the lipid and amide regions, and was also better than the global scores. Notably, the different strains of the OxyCB1 group are distinguishable in the FTIR block scores, as visible in the subplots of Figure 6.19 involving block 4.

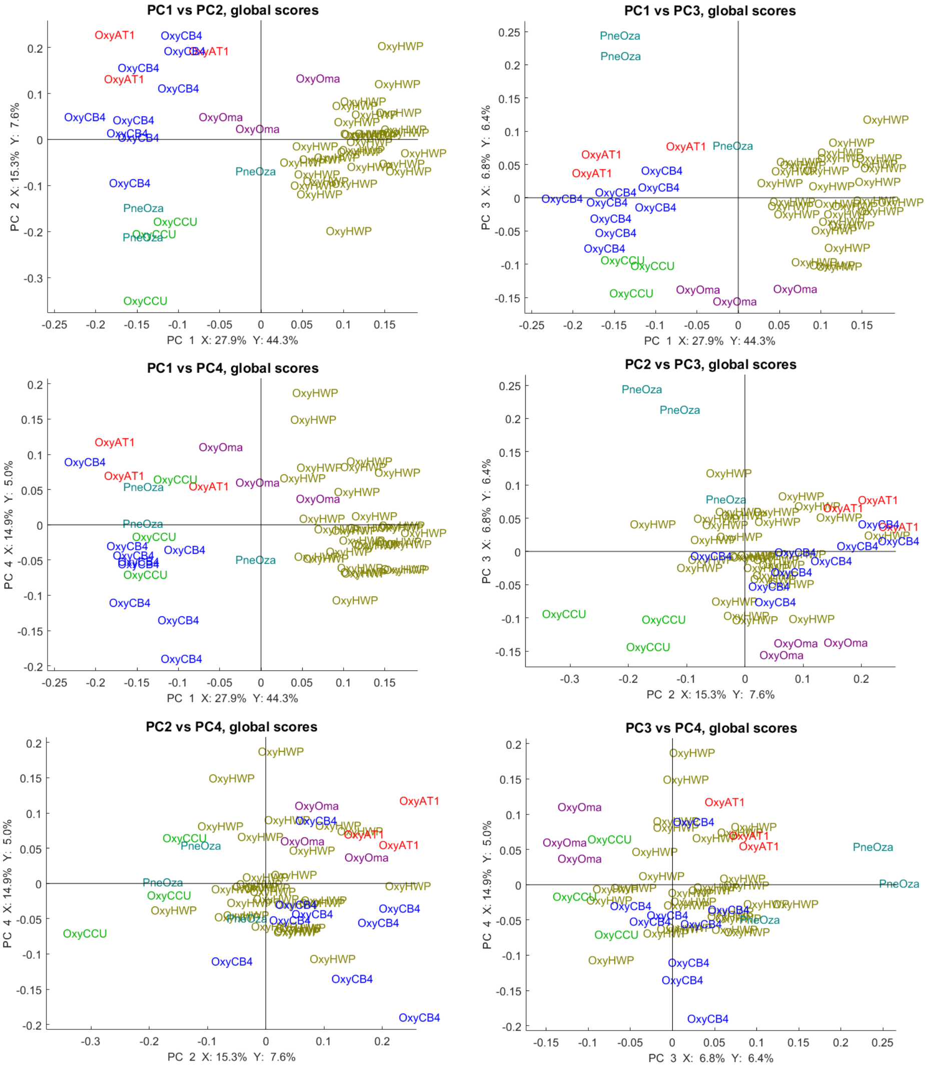


Figure 6.15: Global scores from analysis 4: Split blocks, all data types.

Interestingly, the clearest and most differentiated grouping occurred when the total explained variance between the two PCs plotted against each other was rather low, such as in the subplot of PC3 vs PC4 of FTIR block 4, in the bottom right of Figure 6.19.

6.7.4 Resemblance between MALDI set 1 block scores and global scores

Here we present and compare a selection of block scores from the MALDI block with the global scores of analysis 4. No variables were selected for PC1.

Figure 6.20 shows a selection of block scores from the MALDI block of analysis 4. Some trends in the global scores seemed present here: In both PC2 vs PC3 subplots, OxyCCU and OxyOma were separate and showed clear grouping. OxyHWP was clustered around

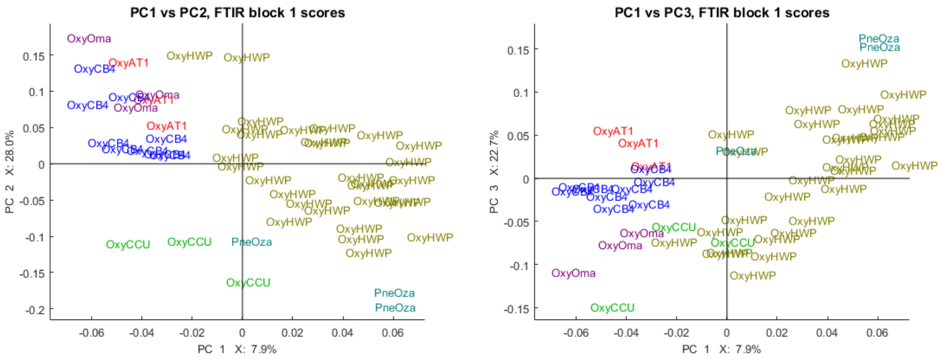


Figure 6.16: Selection of block scores from FTIR multiblock set 1, block 1, corresponding to the lipid region, from analysis 4.

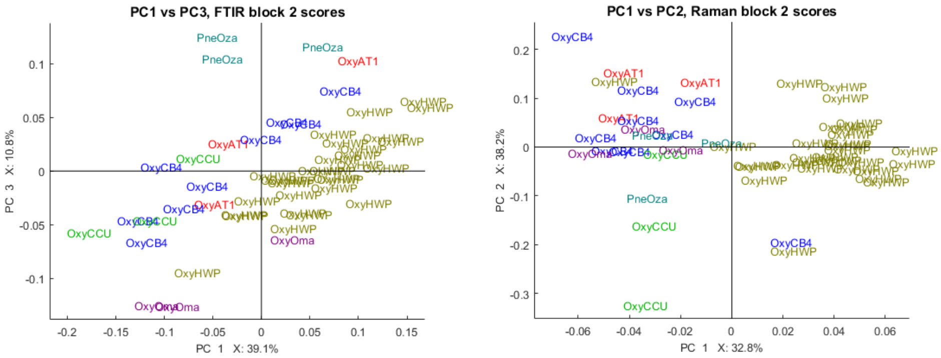


Figure 6.17: Selection of block scores from FTIR and Raman multiblock sets 1, block 2, corresponding to the amide region, from analysis 4.

the origin together with OxyCB4 and OxyAT1, with PneOza slightly separated. In PC2 vs PC4, there was far stronger separation of OxyAT1 and OxyCCU in the block scores than in the corresponding global score plot.

A tendency similar to that seen in blocks 3 and 4 (fingerprint/carbohydrate) of FTIR multiblock set 1 above (Figure 6.19) was also present here. Apart from FTIR multiblock set 1's ability to distinguish between OxyCB1 strains, there was pronounced solitary clustering of most groups in both plots. This was most clear in the subplots of PC3 vs PC4, which showed similar grouping of OxyOma, OxyAT1 and OxyCCU, plus OxyHWP positioned in patterns similar to those in FTIR blocks 3 and 4 with respect to the origin.

6.8 Key results

Our main finding in this chapter was how MALDI was treated differently by the SMBPLSR routine than FTIR/Raman in analysis. Input data from FTIR and Raman seemed to be treated similarly in analysis, while MALDI generated entirely different output. Similar patterns of clustering between groups was stored in different latent variables. This is easily

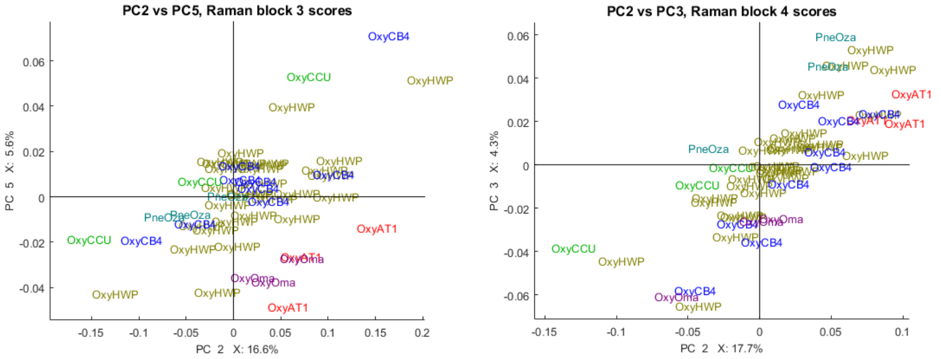


Figure 6.18: Selection of block scores from Raman multiblock set 1, blocks 3 and 4, corresponding to the fingerprint region, from analysis 4.

seen by comparing block score plots from analyses 3 (Section 6.6) and 4 (Section 6.7).

6.8.1 Discrimination

PneOza stood out as being the only group most clearly identified in analysis 3, where especially PC4 of MALDI set 1 (Figure 6.12) and FTIR set 1 (Figure 6.10) encouraged its separation from the other strains in the global scores (Figure 6.9). All other groups were, in one way or another, found more clearly separated and/or clustered in analysis 4.

As briefly mentioned in the final part of Section 6.7, the resemblance was most pronounced between FTIR multiblock set 1, blocks 3 and 4, corresponding to the fingerprint/carbohydrate region, the MALDI block (Figure 6.19), and the global scores. FTIR multiblock set 1, blocks 3 and 4 also showed by far the best discriminative ability of any spectral subregion. This divergence from the general trend of increased scrambling of samples in later latent variables (PC3 and higher) in the global scores, was instead highlighted by Raman multiblock set 1 for most blocks.

The lower fingerprint region corresponding to FTIR multiblock set 1, block 4, showed exceptionally high discriminative ability, being the only block capable of separating the individual OxyCB4 strains.

Closer inspection of Figures 6.15, 6.19 and 6.20 revealed some more detailed correspondence in the latent variables: Grouping of OxyAT1 and its separation from OxyCB4 seemed to be most pronounced whenever PC4 was part of the plot; OxyCCU was most separately grouped from PneOza in PC2 vs PC3; OxyOma showed the least overlap with other groups when PC4 was involved in the FTIR multiblock set 1, blocks 3 and 4 scores, while the best separation by MALDI and global scores was achieved in PC3.

In general, Raman multiblock set 1 data resembled neither its corresponding global scores nor the MALDI set 1 data to any detailed level, at best managing to showcase some clustering, but was unable to separate groups from each other in the block score plots,

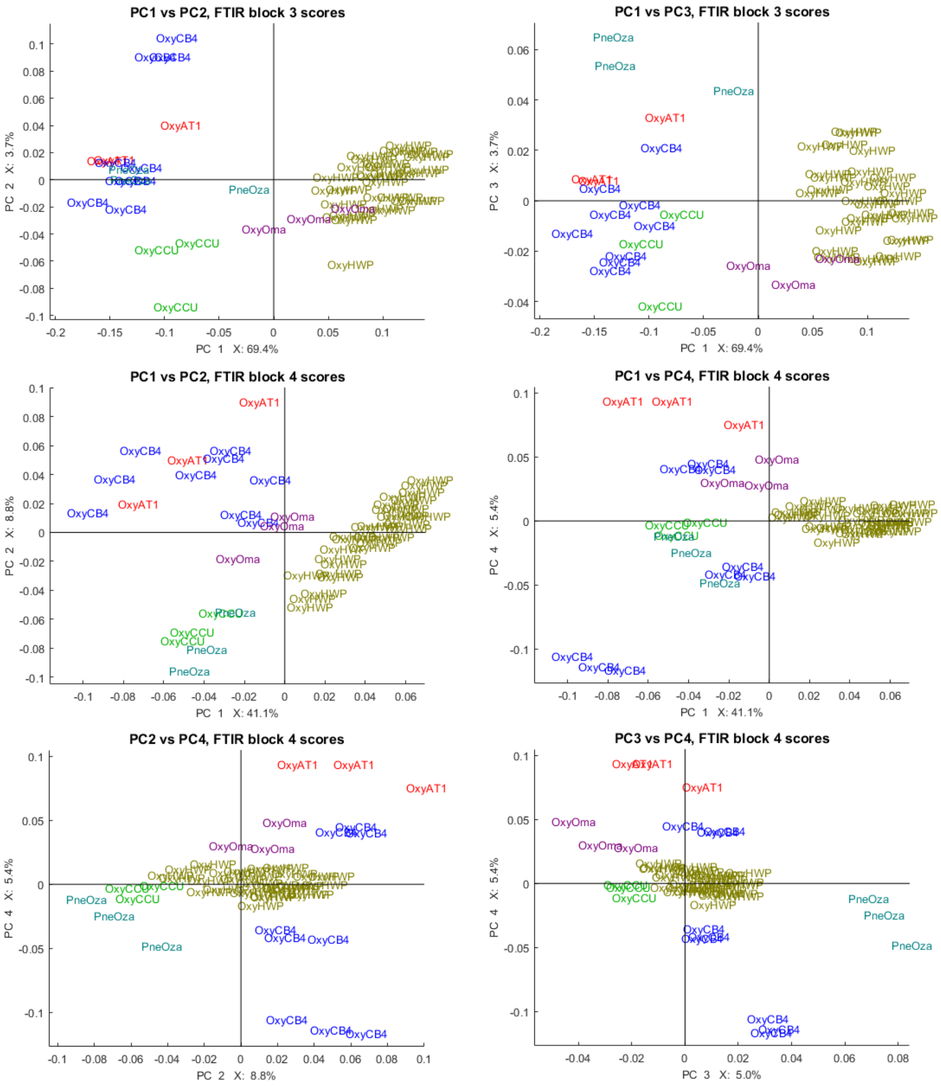


Figure 6.19: Selection of block scores from FTIR multiblock set 1, blocks 3 and 4, corresponding to the fingerprint region, from analysis 4.

not even in the carbohydrate/fingerprint region. Raman multiblock set 1 managed best grouping in the amide region (Figure 6.17, right), a result that corresponded well with the global scores in Figure 6.15, top left.

6.8.2 Success rates

The success rates remained relatively stable for all analyses, although analyses 3 and 4 employed eight-fold CV while analyses 1, 2 and MALDI set 1 only three-fold. Analysis 1 achieved the peak success rate at 96.1%, while success rates were around 85% for the remaining analyses. The high success rate of analysis 1 seems to be owing to the balanced selection of sparse regression coefficients; this was shown to be less balanced in the other analyses.

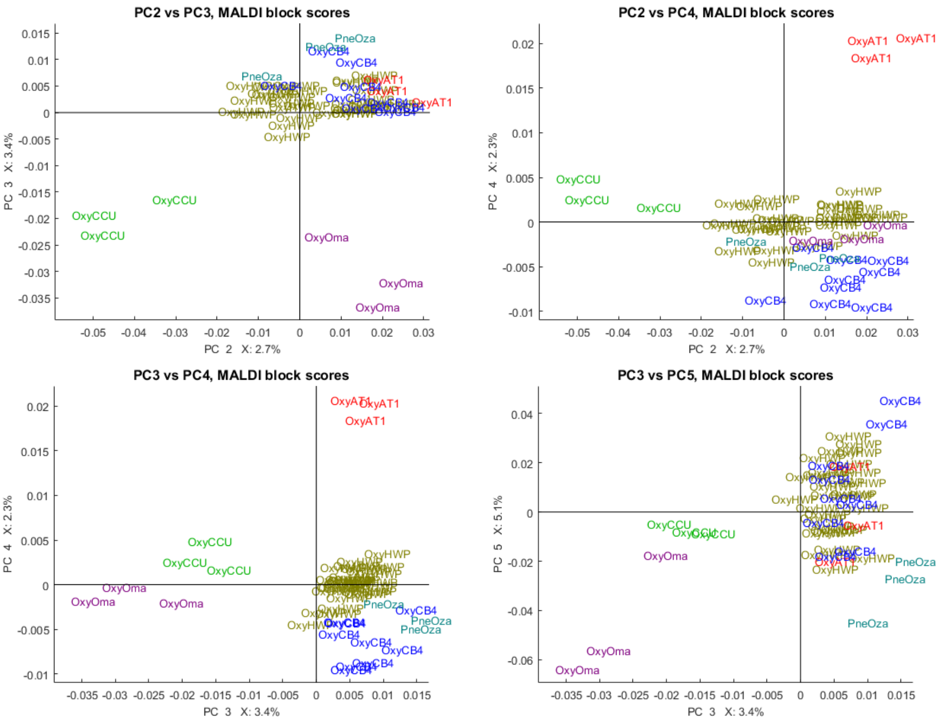


Figure 6.20: Selection of block scores, MALDI set 1, from analysis 4.

6.8.3 Optimal number of latent variables

As shown in Table B.1 in Appendix B below, the optimal number of latent variables were 7 for analysis 3 and 5 for analysis 4. This was also the case for most of the single-type analyses, where AOpt varied from 4 to 7. In this study, we rarely reported plots where latent variables past number 4 were included. Inspecting plots where higher-order PCs were included, revealed little new information. Notable exceptions were PC5 in FTIR and PC6 in MALDI, which both explained OxyAT1 relatively well. Furthermore, OxyCCU was explained well by MALDI PC6 and OxyOma by FTIR PC5. For analysis 8, the only notable information carried by PC5 was how block scores pertaining to the lipid regions of FTIR and MALDI grouped PneOza relatively well.

6.9 Discussion

The reason for the different treatment of MALDI for FTIR/Raman by the SMBPLSR routine, can probably be ascribed to the reduced collinearity in MALDI bar code spectra compared to the FTIR and Raman spectra, i.e. the input data. Peaks in Raman and FTIR possess a certain width, so that even though one or a few samples are slightly displaced, the difference in signal intensity between the samples remains small. In other words: The variance within each variable is small.

The tables are turned completely with the MALDI bar code spectra, for which all peaks are per definition of width one. A displacement of even one unit along the x-axis will remove all collinearity between the samples for the variable and significantly increase its inherent variance (compare unit scales for score plots of MALDI with FTIR/Raman).

6.9.1 Selection of regression coefficients and predictive ability

In Figure 6.4 (analysis 1), relatively strong signals were seen from all groups in several blocks. The success rate of identification was 96.1%. In Figure 6.5 (analysis 2), certain groups, such as OxyOma, only exhibited strong signals for very few variables, whereas strong signals were recorded for many variables in OxyHWP (block 2 and 3). In the corresponding confusion matrix (not shown), no samples were attributed to OxyOma, but two of three OxyOma-samples were misclassified as OxyHWP. The absence of OxyAT1 signals in Figure 6.6 was, similarly to analysis 2, seen as a lack of attributed samples in the corresponding confusion matrix (not shown).

The imbalanced selection of regression coefficients was less pronounced in analysis 3, but still especially evident in the MALDI block, where most strong signals were of the PneOza and OxyCB4 groups. In the corresponding confusion matrix, Figure 6.8, most misclassifications were ascribed to OxyCB4, and PneOza was the only three-sample group completely correctly identified.

Analysis 4 saw an even more balanced selection of regression coefficients (Figure 6.13) than analysis 3. Relatively strong signals were recorded for all blocks and groups, except for the MALDI block, where only signals from OxyHWP and PneOza appeared to be selected. As a result, more groups were completely correctly identified and misclassifications were less scattered, as shown in the corresponding confusion matrix (Figure 6.14.

Having misclassifications ascribed to the largest groups is not entirely unexpected, as the higher number of samples within the group makes the probability of recording a strong signal from the group as a whole a lot higher than gathering them from smaller groups. Still, as the discussion above indicates: The multi-type, multiblock, descriptor data in analysis 4 seemed to stabilise the selection of sparse regression coefficients in disfavour of larger groups, except for the MALDI block.

The group-wise selective selection of regression coefficients in analyses 3 and 4 further emphasises the different treatment of MALDI data from FTIR/Raman data by the SMBPLSR routine. Further study is required to address this problem.

6.9.2 Block variable selection in single-type versus multi-type descriptor data SMBPLSR analyses

In analysis 1, variables were not selected for most of the PCs in lipid and amide regions of FTIR multiblock set 1, and variables for PC1 were only selected in the upper fingerprint region corresponding to block 3; only variables of PCs 1 and 8 were selected for block 3. In general, variables were selected far more scarcely in analysis 1 than in analyses 3 and 4, probably owing to the number of variables being so much lower in the former than the latter two, attributable to the presence of the comparatively large MALDI data set.

The presence of the MALDI data set in practice allowed more than 10% of the variables in FTIR multiblock set 1 to be selected, resulting in a de facto 'less sparse' variable selection in FTIR multiblock set 1. This pattern is repeated in a similar fashion for Raman multiblock set 1.

One could then enquire whether a battery of single data type SMBPLSR analyses similar to analyses 1 and 2 described above, but with a higher percentage of variables retained, say 20 or 30%, would allow for an equally good or better overall solution of the problem posed in this thesis. Or whether the lack of common global scores, loadings and weights in single-type descriptor analyses would obscure the comparative abilities of the multi-type descriptor analysis shown and discussed above.

Chapter 7

Results and discussion, experimental design

This chapter deals with the second part of the topic for this thesis: How sparse multiblock methods perform in comparing phylogenetic similarities and differences assessed by different phenotypic methods. As such, all mention of data sets and spectra in this chapter refers to the *B. subtilis*/*E. coli* data set described in Section 2.2 unless otherwise stated. Data sets are referred to as multiblock sets or simply sets, depending on whether each data type is represented by one block, or their respective number of selected blocks. For example, FTIR multiblock set 2 consists of four blocks, while FTIR set 2 means that these four blocks have been merged into one.

In Sections 7.1 and 7.2, details of the experimental setup are presented. It is explained how blocks were selected and demarcated to increase interpretability and computational speed. In the following four sections, we show and discuss the results of four different SMBPLSR analyses. In Section 7.8, results of the analyses are summarised and the most important findings highlighted, before being discussed in Section 7.9.

7.1 Experimental setup

7.1.1 Preprocessed spectra

All spectra in this section were preprocessed identically to their corresponding *Klebsiella* data set-data types, that is: FTIR by a 2^{nd} order polynomial, 2^{nd} degree derivative SG filter with a window size of 9, followed by standard EMSC; Raman by 2^{nd} degree polynomial SG filter of window size 9 but no derivative, followed by standard EMSC; MALDI low and medium masses by an SG filter of window size 21, followed by normalisation and baseline correction [2]. The FTIR and Raman spectra were processed by in-house or Saisir codes [78]. The preprocessed spectra are shown in Figure 7.1a and b. The MALDI spectra were already preprocessed upon receipt. The preprocessed MALDI spectra constituted an intermediate step in data preprocessing and were not sent to further pretreatment or analysis. Instead, a peak table containing the 30 highest-intensity peaks after preprocessing was transformed into a bar code table; their relative intensities were as such omitted in further analysis. The bar code tables were used in further analysis. The bar code spectra are shown in Figure 7.1c and d.

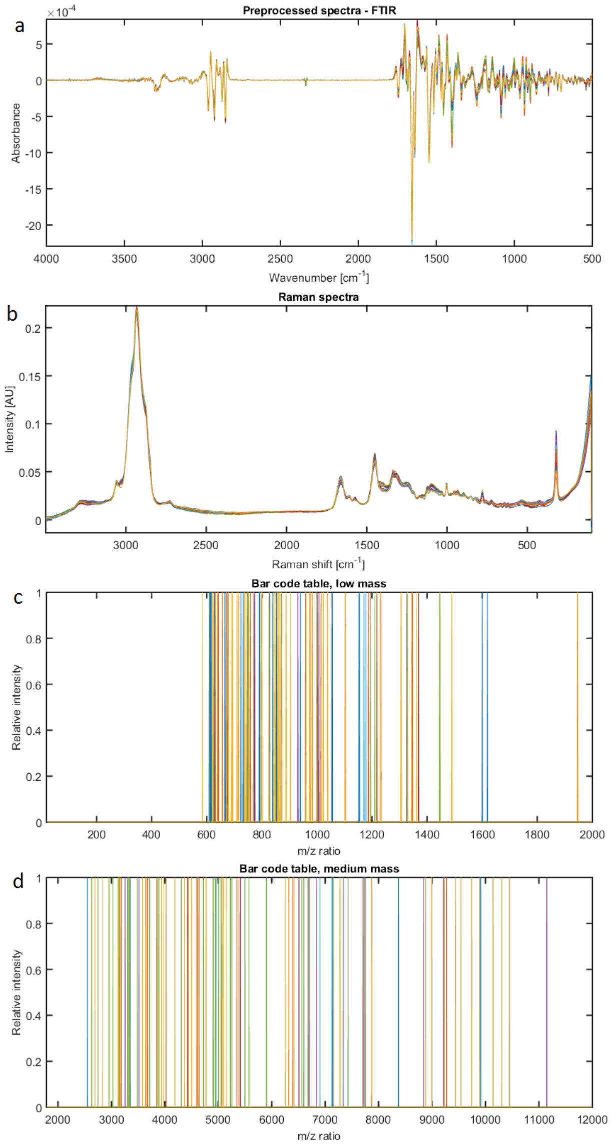


Figure 7.1: Preprocessed spectra (FTIR and Raman), and bar code plots (MALDI, low and medium mass) from the *Bacillus*/*Escherichia* data set. a: FTIR, b: Raman, c: MALDI, low mass, d: MALDI, medium mass.

7.1.2 Selecting relevant spectral regions

All four data types were analysed with SPLSR and frequency plots were set up using full leave-one-out CMV in order to select relevant spectral regions for further analysis. For all data types, the sparsity parameter was set by soft thresholding, so that 90% of the variables were discarded and 10 PCs were calculated. Figure 7.2 shows the resultant frequency plots. Note that the x-axes are reverted for all subplots.

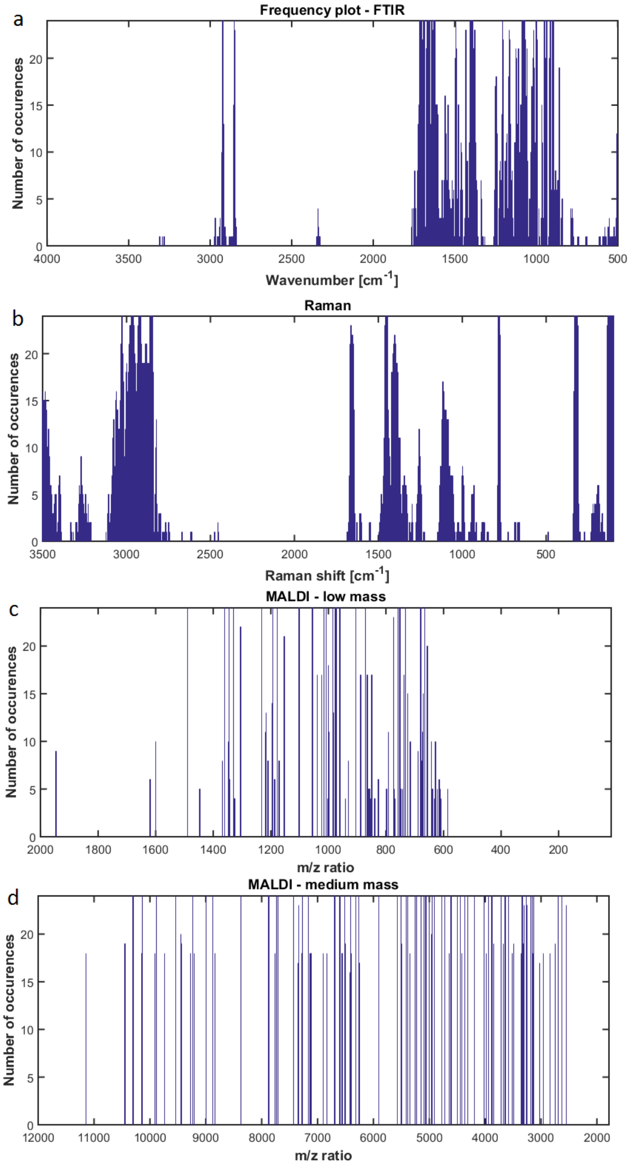


Figure 7.2: Frequency plot from full leave-one-out CMV (24-fold) on all data types. Each data type was analysed by the SPLSR model described in Section 4.3.3. For all analyses, 90% of variables were discarded by soft thresholding. a: FTIR, b: Raman, c: MALDI, low mass, d: MALDI, medium mass.

Figure 7.2a shows the resultant frequency plot from 24-fold CMV of the FTIR data set. Based on the grouping of often selected variables, FTIR multiblock set 2, consisting of four blocks, was assembled:

- Block 1 from 2950-2800 cm^{-1}.
- Block 2 from 1750-1600 cm^{-1}.
- Block 3 from 1500-1350 cm^{-1}.
- Block 4 from 1250-850 cm^{-1}.

Figure 7.2b shows the resultant frequency plot from 24-fold CMV of Raman data. Note that selection of frequencies corresponding to proteins were fewer here than for the FTIR frequency plot, and even for the *Klebsiella* Raman frequency plot, shown in Figure 6.1b. On this basis, Raman multiblock set 2, consisting of four blocks, was assembled:

- Block 1 from 3100-2800 cm^{-1}.
- Block 2 from 1700-1350 cm^{-1}.
- Block 3 from 1010-750 cm^{-1}.
- Block 4 from 350-100 cm^{-1}.

Figure 7.2c shows the resultant frequency plot from 24-fold CMV of MALDI, low mass data. Visual inspection of the plot gave no clear indication of grouping, but did however reveal the redundancy of both extremities of the spectra; the $\frac{m}{z}$ regions 0-500 and 1500-2000, in which almost no variable selection occurred and no variables were selected more than 10/24 times. Nevertheless, we chose not to discard any part of this plot in further analysis.

Figure 7.2d shows the resultant frequency plot from 24-fold CMV of MALDI data, medium mass. No region stood out as particularly information-rich, although there was a slightly higher prevalence of frequently selected variables in the $\frac{m}{z}$ region between 2500 and 5500. The $\frac{m}{z}$ region above 8000 was rather sparsely populated by high-frequency peaks. Also, note that there were no variables selected less than 15 times, making any grouping or incision even harder. As such, no incisions were made and no regions discarded before further analysis.

7.2 Predictive abilities and weighting of regression coefficients

7.2.1 Correspondence between strain and script names

Table A.2 in Appendix A shows the correspondence between the sample strain names shown in Section 2.2.1 and their given names in the data script. Names were altered to ensure concurrence between different data types, and in such a way that important information remained in the name. In the following plots, the first letter of each sample denotes the species: E for *Escherichia* and B for *Bacillus*. The two-digit number denotes the growth temperature of the sample.

7.2.2 Description of input data

After preliminary analysis was performed and blocks selected, we proceeded by analysing sparse multiblock routines on all data types. These analyses were

- Analysis 5: FTIR multiblock set 2[1], four blocks as listed above.
- Analysis 6: Raman multiblock set 2[2], four blocks as listed above.
- Analysis 7: FTIR set 2, Raman set 2, and MALDI data, both low and medium mass, in total four blocks.
- Analysis 8: FTIR multiblock set 2, Raman multiblock set 2 and MALDI data, both low and medium mass, ten blocks. No concatenation.

The MALDI data was neither split into blocks, nor were any regions removed, so we chose not to run additional analyses on these sets. We did, however, gather information about their global scores for further comparison and analysis. For all the analyses described, we kept the sparsity parameters used in Chapter 6, discarding 90% of variables and calculating 10 PCs. Model prediction was performed using three-fold Venetian Blinds CV for all analyses, also those with multiple types of descriptor data. In the following, these will simply be referred to as the MALDI, low or medium mass, analyses. Table B.2 in B lists the global explained variance by the latent variables from analyses 7 and 8.

To shed a first ray of light upon the ability of our phenotyping methods' ability to discriminate between species and their medial growth temperature, we chose to examine and compare the selected regression coefficients and their relative weighting for each analysis specified above.

[1] FTIR set 2 means one data block consisting of the four data blocks in FTIR multiblock set 2.

[2] Raman set 2 means one data block consisting of the four data blocks in Raman multiblock set 2.

7.3 Analysis 5: FTIR multiblock set 2

Figure 7.3 shows the selected variables from analysis 5 on FTIR multiblock set 2. The highest number of variables were selected from block 1 and 2, covering the lipid and amid regions, respectively. The strongest signals were found in the amide and carbohydrate regions. Misclassifications occurred chiefly in the mid-temperature range, but all samples of B43, E43 and E25 were correctly identified (figure not shown). The success rate of identification was 70.8%.

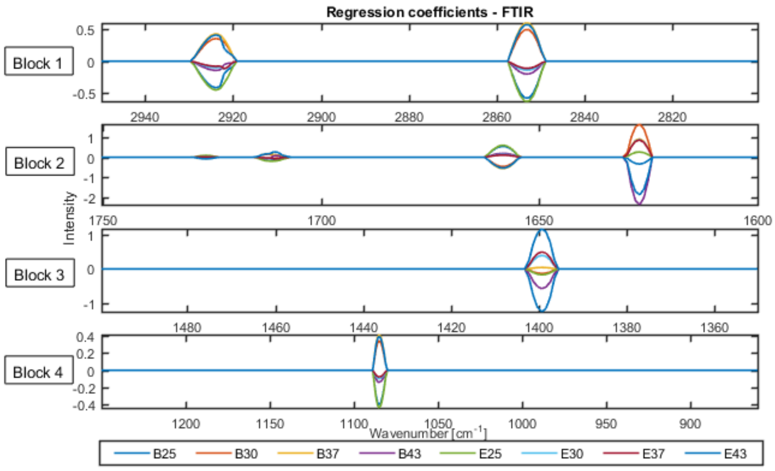


Figure 7.3: Selected sparse regression coefficients for SMBPLSR from analysis 5: FTIR multiblock set 2.

7.4 Analysis 6: Raman multiblock set 2

Figure 7.4 shows that the selected variables of analysis 6 (Raman multiblock set 2) were gathered from all blocks, though in relatively low numbers in block 4. The signal was by far the strongest in the upper fingerprint/carbohydrate region of block 3 and displayed by E43. The E30 group also showed strong signals in both block 1 and 2. The success rate was 54.2%; most misclassifications occurred in *B. subtilis* groups, some of which were incorrectly classified as *E. coli.* All E30 and E43 samples were correctly identified. However, both groups received several additional misclassified samples from *B. subtilis*, suggesting that the dominant signals visible in the regression coefficients of these groups might be the cause of the misclassifications.

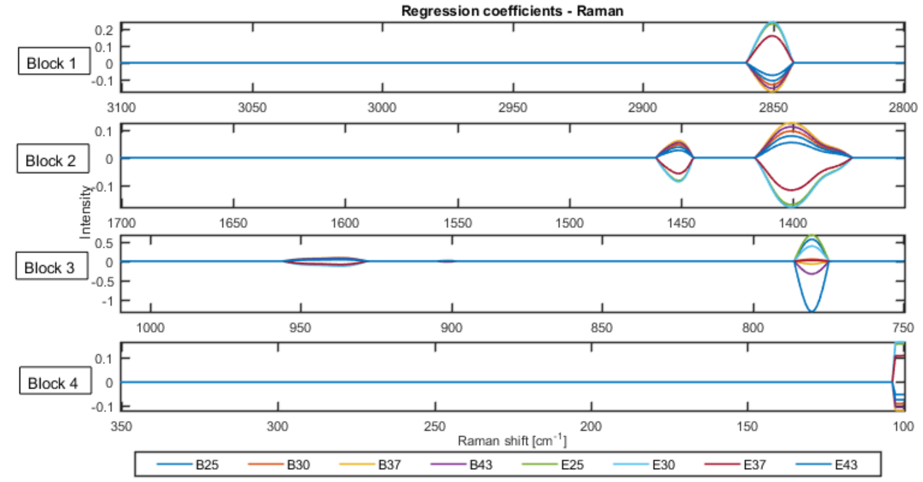


Figure 7.4: Selected sparse regression coefficients for SMBPLSR from analysis 6: Raman multiblock set 2.

7.5 MALDI, low and medium mass analyses

Figure 7.5a and b shows the selected regression coefficients from the MALDI, low and medium mass analyses, respectively. The selection of variables seemed to correlate with high density regions in their respective bar and frequency plots (Figures 7.2c and d, and 7.1).

For both analyses, regression coefficients were only selected from some of the groups. Regression coefficients were primarily attributed to B37, B43, E37 and E43 in subplot a, while B43, E43 and E30 appear to be the only groups for which regression coefficients were selected in subplot b. Interestingly, in the MALDI, low mass, analysis no samples were identified as B43 (figure not shown), while the apparently absent B30 group was completely correctly identified and had misclassifications attributed to it from several other groups. In the analysis of MALDI, medium mass data, no samples were identified as B25, B30, E25 or E37 (figure not shown). Why samples were identified as B37, which was absent in the regression coefficient plots, remains unknown.

The unexpected selection of regression coefficients manifested in particularly low success rates for these analyses: 29.2% for low and 25.0% for medium mass.

7.6 Analysis 7: Concatenated data blocks, all data types

Figure 7.6 shows the resultant variable selection from analysis 7. The vast majority of FTIR variables were - rather unexpectedly, considering especially the relatively weak signal

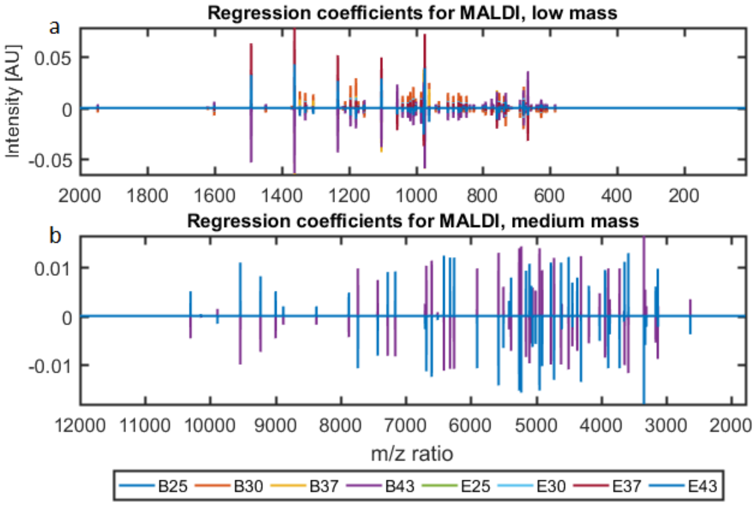


Figure 7.5: Selected sparse regression coefficients for SMBPLSR on MALDI, low and medium mass. a: low mass, b: medium mass.

in the lower fingerprint region of Figure 7.3 - selected from the protein and fingerprint regions. Almost no variables were selected from the lipid region.

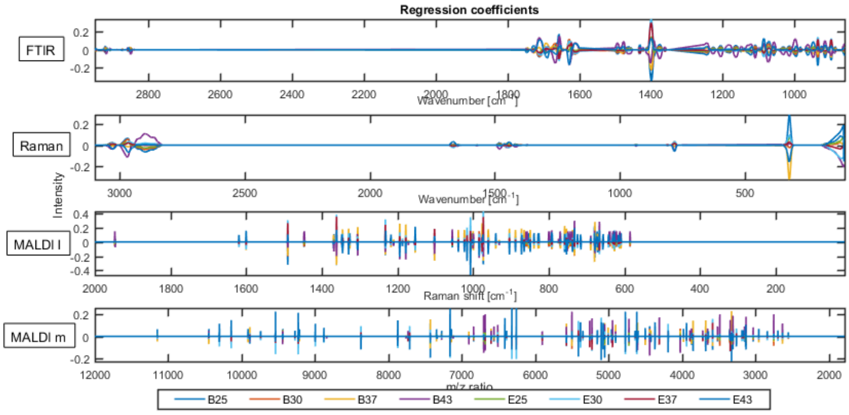
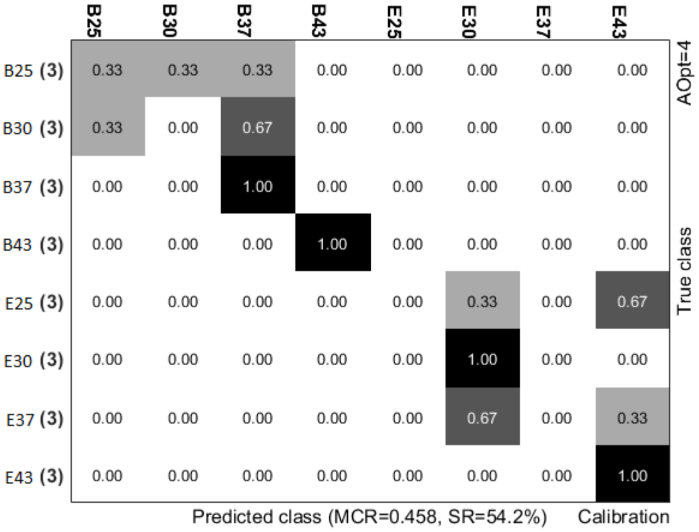


Figure 7.6: Selected sparse regression coefficients for SMBPLSR from analysis 7: FTIR set 2, Raman set 2, and MALDI, low and medium masses.

Variables selected from the Raman block also differed from the selection shown in Figure 7.4; variable selection was now most pronounced in the extreme ends of the spectrum, even though these signals, particularly in block 4, previously were among the weakest. The strong signal of block 3, attributed to be misclassifying *B. subtilis* as *E. coli*, now barely appeared.

E30 and E43 were particularly prominent in the MALDI blocks. This probably caused the numerous misclassifications attributed to them in the corresponding confusion matrix, displayed in Figure 7.7. The weak signals from E25 (light green) and its relative absence from the MALDI blocks in Figure 7.6 saw an effect in the confusion matrix: No samples attributed to this group. Apart from this, regression coefficients appeared to be selected in a relatively balanced manner in the MALDI blocks, as compared to their selection in the previous single-type MALDI runs (Figure 7.5).

The confusion matrix of predicted groups from the analysis 7 is shown in Figure 7.7. The often-correct identification of E43 and B43 from analyses 5 and 6 was continued. Even though the success rate was only 54.2%, there was no misclassification between species, and four groups were completely correctly identified.



	B25	B30	B37	B43	E25	E30	E37	E43
B25 **(3)**	0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00
B30 **(3)**	0.33	0.00	0.67	0.00	0.00	0.00	0.00	0.00
B37 **(3)**	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
B43 **(3)**	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
E25 **(3)**	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.67
E30 **(3)**	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
E37 **(3)**	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.33
E43 **(3)**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 7.7: Confusion matrix from analysis 7. The CV configuration was three-fold Venetian Blinds.

The global scores from analysis 7 are shown in Figure 7.8. There was clear between-species grouping for the first three subplots, but this was discontinued for the final three. B43 was grouped particularly well for all subplots involving PC3.

From Figure 7.9, we see that the block scores of FTIR set 2 showed similar grouping patterns for corresponding subplots, although achieving slightly better clustering of the groups. Visible in subplot 3 is also that involving PC3, such as in the global scores, isolated B43 from the rest of the samples. In this case, however, E43 was also isolated. This complemented the trend from the scores obtained in analysis 5 (not shown).

Only one block score plot was selected from Raman set 2 in analysis 7. The plot is shown in Figure 7.10.

This was the best grouping seen in of all combinations of plots of PC1-4 against one another, yet it barely showed a pattern in the B43/E43 groups, as visible in global scores and FTIR set 2 (Figure 7.9). Grouping was altogether more pronounced in the global scores of analysis 6 (not shown) than for Raman set 2 in analysis 7. A probable

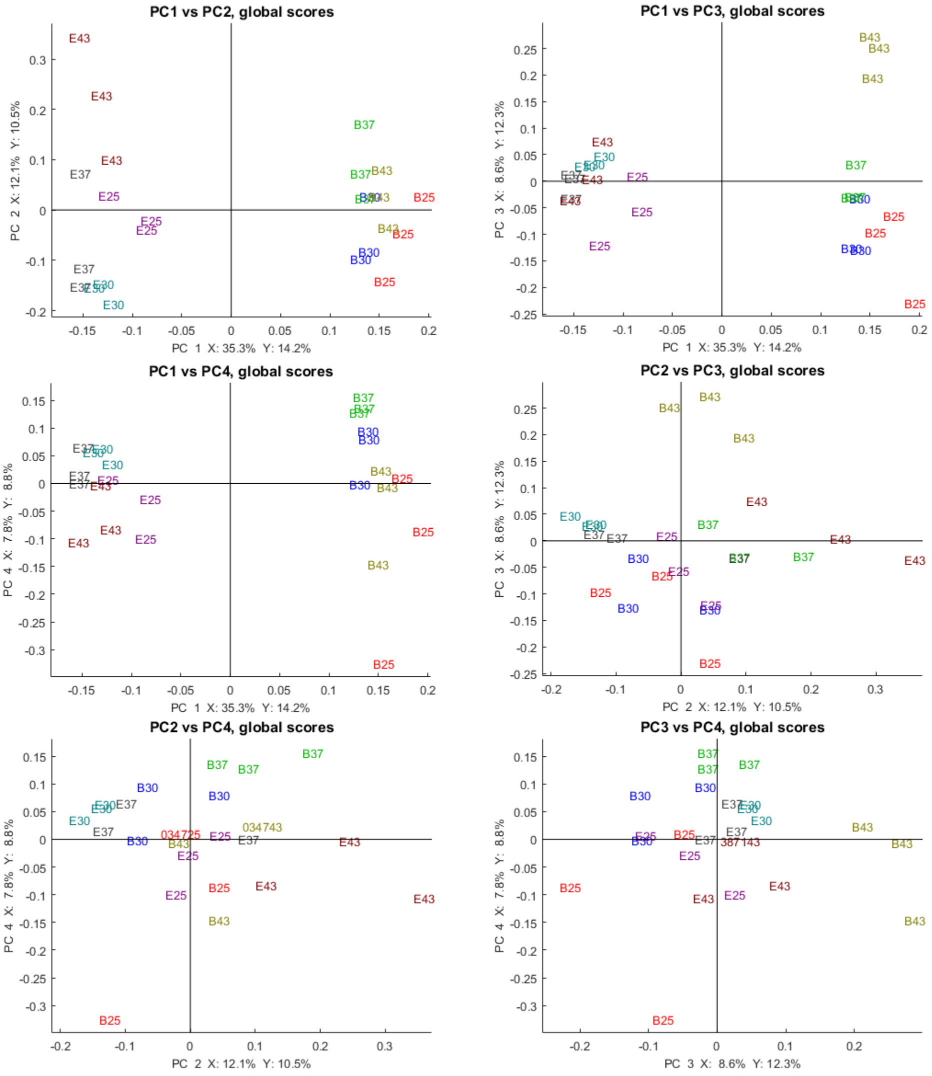


Figure 7.8: Global scores from analysis 7: Concatenated blocks, all data types.

reason for the poor grouping was the absence of a dominant variable. Both PC1 and PC2 explained more than 30% of the total variance, while PC3 explained 15.4%. The resultant diagonal inclination apparent in Figure 7.10 made groups and patterns less discernible than a plot with a dominant variable would have been, and likely also caused the between-species misclassification mentioned in Section 7.4 (analysis 6). In general, the visual correspondence between Figure 7.10 and the global scores of analysis 7 was low.

A recurring pattern with MALDI data was the lack of a dominant latent variable; at most, one or two exceeded 10% of the total explained variance. This trend was less clear in analysis 7, where the first PC explained 13.6% and PC4 19.5% of the total variance, as opposed to less than 8% in the *Klebsiella* data set. Selected block score plots are shown in Figure 7.11. For MALDI, low mass, separation between species was enabled, but intermingled samples or outliers obscured further separation.

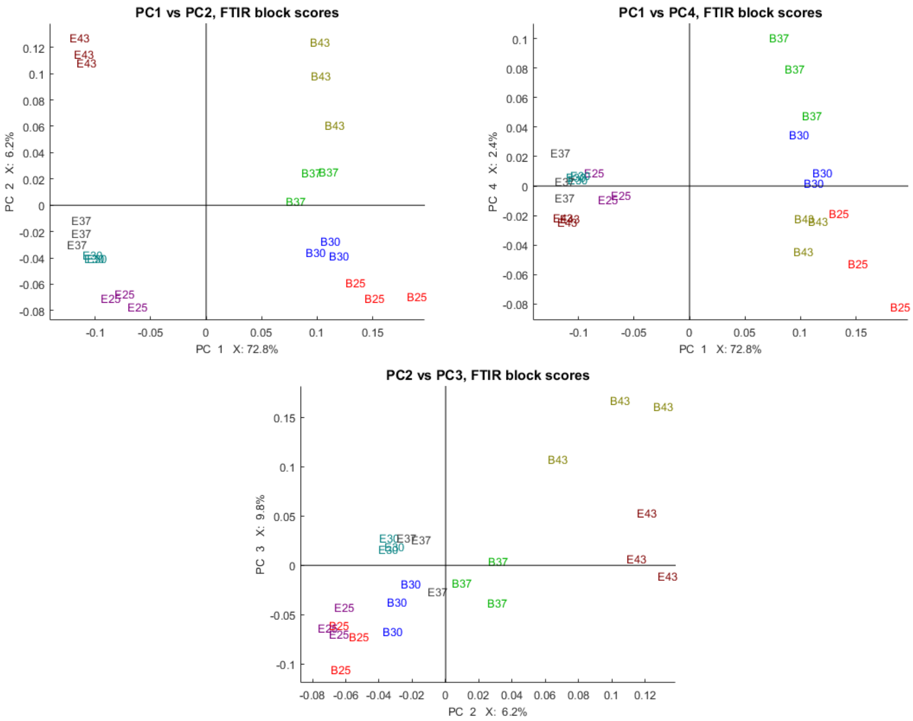


Figure 7.9: Selection of block scores from analysis 7: FTIR set 2.

The predictive results of the MALDI, medium mass, data set was overall better than the low mass one, in accordance with reports of main peaks in the MALDI spectrum of similar microorganisms being found in the $\frac{m}{z}$ range between 3 and 10 kDa [2, 38, 79]. The block score plots of MALDI, medium mass are shown in Figure 7.12. PC1 was relatively dominant, explaining 21.4% of the total variance in the data set alone. However, the total explained variance by the 10 PCs calculated was only 54.3%. Still, grouping was more in accordance with the global scores (Figure 7.8) and FTIR set 2 (Figure 7.9), for example in that B43 was consistently grouped in subplots involving PC3, and that species were separated in subplots 1 and 2.

7.7 Analysis 8: Split data blocks, all data types

The regression coefficients selected from analysis 8 are shown in Figure 7.13. Regression coefficients were selected from the FTIR and Raman blocks (1-8) in such high numbers that the sparsity parameters were almost obviated for these blocks, suggesting a relatively strong signal in this blocks as compared to the MALDI, low and medium mass, blocks. This was allowed by the model because the total number of variables in the first eight blocks made up 11% of the total number of variables, and that most variables in the MALDI blocks was zero and as such could not be selected.

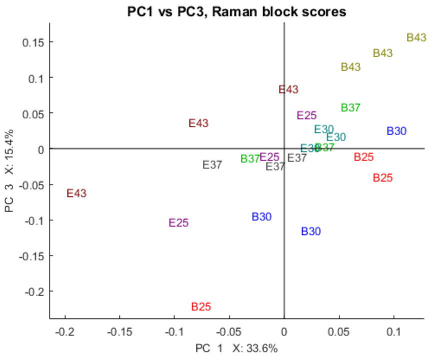


Figure 7.10: Selection of block scores, analysis 7: Raman set 2.

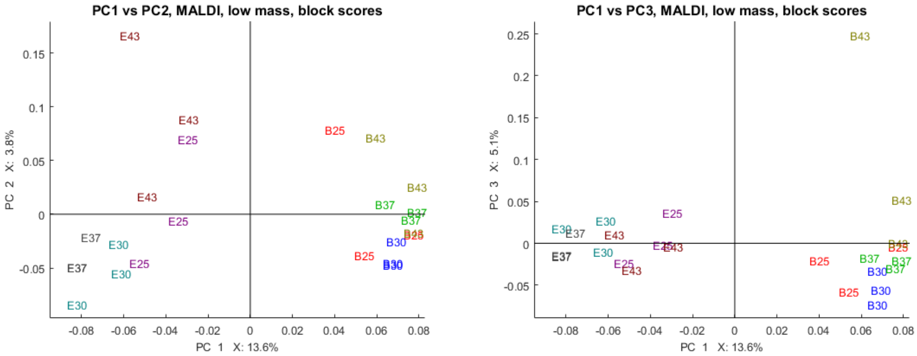
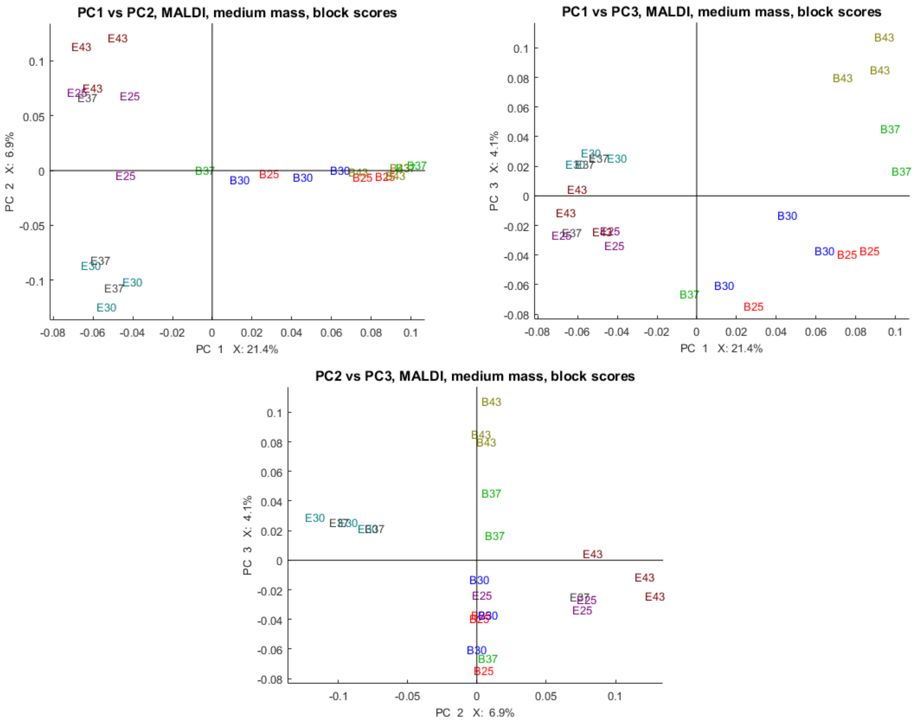


Figure 7.11: Selection of block scores, analysis 7: MALDI, low mass.

The tendency in the *Klebsiella* results Section (6.2) of devaluing MALDI data in a split block-setting, was discontinued here; signals from both low and medium mass spectra were more abundant, and several were among the strongest in the whole set. Furthermore, no groups seem particularly prominent apart from E25 in the MALDI blocks.

Figure 7.14 shows that the prediction using split blocks was fairly similar to that using concatenated blocks (Figure 7.7). A slightly lower success rate at 50.0% and only two groups - again B43 and E43 - completely correctly classified, the prediction still displayed less congregation around certain groups than its concatenated sibling; no groups had zero samples attributed to it and both B25, B37 and E25 had two of three samples correctly identified, suggesting a more balanced outcome for this analysis compared to analysis 7. The argument for stably selected regression coefficients as a result of multi-type multiblock input data made in Section 6.9 is strengthened by this finding.

Comparison of the global scores of analysis 7 (Figure 7.8) and analysis 8 (Figure 7.15) showed similar, but slightly more pronounced grouping patterns in the latter than the former. The clustering of each group was closer and better defined in analysis 8. This could be ascribed to the more dominant first latent variable in analysis 8, explaining 56.9% of the total variance, compared to the first latent variable in analysis 7 explaining only


PC1 vs PC2, MALDI, medium mass, block scores

PC1 vs PC3, MALDI, medium mass, block scores

PC2 vs PC3, MALDI, medium mass, block scores

Figure 7.12: Selection of block scores, analysis 7: MALDI, medium mass.

35.3%. As in earlier analyses, the separation of B43 and E43 was by and large the most often present, although PC4 appeared to carry information making discrimination of B37 possible.

7.7.1 Discriminative abilities of the lipid region of FTIR and Raman multiblock set 2

Figures 7.16 and 7.17 show a selection of block scores from FTIR multiblock set 1 and Raman multiblock set 1, blocks 1, respectively. Raman showed weaker overall grouping and clustering in the lipid region than did the corresponding global scores. The global grouping and clustering patterns were in turn slightly weaker than their FTIR peers.

7.7.2 Discriminative abilities of the amide region of FTIR and Raman multiblock set 2

The amide region (block 2) of the FTIR data appeared to be holding interesting information, particularly about the *Bacillus* groups, all of which were clearly distinguishable, as shown in Figure 7.18. On the other hand, very little valuable information could be obtained from the plot of Raman data in Figure 7.19. No patterns seemed to be especially informative for this plot.

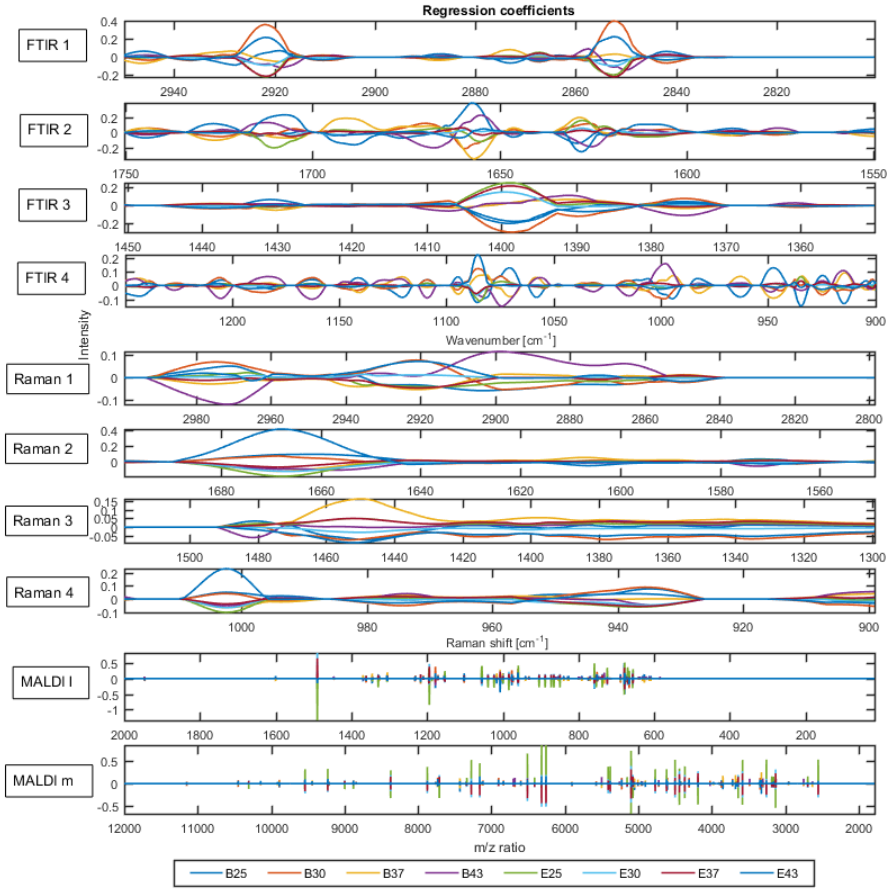
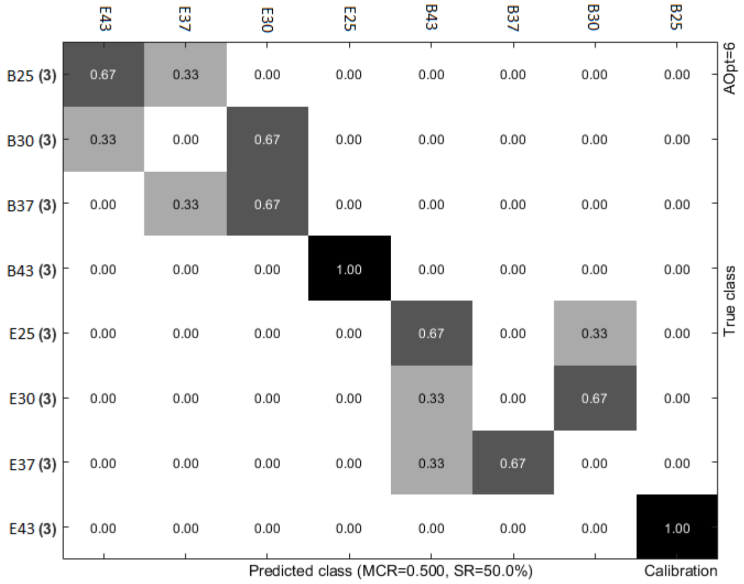


Figure 7.13: Regression coefficients for SMBPLSR from analysis 8: FTIR multiblock set 2, Raman multiblock set 2, and MALDI, low and medium masses.

7.7.3 Discriminative abilities of the carbohydrate/fingerprint region of FTIR and Raman multiblock set 2

Selections of block score plots for block 3 of FTIR multiblock set 2 and Raman multiblock set 2 are shown in Figure 7.20 and 7.21, respectively. For FTIR, information carried by the upper fingerprint/carbohydrate region highlighted B37 as a distinct group. This information was also carried by the global score subplots involving PC3, albeit to a lesser extent (Figure 7.15). No variables were selected from PC2 for any of the blocks; for Raman, nor for PC5. However, PC6 turned out to explain 14.1% of the total block variance and as such a subplot of PC1 vs PC6 was examined and reported. To the extent that any new information was provided by PC6, it seems that it might have forced a slight split between the previously less distinguishable *E. coli* strains.

The block score plots from the lower carbohydrate/fingerprint region, corresponding to block 4 of FTIR and Raman multiblock sets 2, showed the best discriminatory potential



	E43	E37	E30	E25	B43	B37	B30	B25
B25 (3)	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00
B30 (3)	0.33	0.00	0.67	0.00	0.00	0.00	0.00	0.00
B37 (3)	0.00	0.33	0.67	0.00	0.00	0.00	0.00	0.00
B43 (3)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
E25 (3)	0.00	0.00	0.00	0.00	0.67	0.00	0.33	0.00
E30 (3)	0.00	0.00	0.00	0.00	0.33	0.00	0.67	0.00
E37 (3)	0.00	0.00	0.00	0.00	0.33	0.67	0.00	0.00
E43 (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 7.14: Confusion matrix from analysis 8. The CV configuration was three-fold Venetian Blinds.

among the subregions of the FTIR/Raman absorbance spectra. The plots are shown in Figures 7.22 and 7.23.

The strain B37 is distinctly separated in all subplots of Figure 7.22 but especially in the subplot of PC2 vs PC3. Furthermore, E25 and B25 were all also clearly separated and clustered in all subplots. PC2 vs PC3 showed almost perfect grouping, and only struggled to discern between E30 and E37, which have been nearly inseparable in all plots, including the global scores.

Although not as clear-cut as FTIR, the Raman block score plots in Figure 7.23 display good separation compared to other regions, especially in its distinction of B43, and, to some degree E43 and B37. It is also worth noting that PC1 explains 84.2% of the total variance for the block, easily a new record for Raman and an indication of a higher degree of coherence between samples in this spectral region than any other.

The wealth of information provided in the block scores of analysis 8 was contrasted by the lack of such in analyses 5 and 6. Oftentimes, variables of the most highly explanatory PCs of the blocks, and particularly blocks 3 and 4 of the FTIR/Raman multiblock sets 2 examined in analyses 5 and 6, were not selected. For analysis 5, block 4, no variables were selected for PC1, which explained 67.4%; selection only occurred for PCs 3, 8 and 10. These PCs explained 9.1% of the total variance between them. This suggests that the high threshold (90% of the variables to be discarded) removed potentially very valuable information from analyses 5 and 6, some of which has been preserved in analysis 8, and, presumably, analysis 7.

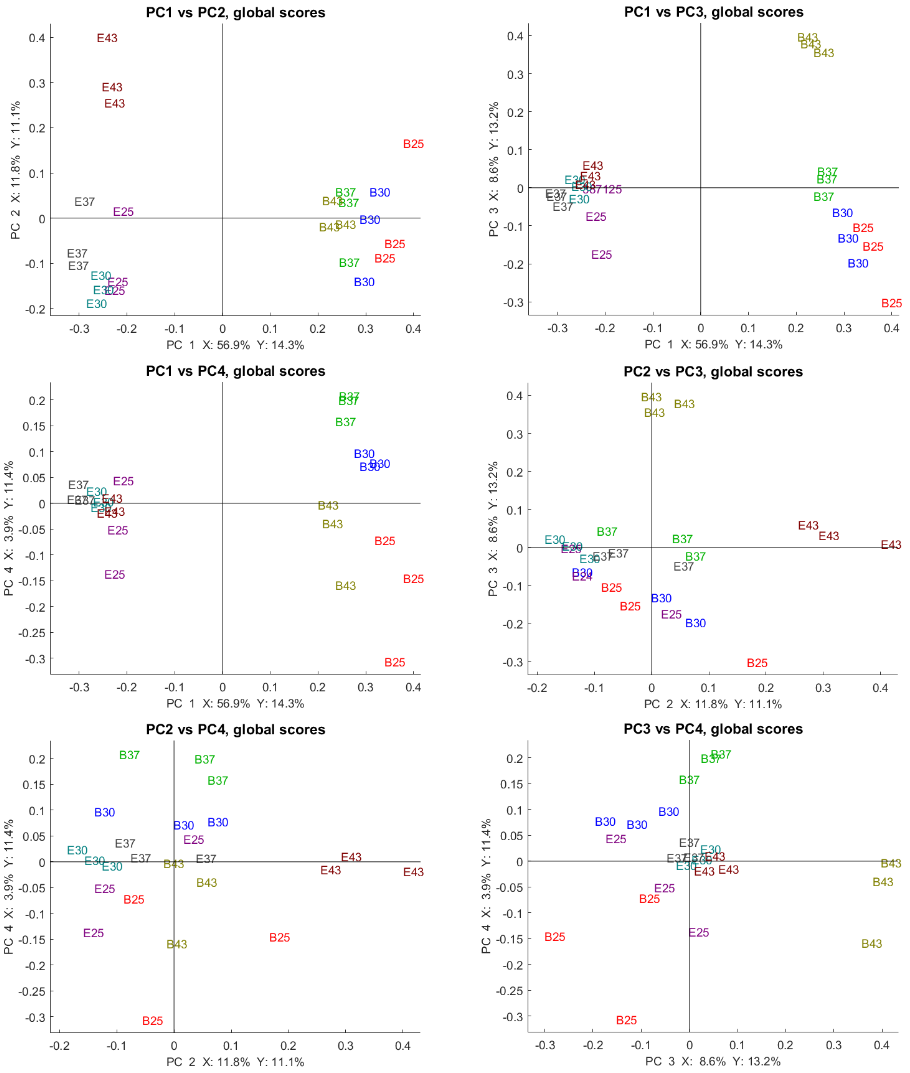


Figure 7.15: Global scores, analysis 8: Split blocks, all data types.

7.7.4 Discriminative abilities of MALDI, low and medium mass

The MALDI low mass block scores in Figure 7.24 were the foremost exponents of a the recurring problem of outliers in in the MALDI blocks. Additionally, we here saw some inexplicable orientation of *Bacillus* strains in both subplots. Not even the pattern of B43 being clearly separated in PC1 vs PC3, present in almost all blocks score plots both in analyses 7 and 8, was continued here. Neither the main (PC1: 13.2%, and PC4: 15.4%) nor the aggregate of 10 PCs (67.3%) seemed to explain the total variance very well. This probably accounted for some of the poor grouping. It was unclear why all *Bacillus* groups were explained almost equally well by both PC1 and PC2 while the *Escherichia* samples behaved normally.

Despite its rather unusual appearance, the selected block score plots for MALDI, medium mass, shown in Figure 7.25, performed the best of any MALDI score plot in

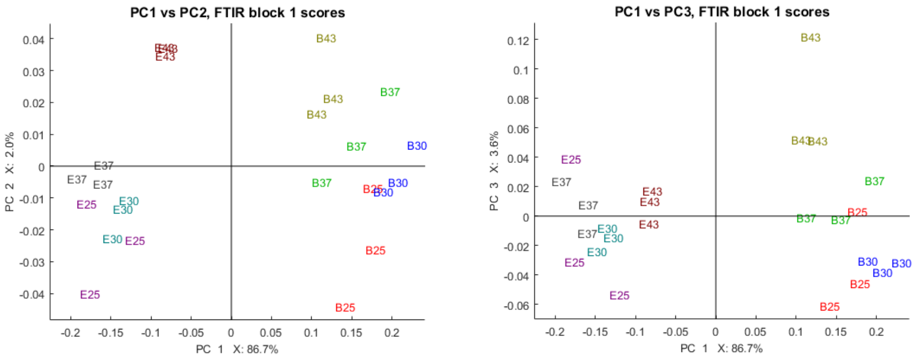


Figure 7.16: Selection of block scores, analysis 8: FTIR multiblock set 2, block 1.

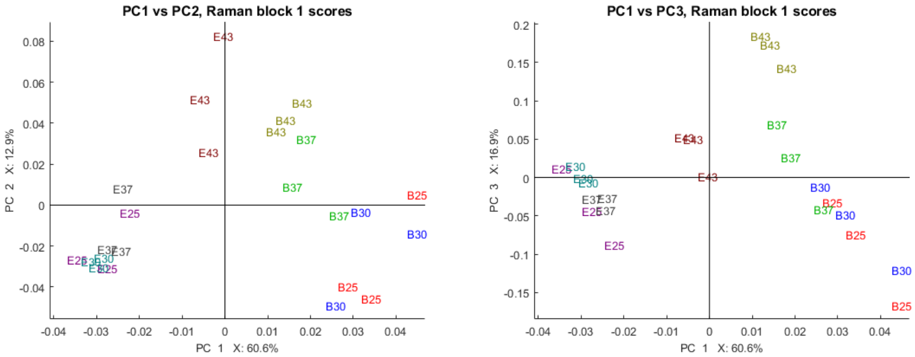


Figure 7.17: Selection of block scores, analysis 8: Raman multiblock set 2, block 1.

analyses 7 and 8 in discriminating between the eight groups. For this block, no variables were selected for PC1, which accounted for 20.8% of the total variance, the highest of any PC of any MALDI block in analyses 7 and 8. Consequently, subplots started from PC2 and upwards. For corresponding global scores, the pattern of grouping was relatively similar, although the planar orientation was somewhat different.

7.7.5 Correlation loading plots

Looking at the design matrices of block correlation loading plots, we see that only a few groups were predicted chiefly by a single latent variable, and that most groups clustered together relatively near the origin. This clustering obscured any otherwise visible connection between the groups of the design matrix, and variables of the spectral subregions. Luckily, there were exceptions: E43 and B43 showed strong correlation in PC2 and PC3, respectively, as best seen in the design matrix correlation loading plot of Figure 7.26.

These two groups were also the most frequently separated groups in the aforementioned block score plots. These two observations combined motivated a search to see what types of biomolecules showed the highest correlation with these groups.

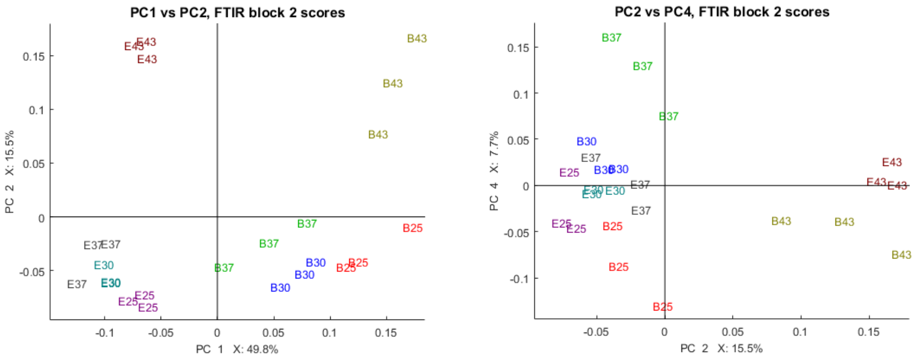


Figure 7.18: Selection of block scores, analysis 8: FTIR multiblock set 2, block 2.

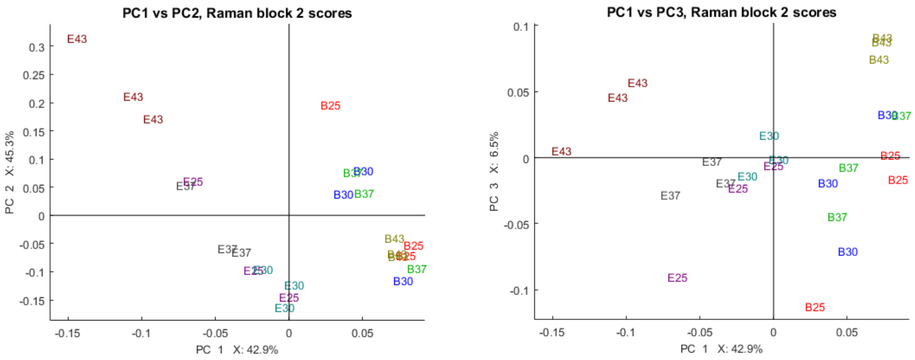


Figure 7.19: Selection of block scores, analysis 8: Raman multiblock set 2, block 2.

The E43 group achieved its highest correlation in the amide and fingerprint regions of the FTIR and Raman spectra, shown in Figure 7.27.

The B43 group achieved its highest correlation in the amide and fingerprint regions of the FTIR spectrum, shown in Figure 7.28; the results were not repeated by the corresponding Raman regions, nor by any of the MALDI plots.

From Figures 7.27 and 7.28, we see clear evidence of collinearity within the blocks. This trend was, as expected, discontinued for the MALDI block correlation loading plots. As a result, the variables grouped together in the general direction of the design matrix variables form the origin, but, in accordance with their explained variance per latent variable, rarely achieved high correlation. Examples are shown in Figure 7.29.

Another interesting finding was the relatively high correlation between design variables for groups E25 and B25, and variables associated with the lipid region, shown in Figure 7.30. This correlation was by not as pronounced as the one described above for E43 and B43, but still worth mentioning. This result was also indicated in the confusion matrix for analysis 8, shown in Figure 7.14, where two out of three samples of both B25 and E25 were correctly identified.

The remaining groups showed high correlation in several spectral subregions, which

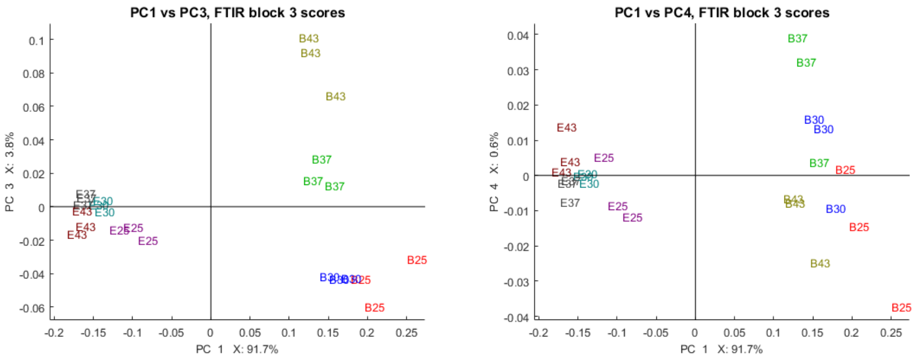


Figure 7.20: Selection of block scores, analysis 8: FTIR multiblock set 2, block 3.

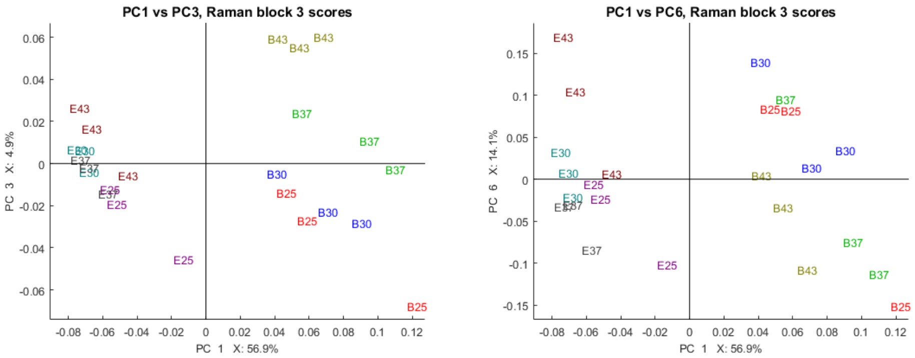


Figure 7.21: Selection of block scores, analysis 8: Raman multiblock set 2, block 3.

made it hard to ascribe their identification to specific types of biomolecules. An example of this (not shown), is the group B37, which, in addition to being located close to other design variables, had pertaining groups that showed high correlation in both blocks 2, 4, 7 and 8, corresponding to the amide (FTIR), carbohydrate (Raman) and fingerprint (both) regions of FTIR and Raman.

7.8 Key results

The main finding in the previous results chapter was repeated. Data were interpreted differently by the SMBPLSR routines, owing to the differences in collinearity between samples. In this data set, we also see further discrepancy between the interpretation of MALDI data of low and medium masses. The global scores of the MALDI, medium mass, data set showed a higher between-sample variance than the low mass one, probably due to the higher amount of variables in the set and consequent increased spread of main peaks.

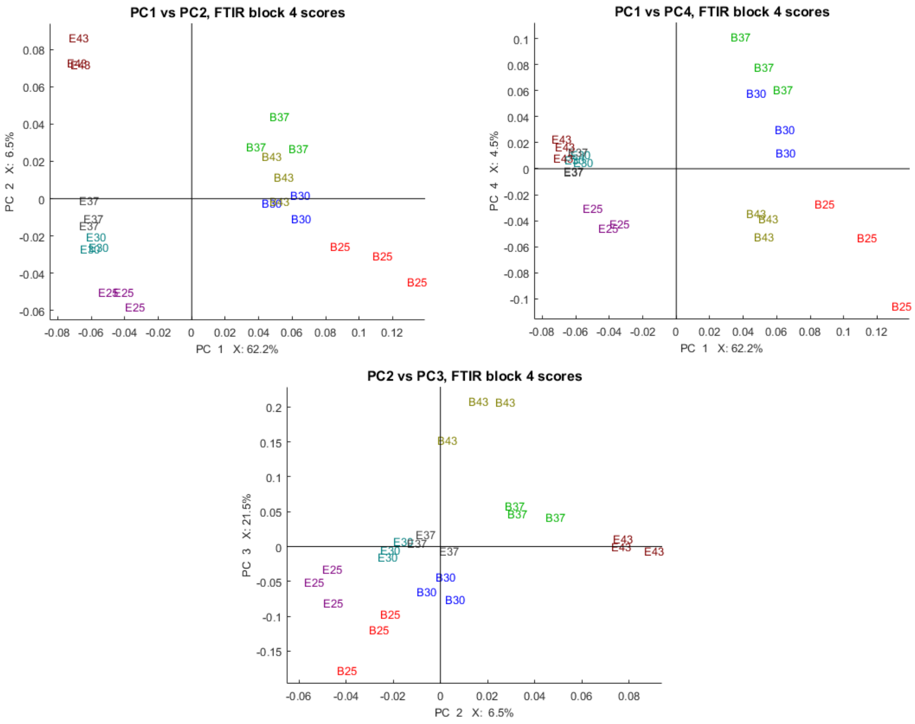


Figure 7.22: Selection of block scores, analysis 8: FTIR multiblock set 2, block 4.

7.8.1 Discrimination

As the summary of results below indicates, the SMBPLSR method showed good all-round discriminative power for this data set, despite the presumably small between-strain variance.

The SMBPLSR method was also able to identify specific temperatures where discrimination between strains of the same species was weak or absent. Inspecting the various score plots, we see that for *Escherichia*, E30 and E37 were seldom far apart, and most times at least partly intermingled. In Figure 7.22, we also see them separated from E25, a frequent companion of theirs.

The *Bacillus* samples showed a similar pattern, but here the B25 and B30 strains were the ones most often clustered, for example on Figures 7.20, 7.23 and in the global scores in Figure 7.13. Both the latter figure and several others bring nuance into this pattern by showing its dependence on certain latent variables. PC4 in particular seemed to be good at separating the two groups.

B30 was consistently well grouped and relatively separated in MALDI, low mass, global scores. This trend was also to some degree present in MALDI, medium mass, global scores (neither plots are shown).

E30 and E37 were separated by PC1 in Raman multiblock set 2 global scores (not shown). This separation was not managed as well elsewhere; the greatest separation

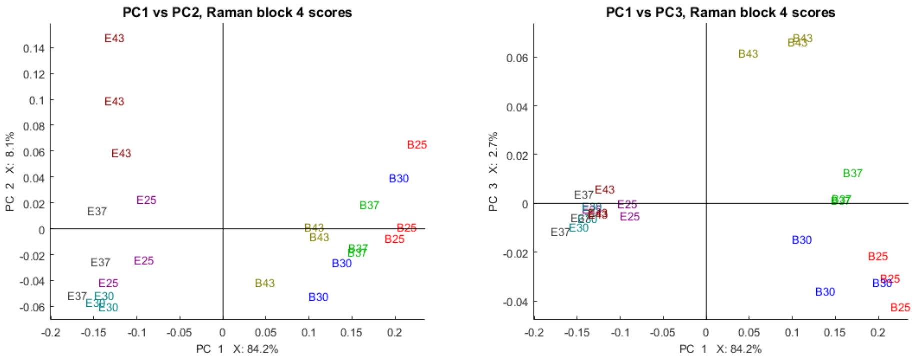


Figure 7.23: Selection of block scores, analysis 8: Raman multiblock set 2, block 4.

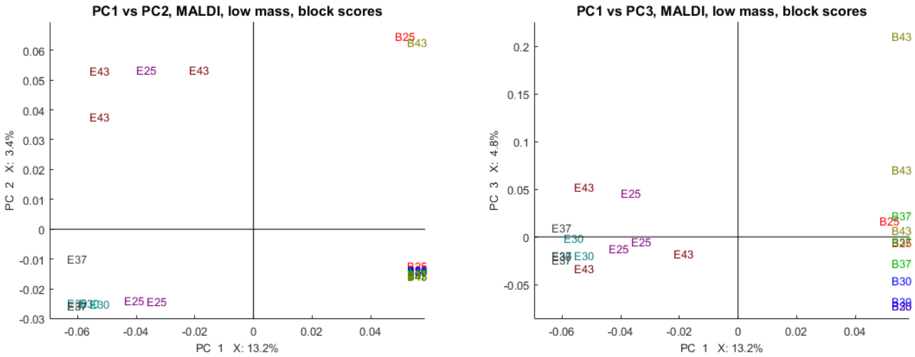


Figure 7.24: Selection of block scores, analysis 8: MALDI, low mass.

of these two strains was in the amide region of Raman multiblock set 2, Figure 7.19, suggesting that a difference in protein composition constituted a separable feature of the strains in question.

In analysis 8, the most easily discriminated groups in the block correlation loading plots were B25, E25, B43 and E43. B43 and E43 showed the highest correlation for variables in the amide and fingerprint regions; B43 only for FTIR, E43 for both FTIR and Raman. For B25 and E25, the lipid regions of both FTIR and MALDI provided the best correlation. Most plots of MALDI, low and medium mass, showed good directional correlation from the origin. The lack of a dominant latent variable muddled high correlation for any variable; correlation above 0.5 was seldom seen.

7.8.2 Success rates

A comparison of success rates between preliminary SPLSR analyses and analyses 5-8 showed a higher success rate for the latter analyses three out of four times. For SPLSR on FTIR (whole spectrum), the success rate was 62.5%, versus 70.8% in analysis 5. Normal PLSR in FTIR data (whole spectrum) also resulted in a success rate of 70.8%. SPLSR on the entire Raman spectrum yielded a success rate of 54.2%, the same as analysis 6. The

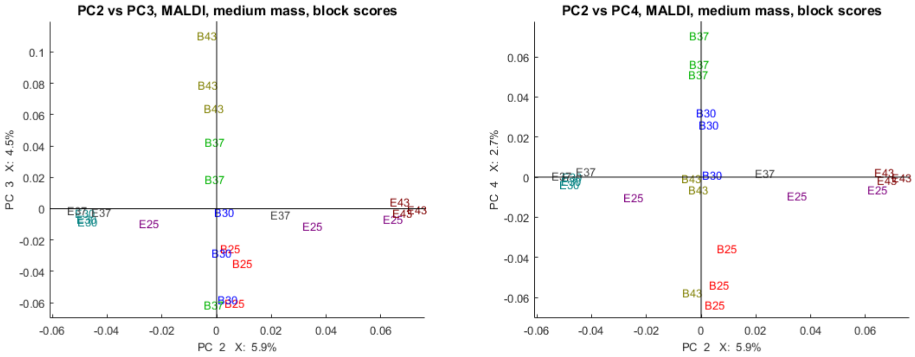


Figure 7.25: Selection of block scores, analysis 8: MALDI, medium mass.

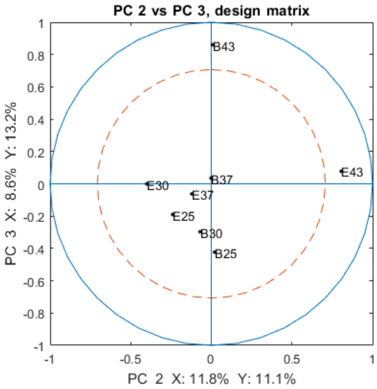


Figure 7.26: Design matrix correlation loading plot for PC2 vs PC3.

success rate for PLSR on the entire Raman spectrum was slightly higher at 62.5%.

The trend with low success rate was, as already mentioned, particularly evident in the MALDI data types when run as single descriptor data types. Dieckmann et al. [2] discussed the relatively poor intra-strain discriminatory power of MALDI as compared to Raman and especially FTIR, naming low degree of variation in ribosomal protein expression as its chief cause. It appears that the variation is no more pronounced as a function of growth temperature, either. The change was bigger in the metabolic fingerprint.

This assumption was backed also by visual inspection of confusion matrices from each single-type analysis (figures not shown). Misclassification across species only occurred in the Raman analysis, for four samples of *B. subtilis* grown at 25 (one sample), 30 (one sample) and 43 (two samples) wrongly being identified as *E. coli.* All other analyses ascribed all strains to the correct species. Given that the primary aim of this chapter was to assess how well SMBPLSR methods highlight phenotypic response to a set of growth conditions, these results are by and large satisfactory.

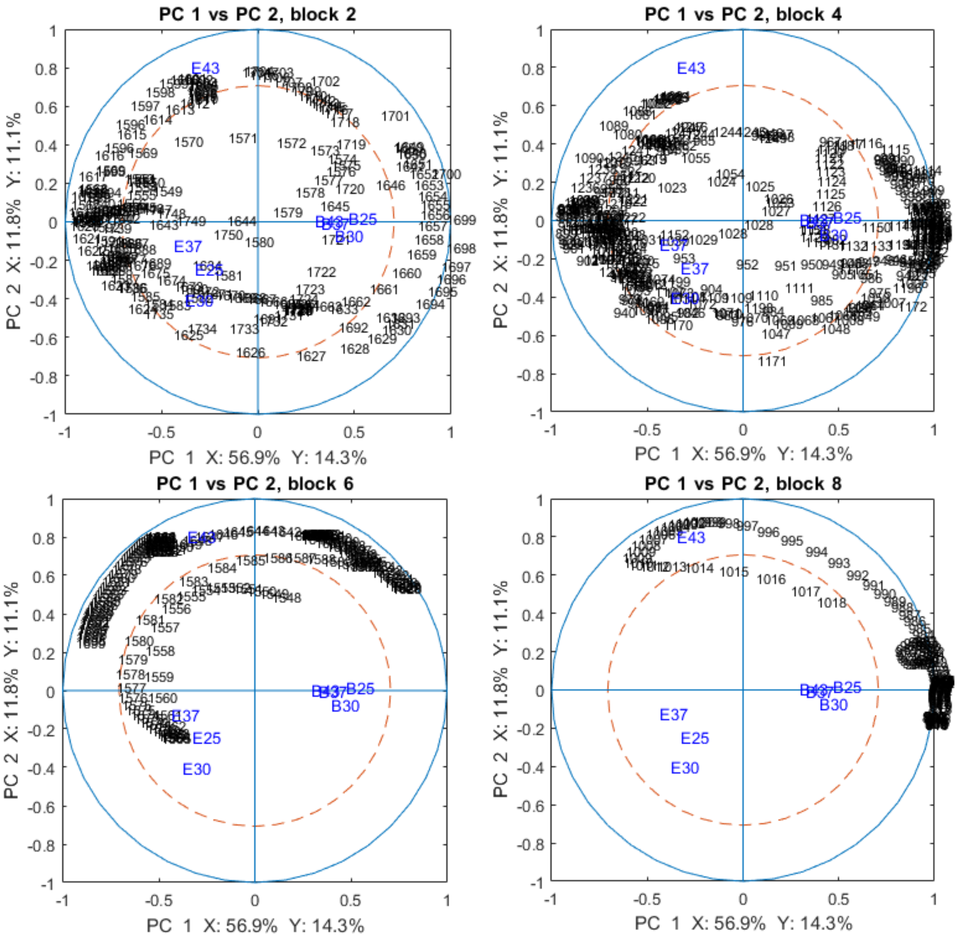


Figure 7.27: Block correlation loading plots of the biomolecular subregions and latent variables displaying the best correlation with the E43 group. FTIR amide block to the left, FTIR lower fingerprint block to the right. FTIR amide (left) and lower fingerprint region (right) on top, corresponding Raman blocks bottom. Design variables enlarged and in blue.

7.8.3 Optimal number of latent variables

The optimal number of latent variables for analysis 8 was 6; 4 for analysis 7. In this study, there were, with one exception in Figure 7.21, no plots reported where latent variables higher than the 4^{th} were shown. Further inspection (not shown) revealed that almost no additional information was extractable from the higher-order latent variables. We did, however, find discrimination of E25 in the 5^{th}. This information was recorded from FTIR multiblock set 2, blocks 1, 2 and 4, and Raman multiblock set 2, block 1, plus both MALDI sets.

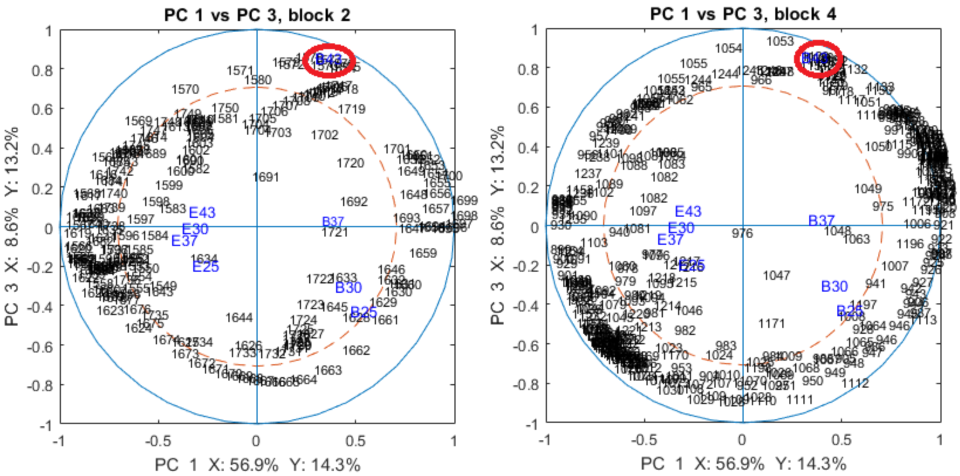


Figure 7.28: Block correlation loading plots of the biomolecular subregions and latent variables displaying the best correlation with the B43 group (encircled in red). FTIR amide block left and lower fingerprint region block right. Design variables enlarged and in blue.

7.8.4 Comparison of global score plots

A comparison between the global scores of analyses 5, 6, the MALDI analyses, and analysis 7 showed that the global scores of analysis 5 (FTIR multiblock set2, plot not shown) resembled those of analysis 7 the most, followed by analysis 6 (not shown) and the MALDI analyses (not shown). This resemblance consisted of equality in predictive ability and subplot-wise strain grouping. As previously discussed, analysis 5 also showed the highest success rate and most pronounced grouping of analyses 5, 6 and MALDI. In a similar comparison between the global scores of analyses 5, 6, MALDI, and analysis 8, grouping and clustering in the global scores was best resembled by analysis 5, and for the same reasons as in the above paragraph.

7.9 Discussion

For all analyses but those of MALDI, low and medium masses, the E43 group was completely correctly identified, though in combination with additional misclassifications attributed to it. In all but Raman and MALDI, both low and medium mass, analyses, B43 was also correctly identified, and always without any additional misclassifications pertaining to it. This suggests that discrimination was easiest and metabolic fingerprint most pronounced at high temperatures. The high scores of the block correlation loading plots reported for E43 and B43 (Figures 7.27 and 7.28) support this suggestion.

One reason for this finding might be that both species are found in mammalian intestines [34, 35], whose core temperatures normally do not exceed 40°C. Core temperatures around

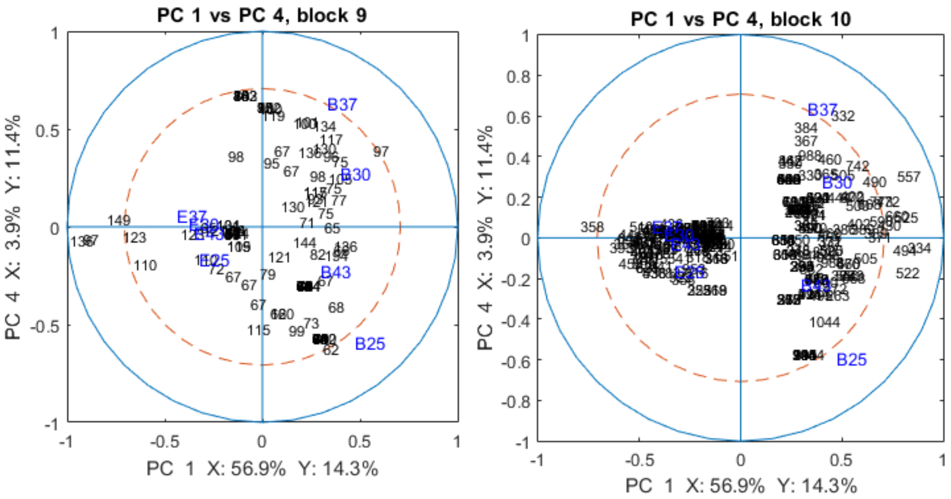


Figure 7.29: Block correlation loading plots of the MALDI block. Low mass to the left, medium mass to the right. Design variables enlarged and in blue.

43^{o}C in humans are lethal due to denaturation of the protein structure in bodily tissue [80]. A similar mechanism might explain the both the change in metabolic fingerprint in FTIR/Raman and in protein expression in MALDI. The global score plots of MALDI, low and medium masses (not shown), strengthen this conclusion by clearly separating E43 (low mass) and B43 (medium mass) in the 4th latent variable. This separation did not occur in corresponding block scores in analyses 7 and 8.

A biological explanation of why the groups B25 and E25 showed high correlation with lipid variables (also mentioned as a digression in Section 7.8.3 above) is the reported temperature dependence of bacterial fatty acid composition [81]. Unsaturated fatty acids have a lower melting point than saturated fatty acids, and as such one would expect an altered ratio in favour of unsaturated fatty acids as the growth temperature declines.

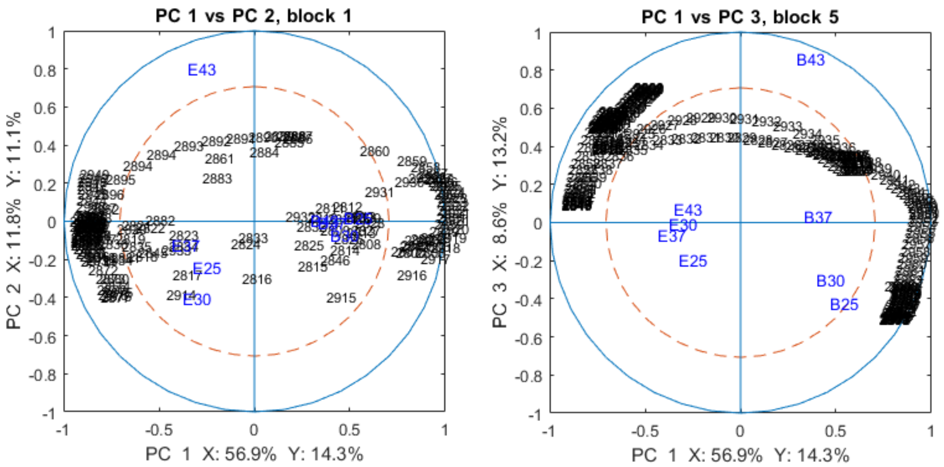


Figure 7.30: Block correlation loading plots of the biomolecular subregions and latent variables displaying the best correlation with the B25 and E25 groups. FTIR lipid block to the left, Raman lipid block to the right. Design variables enlarged and in blue.

Chapter 8

General discussion

8.1 On the different treatment of data types by the SMBPLSR routine

As described in the key results sections of both previous chapters, the SMBPLSR routine treated MALDI data vastly different from FTIR/Raman data. The discrepancy owes its existence to differences in inherent variance within each variable. Consequently, the same set of latent variables produce different grouping among strains and samples between the data blocks. This difference is evident in the block score plots of analyses 3, 4, 7 and 8.

Presumed lower metabolic differences between strains and consequently stronger likelihood for collinearity resulted in more prominent latent variables in the MALDI blocks of the *E. coli/B. subtilis* than in the MALDI block of the *Klebsiella* data set, as seen in Table 8.1 below. However, this did not alter the tendency of the SMBPLSR routines to yield different output for MALDI data than for FTIR/Raman between the two sets of analyses (3 and 4; 7 and 8); the same groups were still explained by different latent variables depending on data type.

Table 8.1: Explained variance by the most prominent latent variable in MALDI blocks in multi-data type analyses. Low and medium denote low and medium masses, respectively.

Analysis #	# Samples in	% of block variance	Latent variable #
	data set	explained	variable #
3	51	5.9	PC1
4	51	6.3	PC1
7	24	19.4 (low)/21.4 (medium)	PC4/PC1
8	24	15.4 (low)/20.8 (medium)	PC4/PC1

Loss of interpretability is a possible undesired side effect of this difference in treatment. Consider a case where the same information is stored in a higher-order latent variable in the FTIR/Raman data, and a lower-order one in MALDI. Because patterns stemming from the FTIR blocks are dominant for lower-order global latent variables and vice versa, additional information carried by lower-order MALDI and higher-order FTIR/Raman latent variables might be suppressed in a multi-type data setting such as analyses 3 and 4, as compared to in single-type settings such as analyses 1, 2 and MALDI set 1. Conversely, the data carried by the presently dominant latent variables are expressed twice. Other reports employing MB methods, sparse or not, selected data types expressing similar collinearity patterns [19, 24, 29].

On the contrary, this tendency can be used advantageously by attempting to establish patterns between the information carried in the latent variables of each data type for future analyses. If successful, these patterns may serve as a source for cross-referencing and quality control between the measurements. That is, for a certain phenotypic expression stored in a specific set of latent variables in FTIR data, one would expect to see similar information carried by a specific, but different set of latent variables in MALDI. Examining an eventual discrepancy between the two may also enhance total understanding of the phenotypic response exhibited, or increase indicative ability. If these attempts are unsuccessful or prove irrelevant to analysis, two measures for increasing collinearity and diminishing columnar variance in MALDI data are suggested below.

Autoscaling, or columnar division by inherent standard deviation [58] in data pretreatment, is one way to perhaps lower the variance within each column and bring it to levels comparable to that of FTIR and Raman. To our knowledge, this has never been done before for MALDI data, but instead with other types of spectral data [29] with high heteroskedasticity. Van den Berg et al. [58] describe several types of data scaling that can be used prior to statistical analysis to maybe tune the MALDI data into a form more visually compatible with FTIR and Raman. Van den Berg et al. [58] also mention inflated signal-to-noise ratio as a drawback with several of the scaling methods; this will not be a problem with bar code plots as all noise has been previously cleared.

While the aforementioned approach of might attune MALDI's appearance to FTIR/Raman, it does not eliminate its problematic lack of collinearity, whose success SMBPLSR methods depend upon [19, 29].

8.2 On utilising varied growth media to enhance discrimination

In the introduction, the proposed [2] advantage of MALDI over FTIR/Raman with regards to the creation of spectral data bases was introduced and commented. Considering the results presented in this thesis, this position should be qualified somewhat.

MALDI results are less susceptible to perturbation by varied growth conditions [11]. Stated in another way: Ribosomal proteins are less sensitive to varied growth conditions as measured by MALDI.

The high sensitivity to growth conditions expressed by both FTIR [2, 25] and Raman [2] metabolic fingerprints makes it possible to examine how minute variations in for instance access to nutrients, can trigger different types of phenotypic response. Knowledge of strain-specific phenotypic response to varied growth conditions can enhance the discriminative potential of analysis.

This potential was briefly addressed in this thesis. We showed how SMBPLSR methods could be utilised to highlight phenotypic response and attribute it to specific biomolecules. Several publications mention overlapping intra-species results as a shortcoming of FTIR methods in characterisation or identification [11, 55]. Including phenotypic response into the analysis of microorganisms at this taxonomic level may provide a higher-resolution picture of their overlap. Consequently, more detailed discrimination between groups can be achieved.

Extensive growth protocols that include broader sets of growth conditions may be more time consuming and costly [5]. However, if recent developments in sample treatment enabling high-throughput FTIR analysis of microbial samples [4, 9] are consolidated with our results to create more robust and detailed models for characterisation and identification, differentiated cultivation might become a desirable option to singular cultivation protocols in the future.

8.3 On the trade-off between predictive ability and interpretability

Throughout the analyses described in this thesis, the predictive ability of single data type analyses (such as 1, 2, 5 and 6) was generally equal to or higher than those of multi-type analyses (3, 4, 7 and 8). For example, analysis 1 (FTIR multiblock set 1) achieved a success rate of 96.1% and analysis 2 (Raman multiblock set 1) achieved 84.3%, equal to analysis 3 (FTIR set 1, Raman set 1, MALDI). The MALDI analysis achieved a success rate of 82.4%.

Achieving a high success rate was not the primary scope in the second set of analyses (5-8); rather examining whether SMBPLSR methods enabled assessment of phenotypic response to varying growth conditions. Nevertheless, the success rates of the set of analyses are mentioned here for comparison.

The success rate was 70.8% for analysis 5 (FTIR multiblock set 2), 54.2% for analysis 6 (Raman multiblock set 2), 54.2% for analysis 7 (FTIR set 2, Raman set 2, MALDI low and medium masses) and 50.0% for analysis 8 (FTIR multiblock set 2, Raman multiblock set 2, MALDI, low and medium masses). The analyses of MALDI, low and medium masses constitute exceptions in this regard, achieving success rates of only 29.2% and 25.0%, respectively. Correspondingly, the global scores of the single type data analyses showed clearer strain grouping than multi-type analyses.

Furthermore, the single data type analyses carried information about how each subregion of the metabolic fingerprint allowed for strain identification (analyses 1, 2) or recorded changes in phenotypic response (analyses 5, 6) for each individual data type. In Section 7.8 above, we reported that success rate, i.e. predictive ability, was comparable for SPLSR and SMBPLSR, given the same input data. In summary, this indicates that SMBPLSR analyses of single data types carry more information in the block loadings than corresponding SPLSR analyses, while not losing significant predictive ability.

The single block analyses achieved higher success rate and grouping, but did not carry information about the relative weighting of each data type. This type of information is, however, easily acquired by analysing multi-type data, such as 3 and 7. The block scores from these described the weighting of each data type for the different latent variables. Examples can be seen in block score plots of higher-order latent variables, where, as MALDI data start dominating the explained variance, patterns similar to MALDI block scores are emerging in the global scores. See the global scores of analysis 3 in Figure 6.9 and corresponding MALDI block scores in Figure 6.12 for an example.

The paragraphs above first describe how single data type analyses (1, 2, 5, 6 etc.) carry information about the weighting of each type of biomolecule in prediction. Then, we saw how analyses 3 and 7 might furnish the interpreter with knowledge of how data types are weighted in a multi-type data analysis. The information provided by analyses 4 and 8 encompasses both these aspects of interpretation: Weighting of data types *and* weighting of biomolecules. SMBPLSR methods possess the capacity to include several data types into the same statistical model, and to highlight and make accessible all the aforementioned information in a meaningful way.

8.4 On the discriminative ability of spectral subregions

In the first set of analyses described in Chapter 6, we see from the block scores of analysis 4 (Figures 6.16 to 6.20) that the single most powerful spectral subregion of all in terms of discriminative ability is the FTIR fingerprint area corresponding to block 4 of FTIR multiblock set 1, shown in Figure 6.19. This result, however, was not found in analysis 1, because the sparsity parameter had set the threshold in such a way that for block 4, block variables were only selected from PCs 3-5, 7 and 9, failing to select variables from the dominant PC1 that explained 40.3% of the total block variance and thus impeding the interpretability of the spectral data.

In the second set of analyses described in Chapter 7, the findings described in the above paragraph were repeated. Again, we saw how the FTIR fingerprint area corresponding to FTIR multiblock set 2, block 4 (Figure 7.22), surpassed all other spectral subregions in discriminative ability in analysis 8. This trend was partly followed in the correlation loading plots, where the variables belonging to fingerprint regions showed high correlation for B43 and E43, although both findings were supported by similar correlation in the amide regions.

According to the score plots, phenotypic response to growth temperature was most pronounced in the fingerprint region. The findings from analysis 1 were also repeated in analysis 5 on FTIR multiblock set 2 alone: Variables were not recorded for several of the dominant latent block variables. The most striking examples were again from block 4, where no variables were recorded for PC1 (67.4% explained block variance) and PC2 (12.7%).

These findings strengthen the case for using multi-type input data in SMBPLSR models for a larger set of variables to deflate upon. For instance, regions of maintained strong signal, such as the FTIR fingerprint region, will be selected at the cost of larger, but less explaining MALDI regions. Another alternative is to impose less strict sparsity requirements in single type analyses. Lê Cao et al. [26] discuss this problem in particular, stating that too strict variable selection can potentially remove subtle but valuable patterns in the data set.

8.5 On different deflation methods

In PCA, the latent variable explaining most of the variation in **X** is always calculated first [57]. The subsequent PCs are then calculated in descending order according to the same quality. This is also the case with PLSR, but in this case, it is the sum of the variance explained for **X** *and* **Y** that determines the position of the PC [59]. Numerous examples of this are listed in Tables B.1 and B.2 in Appendix B. This trend was even more prominent in the score plots, and clearly visible from inspecting the block score plots shown above. An example is shown in Figure 7.19, where, in block 2 of Raman multiblock set 2 from analysis 8, PC1 explained 42.9% and PC2 45.3% of the total block variance.

A related and interesting consequence of deflation using super scores, which was done in these models, was that the dimension chosen for deflation did not necessarily contain strong enough signals in all the blocks; the soft threshold λ was set so high globally that for one or more blocks, no signal was strong enough to exceed it. Consequently, no variables were selected from this block, and there was no way to plot the associated score plots; information had to be obtained by other means, for example correlation loading plots. This weakness was also discussed by Westerhuis and Smilde [27]. Still, as mentioned by Karaman et al. [29], in sparse models, there is essentially no valid alternative to deflation of both **X** and **Y** on super scores, because deflating on block scores would remove information that is not presented; slight mixing of information is preferred to its removal.

8.6 On model validation

While the analytic capability of the SMBPLSR method has enjoyed much attention and devotion in this thesis, less effort has been dedicated to address model stability. This study employs the SMBPLSR routine as it was developed by Karaman et al. [29]. Unlike them, however, we did not address the stability of the model using CMV, only CV. Westad et al. [75] highlighted how CMV could be used deliberately to establish stable and robust connections between the predicted model and specific subregions of the FTIR and Raman spectrum. This validation would have been a particularly useful part of the analyses described in Chapter 7 in this thesis. The reason these methods were not included, was because a CMV function for multiblock models was absent in the Saisir code pack, and that the deadline for the thesis provided the author with insufficient time to assemble one such function. Notwithstanding, other methods of stability-assessment could have been undertaken.

Hassani et al. [24] suggested calculating the Root Mean Square Error (RMSE) to validate the stability of their MBPLSR routine. RMSE is calculated for each the descriptor data block $\mathbf{X}^b$, and the response data $\mathbf{Y}$. These are calculated to highlight the contribution of each block and for each component in the prediction, and to highlight the predictive ability of each descriptor block in the response data, respectively. RMSE is calculated by CV. For mathematical details, see [24].

Another method for addressing model stability is the Q^2 test. This test was used by Karaman et al. [42] to assess predictive ability of each model as part of a CMV routine. This test, however, was found by Szymanska et al. [22] to yield inferior results to the routine of calculating the Number of Misclassifications (NMC) used in this study; according to them, NMC proved better at predicting the model when the difference between groups was small, which is clearly the case in this study for both data sets.

Chapter 9

Conclusions and outlook

In this thesis, a twofold scope has been addressed. First, the ability of sparse multiblock methods to establish a connection between metabolic fingerprints by FTIR and Raman, and protein expression by MALDI in *K. oxytoca* and *K. pneumoniae* has been evaluated. Score plots were used to showcase the methods' ability to accurately characterise different strains in the data set, to ascribe phenotypic similarities and differences between strains to data blocks pertaining to specific regions of the FTIR and Raman spectra, and compare these with protein expression in MALDI.

Second, the methods were used to compare phylogenetic differences and similarities in strains of *E. coli* and *B. subtilis* assessed by the same three phenotyping techniques. Results acquired from multi-type data analyses made it possible to establish connections between growth temperature and phenotypic response, showcasing which types of biomolecules were most involved in identification for different temperatures. Block score plots were used to show that the most pronounced responses were in the amide region for high temperature (43^{o}C), and in the lipid region for low temperature (25^{o}C). These results indicate SMBPLSR methods' ability to utilise differentiated growth protocols to produce more robust and comprehensive models.

For both data sets, there were clear indications that the methods' treatment of MALDI data was completely different from that of FTIR and Raman. While grouping patterns in FTIR and Raman were relatively equivalent, a completely different set of latent variables were responsible for corresponding grouping in MALDI. This discrepancy was due to the lack of collinearity in the MALDI data. As a consequence, global patterns were dominated by the highly collinear FTIR/Raman data in the first 2-3 latent variables, but the dominance shifted towards patterns explained chiefly by the MALDI blocks in higher-order latent variables.

The ability to establish patterns of grouping between FTIR/Raman and MALDI is a possible advantageous consequence of this; a drawback is the risk of the same information being doubly present in dominant latent variables - first in FTIR/Raman - then in MALDI, while other information is suppressed.

The connections described in this thesis are based on few data sets, and for a limited set of phenotypic response. More extensive testing is required to strengthen the conclusions. The robustness of the method should also be further examined by application on strains subject to other variations in growth conditions, such as nutrition deficiency [6] or cultivation time [11].

Appendices

Appendix A

Correspondence tables for strain names

Table A.1: Overview of *K. oxytoca* and *K. pneumoniae* strain names, the names with which they appear in the figures in this thesis, and their corresponding full script names. r = 0, 1 or 2 depending on biological replicate.

Strain name	Name appearing in figures	Full name in script
PHS-890^a	OxyHWP	OxyHWP8000r
PHS-891^a	OxyHWP	OxyHWP8010r
PHS-892^a	OxyHWP	OxyHWP8020r
PHS-893^a	OxyHWP	OxyHWP8030r
PHS-894^a	OxyHWP	OxyHWP8040r
PHS-895^a	OxyHWP	OxyHWP8050r
PHS-896^a	OxyHWP	OxyHWP8060r
PHS-897^a	OxyHWP	OxyHWP8070r
PHS-898^a	OxyHWP	OxyHWP8080r
PHS-899^a	OxyHWP	OxyHWP8090r
CB4063^a	OxyCB4	OxyCB40000r
CB4074^a	OxyCB4	OxyCB40010r
CB4072^a	OxyCB4	OxyCB40020r
CCUG 15788^a	OxyCCU	OxyCCUG000r
Oman 61^a	OxyOma	OxyOman000r
ATCC 13182^a	OxyAT1	OxyAT13000r
ATCC 25926^b	PneOza	PneOzae000r

^{a} *K. oxytoca*, ^{b} *K. pneumonia*

Table A.2: Overview of *B. subtilis* and *E. coli* strain names and, the names with which they appear in the figures in this thesis, and their corresponding script names. r = 0, 1 or 2 depending on biological replicate.

Strain name	Growth temperature	Name appearing	Full name
	[^{o}C]	in figures	in script
DSM 347[a]	25	B25	BacSubDSM0347250r
DSM 347[a]	30	B30	BacSubDSM0347300r
DSM 347[a]	37	B37	BacSubDSM0347370r
DSM 347[a]	43	B43	BacSubDSM0347430r
K12 DSM 3871[b]	25	E25	EscColDSM3871250r
K12 DSM 3871[b]	30	E30	EscColDSM3871300r
K12 DSM 3871[b]	37	E37	EscColDSM3871370r
K12 DSM 3871[b]	43	E43	EscColDSM3871430r

[a] *B. subtilis*, [b] *E. coli*

Appendix B

Globally explained variance

Table B.1: Explained global variance for the 10 PCs calculated as part of the *K.oxytoca/K. pneumoniae* analyses a: 3 and b: 4. Optimal number of PCs were 7 for analysis 3 and 5 for analysis 4.

PC#	$\mathbf{X}_a$ [%]	$\mathbf{Y}_a$ [%]	$\mathbf{X}_b$ [%]	$\mathbf{Y}_b$ [%]
1	27.2	38.4	27.9	44.3
2	15.2	7.4	15.3	7.6
3	9.1	8.3	6.8	6.4
4	4.3	12.1	14.9	5.0
5	3.5	11.0	4.0	6.5
6	5.5	4.9	6.5	3.9
7	3.5	3.9	3.9	4.5
8	3.2	4.7	2.5	4.1
9	2.4	2.2	2.6	2.1
10	2.7	1.5	1.8	3.8
Total:	**76.6**	**94.4**	**86.2**	**88.2**

Table B.2: Explained global variance for the 10 PCs calculated as part of the *E. coli*/*B. subtilis* SMBPLSR analyses a: 7, and b: 8. Optimal number of PCs was 4 for analysis 7 and 6 for analysis 8.

PC#	$\mathbf{X}_a$ [%]	$\mathbf{Y}_a$ [%]	$\mathbf{X}_b$ [%]	$\mathbf{Y}_b$ [%]
1	35.3	14.2	56.9	14.3
2	12.1	10.5	11.8	11.1
3	8.6	12.3	8.5	13.2
4	7.8	8.8	3.9	11.4
5	3.9	11.7	1.7	12.0
6	2.9	12.2	3.5	8.3
7	3.3	7.1	1.9	9.2
8	1.9	9.6	0.8	10.6
9	3.1	3.7	1.5	3.4
10	2.2	3.6	0.8	1.9
Total:	**81.1**	**93.7**	**91.3**	**95.4**

Appendix C

MATLAB scripts

In this chapter, the two main scripts used in this thesis are presented. Code and functions referred to in the script may be shared upon request. Currently, both scripts are set up for analysing data corresponding to runs 1 and 5 (FTIR multiblock sets 1 and 2, respectively).

C.1 Identification

The following script is the one used to analyse the data on *K. oxytoca* and *K. pneumoniae* described in Chapter 2.1.

```
% Master thesis script tailored for analysis of the data
% set of Klebsiella oxytoca and K. pneumoniae, provided by
% Peter Lasch of Robert Koch-Institut, Berlin, Germany.
%
% Original script setup by Valeria Tafintseva, Norwegian
  University of
% Life Sciences. Tailoring and extensions set up by Tor Einar
  M ller.

clear all;
close all;

MALDI = 0; % If MALDI data are included in the analysis
CMV = 0; % If CMV is to be performed

tic

%% Import data. Dataset from Robert Koch Inst./Peter Lasch
```

```
Akos=genpath('C:\Users\Bruker\Desktop\Master\2TorEinar\');
addpath(Akos,'C:\Users\Bruker\Desktop\Master\scripts\saisir');
addpath(Akos,'C:\Users\Bruker\Desktop\Master\scripts');
DirNameData='C:\Users\Bruker\Desktop\Master\DataRobertKoch\'; %
    Set paths

[ZSaisir1]=LoadFromUnscrambler_v1003(DirNameData,'KlebsiellaFTIR
    ');
ZX = ZSaisir1; % FTIR

ZX_raw = load(strcat(DirNameData,'KlebsiellaRaman'));
ZSaisir.v = ZX_raw.VarLabels0;
ZSaisir.d = ZX_raw.NIR_FTRaman;
ZSaisir.i = ZX_raw.ObjLabels;
RX = ZSaisir; % Raman

maldi_loc = '\MALDI spectra\Klebsiella-oxytoca.muf';
MX = load(strcat(DirNameData, maldi_loc),'-mat');
MX = MX.spec; % MALDI

%% Replace comma by point
Nx=size(ZX.v,1); % FTIR
for j=1:Nx
    oldName=ZSaisir1.v(j,:);
    modifiedStr = strrep(oldName,',','.');
    ZX.v(j,:)=modifiedStr;
end

Nx=size(RX.v,1); % Raman
for j=1:Nx
    oldName=ZSaisir.v(j,:);
    modifiedStr = strrep(oldName,',','.');
    RX.v(j,:)=modifiedStr;
end

%% Preprocess spectra by EMSC
ws = 9; % Set window size of SG filter
```

```
RX_2ndDer = saisir_derivative(RX,2,ws,0); % Raman
[RX_EMSCModel]=make_emsc_modfunc(RX_2ndDer);
[RX_cor,~,~]=cal_emsc(RX_2ndDer,RX_EMSCModel);

ZX_2ndDer = saisir_derivative(ZX,2,ws,2); % FTIR
[ZX_EMSCModel]=make_emsc_modfunc(ZX_2ndDer);
[ZX_cor,~,~]=cal_emsc(ZX_2ndDer,ZX_EMSCModel);

%% Select blocks of FTIR and Raman data
% Select FTIR blocks
ZX_cor_sel{1} = selectcol(ZX_cor, 1090:1245); % 2950-2800 wn
ZX_cor_sel{2} = selectcol(ZX_cor, 2334:2542); % 1750-1550 wn
ZX_cor_sel{3} = selectcol(ZX_cor, 2645:2749); % 1450-1350 wn
ZX_cor_sel{4} = selectcol(ZX_cor, 2853:3216); % 1250-900 wn

% Concatenate FTIR blocks
ZX_cor_con = appendcol(ZX_cor_sel{1}, ZX_cor_sel{2});
ZX_cor_con = appendcol(ZX_cor_con, ZX_cor_sel{3});
ZX_cor_con = appendcol(ZX_cor_con, ZX_cor_sel{4});

% Select Raman blocks
RX_cor_sel{1} = selectcol(RX_cor, 468:728); % 3050-2800 wn
RX_cor_sel{2} = selectcol(RX_cor, 1868:2024); % 1700-1550 wn
RX_cor_sel{3} = selectcol(RX_cor, 2076:2283); % 1500-1300 wn
RX_cor_sel{4} = selectcol(RX_cor, 2387:2854); % 1200-750 wn
RX_cor_sel{5} = selectcol(RX_cor, 3269:3526); % 350-100 wn

% Concatenate Raman blocks
RX_cor_con = appendcol(RX_cor_sel{1}, RX_cor_sel{2});
RX_cor_con = appendcol(RX_cor_con, RX_cor_sel{3});
RX_cor_con = appendcol(RX_cor_con, RX_cor_sel{4});
RX_cor_con = appendcol(RX_cor_con, RX_cor_sel{5});

%% Set up MALDI data
% Set up names: Constructs and stores names identical to those
   used in
```

```
% FTIR in a new MX field called stn (stored names).
for row = 1:length(MX)
    species = strcat(upper(MX(row).spe(1)),MX(row).spe(2:3));
    if strcmp(MX(row).str(1:2), '13')
        strain = strcat('HWP80', MX(row).str(9));
    elseif strcmp(MX(row).str(1:6), 'ATCC 1')
        strain = 'AT1300';
    elseif strcmp(MX(row).str(1:6), 'ATCC 2')
        strain = 'Ozae00';
    elseif strcmp(MX(row).str(1:4), 'Oman')
        strain = 'Oman00';
    elseif strcmp(MX(row).str(1:4), 'CCUG')
        strain = 'CCUG00';
    elseif strcmp(MX(row).str(1:6), 'CB4063')
        strain = 'CB4000';
    elseif strcmp(MX(row).str(1:6), 'CB4072')
        strain = 'CB4001';
    elseif strcmp(MX(row).str(1:6), 'CB4074')
        strain = 'CB4002';
    else
        disp('Unknown strain.')
    end
    replicate = str2num(MX(row).typ(5:6))-1;
    rep = strcat('0', num2str(replicate));
    MX(row).stn = strcat(species, strain, rep); % Creates a new
       field in MX
end

if MALDI
    MX_bar = make_bar_code(MX); % Create MALDI bar code plot
    MX_bar = selectcol(MX_bar, 1222:9222); % Incise in spectrum
end


%% Concatenate data types
% One block for all data types
X_con_data = appendcol(ZX_cor_con, RX_cor_con);
```

```
if MALDI
    X_con_data = appendcol(X_con_data, MX_bar);
end

% Multiblock, one block for each data type
X_multi_data{1} = ZX_cor_con;
X_multi_data{2} = RX_cor_con;
if MALDI
    X_multi_data{3} = MX_bar;
end

% Multiblock, four FTIR blocks, five Raman blocks, and one MALDI
     block
XM{1} = ZX_cor_sel{1};
XM{2} = ZX_cor_sel{2};
XM{3} = ZX_cor_sel{3};
XM{4} = ZX_cor_sel{4};
XM{5} = RX_cor_sel{1};
XM{6} = RX_cor_sel{2};
XM{7} = RX_cor_sel{3};
XM{8} = RX_cor_sel{4};
XM{9} = RX_cor_sel{5};
if MALDI
     XM{10} = MX_con;
end

%% Define parameters
% Classification Tree specification
Levels.Names =  'Strain';
Levels.NameRange{1,1} = 1:6;
Levels.NumLevels = size(Levels.Names,1);

% Define CV type:  'Full CV','Contiguous Blocks','Random Subsets
    ',
% 'Venetian Blinds', 'Spefified'.
CVPar.CVType = 'Venetian Blinds';
CVPar.cvfold = 3;
CVPar.cvind = [1,11]; % Relevant for 'Specified' configuration.
CVPar = CVParGenerator(CVPar,ZX_cor_con);

% Define PLS method
PLSPar.PLSMethod = 'pls';
PLSPar.Sparse = 'on';
PLSPar.Multiblock = 'off';
PLSPar.PLSType = 'DA'; % Discriminant Analysis
PLSPar.SparsInterval = [0.90, 0.98]; % percent of variables
    thrown away
PLSPar.NSpars = 5; % 5 number of values for sparsity
PLSPar.p_critical = 0.05;      % Significance level for
    sensitivity in the
                               % model selection: measured in
                                   precent of
                               % optimal error.
PLSPar.pc = 10;                % Maximum number of PC's

%% MAIN PART
%% Training
N = size(ZX_cor_con,1);
level = 1;
which = (1:N)';
PlotTrain.ConfM = 1; % Plot confusion matrix
PlotTrain.RegrCoef = 1; % Plot regression coefficients

% Set up Y
[ZY,ZYAdd] = split2levels(Levels.NameRange,ZX_cor_con,'off');

% Perform PLS
[ClasModel] = EstimateClassTree(level,ZX_cor_con,ZY,ZYAdd,which
    ,...
                PLSPar,CVPar,Levels,PlotTrain);

%% CMV
if CMV
    Vote.switch = 'off';
    Vote.ind = 1:20;

    CMVPar.CVType = 'Full CV'; % Options as for CV described
       above
    CMVPar.cvfold = 9;
    CMVPar.cvind = 7:8;
    CMVPar = CVParGenerator(CMVPar,RX_cor);
    CMVPar.cvfold = CMVPar.cvfold;
    CMVPar.freplt = 1; % Create frequency plot
    CMVData = CMVDA(RX_cor,PLSPar,CVPar,CMVPar,Levels,Vote,'off'
       );
end

toc
```

C.2 Experimental design

The following script is the one used to analyse the data on *E. coli* and *B. subtilis* described in Chapter 2.2.

```
% Master thesis script tailored for analysis of the data
% set 'Study2008' of Escherichia coli and Bacillus subtilis,
   provided by
% Peter Lasch of Robert Koch-Institut, Berlin, Germany.
%
% Original script setup by Valeria Tafintseva, Norwegian
   University of
% Life Sciences. Tailoring and extensions set up by Tor Einar
   M ller.

clear all;
close all;
MALDI = 0; % If MALDI data are included in the analysis
CMV = 0; % If CMV is to be run

tic

%% Import data. Dataset from Robert Koch Inst./Peter Lasch
```

```
Akos=genpath('C:\Users\Bruker\Desktop\Master');
addpath(Akos,'C:\Users\Bruker\Desktop\Master\scripts');
DirNameData='C:\Users\Bruker\Desktop\Master\Data2008\'; % Set paths

[ZSaisir]=LoadFromUnscrambler_v1003(DirNameData,'FTIRData');
ZX = ZSaisir; % FTIR

[RSaisir]=LoadFromUnscrambler_v1003(DirNameData,'RamanData');
RX = RSaisir; % Raman

MX_l = load(strcat(DirNameData, 'low-mass'));
MX_l = MX_l.spec; % MALDI low mass

MX_m = load(strcat(DirNameData, 'medium-mass'));
MX_m = MX_m.spec; % MALDI low mass

%% Replace comma by point
Nx=size(ZX.v,1); % FTIR
for j=1:Nx
     oldName=ZSaisir.v(j,:);
     modifiedStr = strrep(oldName,',','.');
     ZX.v(j,:)=modifiedStr;
end

Nx=size(RX.v,1); % Raman
for j=1:Nx
     oldName=RSaisir.v(j,:);
     modifiedStr = strrep(oldName,',','.');
     RX.v(j,:)=modifiedStr;
end

%% Preprocess spectra
ws = 9; % Set window size of SG filter.

ZX_2ndDer = saisir_derivative(ZX,2,ws,2); % FTIR
[ZX_EMSCModel]=make_emsc_modfunc(ZX_2ndDer);
```

```
[ZX_cor,~,~]=cal_emsc(ZX_2ndDer,ZX_EMSCModel);

RX_2ndDer = saisir_derivative(RX,2,ws,0); % Raman
[RX_EMSCModel]=make_emsc_modfunc(RX_2ndDer);
[RX_cor,~,~]=cal_emsc(RX_2ndDer,RX_EMSCModel);

%% Select blocks of FTIR and Raman data
% Select FTIR blocks
ZX_cor_sel{1} = selectcol(ZX_cor, 1090:1245); % 2950-2800 wn
ZX_cor_sel{2} = selectcol(ZX_cor, 2334:2490); % 1750-1600 wn
ZX_cor_sel{3} = selectcol(ZX_cor, 2594:2749); % 1500-1350 wn
ZX_cor_sel{4} = selectcol(ZX_cor, 2853:3257); % 1250-850 wn

% Concatenate FTIR blocks
ZX_cor_con = appendcol(ZX_cor_sel{1}, ZX_cor_sel{2});
ZX_cor_con = appendcol(ZX_cor_con, ZX_cor_sel{3});
ZX_cor_con = appendcol(ZX_cor_con, ZX_cor_sel{4});

% Select Raman blocks
RX_cor_sel{1} = selectcol(RX_cor, 415:727); % 3100-2800 wn
RX_cor_sel{2} = selectcol(RX_cor, 1867:2230); % 1700-1350 wn
RX_cor_sel{3} = selectcol(RX_cor, 2583:2853); % 1010-750 wn
RX_cor_sel{4} = selectcol(RX_cor, 3267:3527); % 350-100 wn

% Concatenate Raman blocks
RX_cor_con = appendcol(RX_cor_sel{1}, RX_cor_sel{2});
RX_cor_con = appendcol(RX_cor_con, RX_cor_sel{3});
RX_cor_con = appendcol(RX_cor_con, RX_cor_sel{4});



%% Set up MALDI data
% Set up names: Constructs and stores names identical to those
   used in
% FTIR in a new MX field called stn (stored names).

for row = 1:length(MX_l)
    gen = MX_l(row).gen(1:3);
    spe = strcat(upper(MX_l(row).spe(1)), MX_l(row).spe(2:3));
    if strcmp(gen,'Bac')
        str = strcat(MX_l(row).str(1:3),'0',MX_l(row).str(5:7));
    elseif strcmp(gen,'Esc')
        str = strcat(MX_l(row).str(6:8),MX_l(row).str(10:13));
    else disp('unknown species!')
        break
    end
    temp = MX_l(row).tem(1:2);
    if rem(row, 3) == 1
        tech = '00';
    elseif rem(row, 3) == 2
        tech = '01';
    elseif rem(row, 3) == 0
        tech = '02';
    end
    MX_l(row).stn = strcat(gen, spe, str, temp, tech);
    MX_m(row).stn = strcat(gen, spe, str, temp, tech);
end

if MALDI
    MX_l_bar = make_bar_code(MX_l);
    MX_l_bar = selectcol(MX_l_bar,1:1983);
    MX_m_bar = make_bar_code(MX_m);
    MX_m_bar = selectcol(MX_m_bar,1:10216);
end

%% Concatenate data types
% One block containing all data
X_con_data = appendcol(ZX_cor_con, RX_cor_con);
if MALDI
    X_con_data = appendcol(X_con_data, MX_l_bar);
    X_con_data = appendcol(X_con_data, MX_m_bar);
end

% Multiblock, one block per data type
X_multi_data{1} = ZX_cor_con;
```

```
X_multi_data{2} = RX_cor_con;
if MALDI
    X_multi_data{3} = MX_l_bar;
    X_multi_data{4} = MX_m_bar;
end

% Multiblock, four FTIR-blocks, four Raman-blocks and two MALDI-
   blocks
XM{1} = ZX_cor_sel{1};
XM{2} = ZX_cor_sel{2};
XM{3} = ZX_cor_sel{3};
XM{4} = ZX_cor_sel{4};
XM{5} = RX_cor_sel{1};
XM{6} = RX_cor_sel{2};
XM{7} = RX_cor_sel{3};
XM{8} = RX_cor_sel{4};
if MALDI
    XM{9} = MX_l_bar;
    XM{10} = MX_m_bar;
end


%% Define parameters
% Classification Tree specification
Levels.Names = 'Design';
Levels.NameRange{1,1} = 10:15;
Levels.NumLevels = size(Levels.Names,1);

% Define CV type: 'Full CV','Contiguous Blocks','Random Subsets
   ',
% 'Venetian Blinds', 'Specified'.
CVPar.CVType = 'Venetian Blinds';
CVPar.cvfold = 3;
CVPar.cvind = [1,11]; % Relevant for 'Specified' configuration.
CVPar = CVParGenerator(CVPar,ZX_cor_con);

% Define PLS method
```

```
PLSPar.PLSMethod = 'pls';
PLSPar.Sparse = 'off';
PLSPar.Multiblock = 'on';
PLSPar.PLSType = 'DA'; % Discriminant Analysis
PLSPar.SparsInterval = [0.90, 0.99]; % percent of variables
   thrown away
PLSPar.NSpars = 3; % 5 number of values for sparsity
PLSPar.p_critical = 0.05;      % Significance level for
   sensitivity in the
                               % model selection: measured in
                                  precent of
                               % optimal error.
PLSPar.pc = 10;                % Maximum number of PC's

%% MAIN PART
%% Training
N = size(ZX_cor_con,1);
level = 1;
which = (1:N)';
PlotTrain.ConfM = 1; % Plot confusion matrix
PlotTrain.RegrCoef = 1; % Plot regression coefficients

% Set up Y
[ZY,ZYAdd] = split2levels(Levels.NameRange,ZX_cor_con,'off');

% Perform PLSR
[ClasModel] = EstimateClassTree(level,ZX_cor_sel,ZY,ZYAdd,which
   ,...
                 PLSPar,CVPar,Levels,PlotTrain);

%% CMV
if CMV;
    Vote.switch = 'off';
    Vote.ind = 1:20;

    CMVPar.CVType = 'Venetian Blinds'; % Options as for CV
    CMVPar.cvfold = 8;
    CMVPar.cvind = 7:8;
    CMVPar = CVParGenerator(CMVPar,ZX_cor);
    CMVPar.cvfold = CMVPar.cvfold;
    CMVPar.freplt = 1; % Plot frequency plot
    CMVData = CMVDA(ZX_cor,PLSPar,CVPar,CMVPar,Levels,Vote,'off'
        );
end

toc
```

Bibliography

[1] D. L. Hartzl and E. W. Jones, *Genetics. Analysis of genes and genomes.* Jones and Bartlett Publishers, Sudbury, MA, USA, 6^{th} ed., 2005.

[2] R. Dieckmann, J. A. Hammerl, H. Hahmann, A. Wicke, S. Kleta, P. W. Dabrowski, A. Nitchse, M. Stämmler, S. A. Dahouk, and P. Lasch, “Rapid characterisation of *Klebsiella oxytoca* isolates from contaminated liquid hand soap using mass spectrometry, FTIR and Raman spectroscopy,” *Faraday Discussions*, vol. 187, pp. 353–375, 2016.

[3] P. Nordmann, G. Cuzon, and T. Naas, “The real threat of *Klebsiella pneumoniae* carbapenemase-producing bacteria,” *The Lancet Infectious Diseases*, vol. 9, pp. 228–236, 2009.

[4] V. Shapaval, J. Schmitt, T. Møretrø, H. P. Suso, I. Skaar, A. W. Åsli, D. Lillehaug, and A. Kohler, “Characterization of food spoilage fungi by FTIR spectroscopy,” *Journal of Applied Microbiology*, vol. 3, pp. 788–796, 2013.

[5] V. Shapaval, B. Walczak, S. Gognies, T. Møretrø, H. Suso, A. W. Åsli, A. Belarbi, and A. Kohler, “FTIR spectroscpic characterization of differently cultivated food related yeasts,” *Analyst*, vol. 138, pp. 4129–4138, 2013.

[6] V. Shapaval, N. K. Afseth, G. Vogt, and A. Kohler, “Fourier transform infrared spectroscopy for the prediction of fatty acid profiles in *Mucor* fungi grown in media with different carbon sources,” *Microbial Cell Factories*, vol. 1, 2014.

[7] M. Wenning, H. Seiler, and S. Scherer, “Fourier-Transform Infrared Microspectroscopy, a Novel and Rapid Tool for Identification of Yeasts,” *Applied and Environmental Microbiology*, vol. 68 (10), pp. 4717–4721, 2002.

[8] B. R. Bochner, “Global phenotypic characterization of bacteria,” *FEMS Microbiology Reviews*, vol. 33, pp. 191–205, 2009.

[9] A. Kohler, U. Böcker, V. Shapaval, A. Forsmark, M. Andersson, J. Warringer, H. Martens, S. W. Omholt, and A. Blomberg, "High-Throughput Biochemical Fingerprinting of *Saccharomyces cerevisiae* by Fourier Transform Infrared Spectroscopy," *PLOS One*, vol. 10 (2), 2015.

[10] K. Kneipp, H. Kneipp, I. Itzkan, R. R. Dasari, and M. S. Feld, "Ultrasensitive Chemical Analysis by Raman Spectroscopy," *Chemical Reviews*, vol. 99 (10), pp. 2957–2975, 1999.

[11] M. Wenning, F. Breitenwieser, R. Konrad, I. Huber, U. Busch, and S. Scherer, "Identification and differentiation of food-related bacteria: A comparison of FTIR spectroscopy and MALDI-TOF mass spectrometry," *Journal of Microbiological Methods*, vol. 103, pp. 44–52, 2014.

[12] P. Lasch, M. Drevinek, H. Nattermann, R. Grunow, M. Stämmler, R. Dieckmann, T. Schwenke, and D. Naumann, "Characterization of *Yersinia* Using MALDI-TOF Mass Spectrometry and Chemometrics," *Analytical Chemistry*, vol. 82 (20), pp. 8464–8475, 2010.

[13] D. Williams and I. Fleming, *Spectroscopic Methods in Organic Chemistry*. McGraw-Hill Higher Education, Shoppenhangers Road, Maidenhead, Berkshire, 6th ed., 2008. ISBN-13 978-0-07-711812-9.

[14] T. J. Griffin and L. M. Smith, "Single-nucleotide polymorphism analysis by MALDI–TOF mass spectrometry," *Trends in Biotechnology*, vol. 18, pp. 77–84, 2000.

[15] P. Lasch, T. Wahab, S. Well, B. Pályi, H. Tomaso, S. Zange, B. K. Granerud, M. Drevinek, B. Kokotovic, M. Wittwer, V. Pflüger, A. Di Caro, M. Stämmler, R. Grunow, and D. Jacob, "Identification of Highly Pathogenic Microorganisms by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry: Results of an Interlaboratory Ring Trial," *Journal of Clinical Microbiology*, vol. 53 (8), pp. 2632–2640, 2015.

[16] NIST, "NIST Chemistry Webbook." `http://webbook.nist.gov/chemistry/`. [Online; accessed 22 November, 2016].

[17] Internetchemistry, "Spectral databases." `http://www.internetchemistry.com/chemistry/spectral_database.htm`. [Online; accessed 22 November, 2016].

[18] L. Weisenberger, "NMR," in *Handbook of Spectroscopy* (G. Gauglitz and T. Vo-Dinh, eds.), WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003. ISBN 3-527-29782-0.

[19] A. Kohler, M. Hanafi, D. Bertrand, E. M. Qannari, A. O. Janbu, T. Møretrø, K. Naterstad, and H. Martens, “Interpreting several types of measurements in bioscience,” in *Biomedical Vibrational Spectroscopy*, John Wiley & Sons, Ltd, 2008.

[20] B. Zimmermann, V. Tafintseva, M. Bagcıoglu, M. Berdahl, and A. Kohler, “Analysis of Allergenic Pollen by FTIR Microspectroscopy,” *Analytical Chemistry*, vol. 88, pp. 803–811, 2015.

[21] E. Y. Jiang, *Advanced FT-IR Spectroscopy.* Thermo Electron Corporation, 2003.

[22] E. Szymanska, E. Saccenti, A. K. Smilde, and J. A. Westerhuis, “Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies,” *Metabolomics*, vol. 8, pp. 3–16, 2012.

[23] B. Kuehl, S. Marten, Y. Bischoff, G. Brenner-Weiss, and U. Obst, “MALDI-ToF mass spectrometry-multivariate data analysis as a tool for classification of reactivation and non-culturable states of bacteria,” *Analytical and Bioanalytical Chemistry*, vol. 401, pp. 1593–1600, 2011.

[24] S. Hassani, H. Martens, E. M. Qannari, and A. Kohler, “Model validation and error estimation in multiblock partial least squares regression,” *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 42–53, 2012.

[25] J. L. R. Arrondo and F. M. Goñi, “Structure and dynamics of membrane proteins as studied by infrared spectroscopy,” *Progress in Biophysics & Molecular Biology*, vol. 72, pp. 367–405, 1999.

[26] K. A. Lê Cao, D. Rossow, C. Robert-Granié, and P. Besse, “A Sparse PLS for Variable Selection when Integrating Omics data,” *Statistical Applications in Genetics and Molecular Biology*, vol. 7 (1), 2008.

[27] J. A. Westerhuis and A. K. Smilde, “Short communication: Deflation in multiblock PLS,” *Journal of Chemometrics*, vol. 15, pp. 485–493, 2001.

[28] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, “Analysis of multiblock and hierarchical pca and pls models,” *Journal of Chemometrics*, vol. 12, pp. 301–321, 1998.

[29] I. Karaman, N. P. Nørskov, C. C. Yde, M. S. Hedemann, K. E. B. Knudsen, and A. Kohler, “Sparse multi-block PLSR for biomarker discovery when integrating data from LC-MS and NMR metabolomics,” *Metabolomics*, vol. 11, p. 367–379, 2014.

[30] Lasch, P., “Data Format of Spectral Multifiles.” http://www.wiki.microbe-ms.com/index.php?title=Data_Format_of_Spectral_Multifiles. [Online; accessed 21 September, 2016].

[31] The Editors of Encyclopædia Britannica, “Klebsiella — bacteria genus — britannica.com.” https://global.britannica.com/science/Klebsiella. [Online; accessed 19 September, 2016].

[32] S. D. Eiref, M. Leitman, and W. Riley, “Hand sanitizer dispensers and associated hospital-acquired infections: friend or fomite?,” *Surgical Infections*, vol. 13 (3), pp. 137–140, 2012.

[33] D. Weber, W. A. Rutala, and E. S. Sickbert-Bennett, “Outbreaks Associated with Contaminated Antiseptics and Disinfectants,” *Antimicrobial Agents and Chemotherapy*, vol. 51 (12), pp. 4217–4224, 2007.

[34] E. Kirk, “*Bacillus subtilis*.” http://web.mst.edu/~microbio/BIO221_2009/B_subtilis.html. [Online; accessed 09 November, 2016].

[35] Wikipedia, the free encyclopedia, “*Escherichia coli*.” https://en.wikipedia.org/wiki/Escherichia_coli. [Online; accessed 09 November, 2016; Text is available under the Creative Commons Attribution-ShareAlike License].

[36] Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, “Details: DSM-3871.” https://www.dsmz.de/catalogues/details/culture/DSM-3871.html. [Online; accessed 09 November, 2016].

[37] Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, “Details: DSM-347.” https://www.dsmz.de/catalogues/details/culture/dsm-347.html. [Online; accessed 26 November, 2016].

[38] P. Lasch, H. Nattermann, M. Erhard, M. Stämmler, R. Grunow, N. Bannert, B. Appel, and D. Naumann, “MALDI-TOF Mass Spectrometry Compatible Inactivation Method for Highly Pathogenic Microbial Cells and Spores,” *Analytical Chemistry*, vol. 80, pp. 2026–2034, 2008.

[39] W. M. Albers, A. Annila, N. J. Goddard, G. Patonay, and E. Soini, “Bioanalysis,” in *Handbook of Spectroscopy* (G. Gauglitz and T. Vo-Dinh, eds.), WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003. ISBN 3-527-29782-0.

[40] M. Hof, “Basics of Optical Spectroscopy,” in *Handbook of Spectroscopy* (G. Gauglitz and T. Vo-Dinh, eds.), WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003. ISBN 3-527-29782-0.

[41] Ellesmere OCR A level Chemistry, "Infrared spectroscopy." https://sites.google.com/site/ellesmerealevelchemistry/module-4-core-organic-chemistry/4-2-alcohols-haloalkanes-and-analysis/4-2-4-analytical-techniques/-a-b-c-d-e-infrared-spectroscopy. [Online; accessed 01 September, 2016].

[42] I. Karaman, E. M. Qannari, H. Martens, M. S. Hedemann, K. E. B. Knudsen, and A. Kohler, "Comparison of Sparse and Jack-knife partial least squares regression methods for variable selection," *Chemomectrics and Intelligent Laboratory Systems*, vol. 122, pp. 65–77, 2013.

[43] L. F. Leopold, N. Leopold, H. Diehl, and C. Socaciu, "Quantification of carbohydrates in fruit juices using FTIR spectroscopy and multivariate analysis," *Spectroscopy*, vol. 26, pp. 93–104, 2011.

[44] K. Janssens, "X-ray Fluoresence Analysis," in *Handbook of Spectroscopy* (G. Gauglitz and T. Vo-Dinh, eds.), WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2003. ISBN 3-527-29782-0.

[45] Moxfyre, based on work of User:Pavlina2.0, "Raman energy levels." https://commons.wikimedia.org/w/index.php?curid=7845122. [Online; accessed 26 October, 2016, Published under a Creative Commons lisence: CC BY-SA 3.0].

[46] P. Zhou, "Choosing the Most Suitable Laser Wavelength For Your Raman Application." From https://www.google.no/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwicp5ffxOLQAhXFUhQKHX3BCKcQFggdMAA&url=http%3A%2F%2Fbwtek.com%2Fwp-content%2Fuploads%2F2015%2F07%2Framan-laser-selection-application-note.pdf&usg=AFQjCNGQdp2Xjnu99sx3jiwsp7ya-FVhjw&sig2=HGdQbdUd4RosILWJ5QZuyQ (pdf file); accessed 07.12.2016.

[47] Wikipedia, the free encyclopedia, "Raman Spectroscopy." https://en.wikipedia.org/wiki/Raman_spectroscopy. [Online; accessed 09 November, 2016; Text is available under the Creative Commons Attribution-ShareAlike License].

[48] M. W. Duncan, H. Roder, and S. W. Hunsucker, "Quantitative matrix-assisted laser desorption/ionization mass spectrometry," *Briefings in functional genomics and proteomics*, vol. 7 (5), pp. 355–370, 2008.

[49] J. P. Dekker and J. A. Branda, "MALDI-TOF Mass Spectrometry in the Clinical Microbiology Laboratory," *Clinical Microbiology Newsletter*, vol. 33 (12), pp. 87–93, 2011.

[50] E. Carbonelle, C. Mesquita, E. Bille, N. Day, B. Dauphin, J. Beretti, A. Ferroni, L. Gutmann, and X. Nassif, "MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory," *Clinical Biochemistry*, vol. 44, pp. 104–109, 2011.

[51] E. Carbonelle, P. Grohs, H. Jacquier, N. Day, S. Tenza, A. Dewailly, O. Vissouarn, M. Rottman, J. Herrmann, I. Podglajen, and L. Raskine, "Robustness of two MALDI-TOF mass spectrometry systems for bacterial identification," *Journal of Microbiological Methods*, vol. 89, pp. 133–136, 2012.

[52] J. O. Lay Jr., "Maldi-tof spectrometry of bacteria," *Mass Spectrometry Reviews*, vol. 20, pp. 172–194, 2001.

[53] N. Nicolaou, Y. Xu, and R. Goodacre, "Detection and Quantification of Bacterial Spoilage in Milk and Pork Meat Using MALDI-TOF-MS and Multivariate Analysis," *Analytical Chemistry*, vol. 84, pp. 5951–5958, 2012.

[54] T. R. Sandrin, J. E. Goldstein, and S. Schumaker, "MALDI TOF MS Profiling of bacteria at the strain level: A review," *Mass Spectrometry Reviews*, vol. 32, pp. 188–217, 2013.

[55] R. Dieckmann, R. Helmuth, M. Erhard, and B. Malorny, "Rapid Classification and Identification of Salmonellae at the Species and Subspecies Levels by Whole-Cell Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry," *Applied and Environmental Microbiology*, vol. 72 (24), pp. 7767–7778, 2008.

[56] A. Kohler, C. Kirschner, A. Oust, and H. Martens, "Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-sections of Beef Loin," *Applied Spectroscopy*, vol. 59, pp. 707–716, 2005.

[57] J. Shlens, "A tutorial on principal component analysis. derivation, discussion and singular value decomposition." Version 1, 25 March, 2003.

[58] R. van den Berg, H. J. C. Hoesloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7 (142), 2006.

[59] A. Kohler, N. K. Afseth, and H. Martens, "Chemometrics in biospectroscopy," in *Applications of Vibrational Spectroscopy in Food Science*, John Wiley & Sons, Ltd, 2010.

[60] B. Zimmermann and A. Kohler, "Optimizing Savitzky-Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy," *Applied Spectroscopy*, vol. 67, pp. 892–902, 2013.

[61] A. Rinnan, F. W. J. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *Trends in Analytical Chemistry*, vol. 28 (10), pp. 1201–1222, 2009.

[62] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36 (8), pp. 1627–1639, 1964.

[63] P. A. Gorry, "General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay)-Method," *Analytical Chemistry*, vol. 62 (5), pp. 570–573, 1990.

[64] Technische Universität München, Chair for Computer Aided Medical Procedures & Augmented Reality, "1D and 2D Gaussian Derivatives." `http://campar.in.tum.de/twiki/pub/Chair/HaukeHeibelGaussianDerivatives/gauss1d1.png`. [Online; accessed 29 August, 2016].

[65] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemomectrics and Intelligent Laboratory Systems*, vol. 112, pp. 92–99, 2012.

[66] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39 (3), pp. 491–500, 1985.

[67] H. Martens, J. P. Nielsen, and S. B. Engelsen, "Light scattering and light absorbance separated by extended multiplicative signal correction. application to near-infrared transmission analysis of powder mixtures," *Analytical Chemistry*, vol. 75 (3), pp. 394–404, 2003.

[68] W. Wu, D. L. Massart, and S. de Jong, "The kernel PCA algorithms for wide data. part I: theory and algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 36, pp. 165–172, 1997.

[69] F. Vogt and M. Tacke, "Fast principal component analysis of large data sets," *Chemometrics and Intelligent Laboratory Systems*, vol. 59, pp. 1–18, 2001.

[70] A. Höskuldsson, "PLS Regression Methods," *Journal of Chemometrics*, vol. 2, pp. 211–228, 1988.

[71] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15 (2), pp. 265–286, 2006.

[72] L. E. Wangen and B. R. Kowalski, "A multiblock partial least squares algorithm for investigating complex chemical systems," *Journal of Chemometrics*, vol. 3, pp. 3–20, 1988.

[73] S. Hassani, M. Hanafi, E. M. Qannari, and A. Kohler, "Deflation strategies for multi-block principal component analysis revisited," *Chemometrics and Intelligent Laboratory Systems*, vol. 120, pp. 154–168, 2013.

[74] Eigenvector Research staff and associates, "Using Cross-Validation." `http://wiki.eigenvector.com/index.php?title=Using_Cross-Validation`. [Online; accessed 3 October, 2016].

[75] F. Westad, N. K. Afseth, and R. Bro, "Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression," *Analytica Chimica Acta*, vol. 595, pp. 323–327, 2007.

[76] D. I. Broadhurst and D. B. Kell, "Statistical strategies for avoiding false discoveries in metabolomics and related experiments," *Metabolomics*, vol. 2 (4), pp. 171–196, 2006.

[77] S. Hassani, H. Martens, E. M. Qannari, G. I. Borge, and A. Kohler, "Analysis of -omics data: Graphical interpretation- and validation tools in multi-block methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 104, pp. 140–153, 2010.

[78] D. Bertrand and C. Cordella, "Saisir Webpage." `http://www.chimiometrie.fr/saisir_webpage.html`. [Online; accessed 03 October, 2016].

[79] P. Lasch, W. Beyer, H. Nattermann, M. Stämmler, E. Siegbrecht, R. Grunow, and D. Naumann, "Identification of *Bacillus anthracis* by Using Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry and Artificial Neural Networks," *Applied and Environmental Microbiology*, vol. 75 (22), pp. 7299–7242, 2009.

[80] T. A. Holme, "Denaturation - Chemistry Encyclopedia." `http://www.chemistryexplained.com/Co-Di/Denaturation.html`. [Online; accessed 29 November, 2016].

[81] J. C. Patton, E. J. MuMurchie, B. K. May, and W. H. Elliott, "Effect of Growth Temperature on Membrane Fatty Acid Composition and Susceptibility to Cold Shock of *Bacillus amyloliquefaciens*," *Journal of Bacteriology*, vol. 135 (3), pp. 754–759, 1978.

Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway