ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

CESENA CAMPUS

DEPARTMENT OF ELECTRICAL, ELECTRONIC, AND INFORMATION ENGINEERING

“GUGLIELMO MARCONI”

SECOND CYCLE DEGREE IN BIOMEDICAL ENGINEERING
Class: LM-21

THESIS TITLE

Machine Learning Techniques for Analysis of Biochemical Data

Graduation thesis in

SENSORS AND NANOTECHNOLOGY

Supervisor
Prof. Marco Tartagni
Co-Supervisor
Prof. Emanuele Domenico Giordano
Co-Supervisor
Prof. Joseph Lovecchio

Candidate
Matilde Arfilli

Academic Year 2022/2023

Contents

Introduction

This thesis, titled ”Machine Learning Techniques for Analysis of Biochemical Data,” represents a preliminary research step within a larger project aiming to quantify lactate in blood samples using non-invasive sensors integrated into diagnostic machinery used in extracorporeal flows to support artificial organs. This approach suggests the use of technologies such as optical spectral analysis, NIR, and MIR, deemed likely to succeed in this context.

The primary objective of this research is to address the long-term challenge of monitoring lactate concentration in the blood. Initially, predictive multivariate analysis systems, a subset of Machine Learning that does not use neural networks, will be tested on spectra acquired through a spectrophotometer. This approach, specifically designed for quantitative biological analysis, focuses on the accurate measurement of lactate, even in the presence of disturbance variables. Creating a Machine Learning model is a crucial tool to achieve the goal of monitoring lactate concentration in the blood over the long term, providing precise predictions based on biochemical spectra acquired through the spectrophotometer.

Initially, the research explored the behavior of the main components present in the blood, considering their spectrum in the visible, NIR, and MIR ranges. A specific chapter was dedicated to studying various blood components such as CO_2, hemoglobin, and oxygen to understand their behavior in the areas of interest.

Subsequently, Machine Learning techniques were further explored, ranging from Multiple Linear Regression (MLR) to latent variable modeling, with a particular focus on Principal Component Analysis (PCA), Principal Component Regression (PCR), and especially Partial Least Squares (PLS). The latter was identified as crucial for the type of dataset addressed, allowing the handling of the complexity of biochemical analyses.

The experimental phase is divided into two distinctive parts: the first focused on quantitative analysis using dyes, and the second dedicated to exploring lactate concentrations in cell cultures.

In the initial part of the experiment, samples containing Red Alizarin at various concentrations were analyzed, dissolved in a water solution, resulting in a predictive model based on the acquired spectra. The same approach was subsequently applied to samples containing PrestoBlue. Later, a third experiment was conducted using samples with variable concentrations of water, Alizarin, and PrestoBlue. This additional phase was designed to assess the model’s ability to handle complex mixtures of substances and verify its robustness in more realistic scenarios. The analysis of the results from this third experiment contributed to consolidating the model's effectiveness in addressing more complex and heterogeneous situations.

Following that, the same Machine Learning technique was applied to identify lactate concentrations in cell cultures. The extension of the Machine Learning technique's application to biological samples represents a significant step in exploring its effectiveness in realistic contexts. Initially, a preliminary test was conducted with nine biological samples, followed by a larger and more articulated experiment with 54 samples. This phase was designed to evaluate the model's ability to adapt to more complex samples, typical of real situations where substance concentrations can vary more dynamically and heterogeneously. This step serves as a preliminary phase for subsequent experiments to be conducted on blood samples.

Both parts of the experiment contributed positively, highlighting the versatility and effectiveness of the proposed model.

The complete set of these experimental phases provides a robust and fundamental foundation for the future development of the project. It is crucial to emphasize that this thesis represents only the beginning of a broader and more ambitious journey. The future goal is to develop an advanced application capable of directly measuring lactate concentration in blood spectra. This next step implies a more complex challenge and a more specific application of the model, opening the way to new research frontiers and contributing significantly to the development of innovative approaches in the field of biochemical analysis based on machine learning techniques and spectroscopy.

Chapter 1

Fundamentals of Spectroscopy: UV/Visible, Near Infrared (NIR), and Mid Infrared (MIR)

1.1 General introduction

Spectroscopy is a fundamental discipline in many scientific fields, offering a detailed analysis of the interactions between light and matter. This chapter explores three fundamental spectroscopic technologies: UV-Visible spectroscopy, Near-Infrared (NIR) spectroscopy, and Mid-Infrared (MIR) spectroscopy. For our specific application, these three spectroscopic techniques are of paramount importance. Each of these technologies possesses unique characteristics and diverse applications, making them essential tools in various scientific disciplines.

UV-Visible spectroscopy is based on the measurement of the absorption or transmission of UV or visible light by a sample. This technique provides valuable information about the composition of the sample and is widely used in areas such as quantitative analysis, pharmaceutical research, and environmental monitoring.

NIR spectroscopy extends into the near-infrared range of the electromagnetic spectrum, offering a detailed analysis of molecular vibrations. This technology is particularly useful for the study of biomolecules such as DNA and proteins, in addition to finding applications in environmental analysis and the characterization of nanostructured materials.

MIR spectroscopy, operating in the mid-infrared range, focuses on fundamental molecular vibrations, enabling precise identification of chemical functionalities in samples. Despite some limitations, such as the need to prepare samples in the form of powders or thin films, MIR spectroscopy remains a powerful tool in chemistry, pharmaceuticals, and biological research. This chapter will delve into the operating principles, key applications, and distinctive characteristics of each of these spectroscopic technologies, highlighting their advantages and limitations. The goal is to provide a comprehensive overview of spectroscopic technologies and underscore their importance in various scientific fields.

1.2 Physical Principles of Spectroscopy

Light can be thought of as particles of energy moving through space, exhibiting wave-like properties. This representation suggests that the energy associated with a beam of light is not uniformly distributed along the electric and magnetic fields of the wave but is concentrated in discrete packets known as photons. The duality that emerges indicates that light exhibits both particle and wave behaviors.

Phenomena related to the propagation of light, such as interference, diffraction, and refraction, find clearer explanations through the adoption of the wave theory of electromagnetic radiation. However, to better understand how light interacts with matter in processes like absorption and emission spectroscopy, it's important to recognize that light behaves like particles. This duality isn't exclusive to light; even fundamental particles of matter, such as electrons, protons, and neutrons, exhibit wave-like behaviors.

The wave properties of electromagnetic radiation are described in terms of frequency, wavelength, and amplitude. Frequency (ν), typically measured in Hertz (Hz), is the number of complete cycles of a wave passing through a given point in one second. It is the reciprocal of the period (p), the time needed for successive wave maxima to pass through a fixed point. Wavelength (λ) is the distance between successive maxima on any wave. Spectroscopic data are sometimes reported in terms of frequency or reciprocal wavelengths in units of cm^{-1}, commonly encountered in IR spectroscopy.

The wavelength and frequency are related by the following equation:

$$\nu\lambda = c$$

where c is the speed of light in a vacuum. If, however, we consider a material medium, the propagation speed (v_i) of an electromagnetic wave in a specific medium ”i” can be calculated using the formula:

$$v_i = \nu\lambda_i$$

where ν is the frequency of the wave, and λ_i is its length in the i-th material medium.
The frequency of an electromagnetic wave is intrinsically tied to the radiation source, remaining constant along its path. However, the propagation speed can undergo slight variations depending on the medium, causing proportional changes in the wavelength of the radiation.
The amplitude of the wave, indicative of the magnitude of the electric vector at the wave maxima, plays a crucial role in defining the radiant power and radiant intensity of a radiation beam as they are proportional to the square of the amplitude of the associated waves. Electromagnetic waves consist of oscillating magnetic and electric fields, in phase with each other and perpendicular to the direction of wave propagation.

Wave Fronts, In-Phase Waves, and Interference in Electromagnetic Wave Propagation

The propagation of electromagnetic waves unveils interesting phenomena, including interference, a crucial aspect for fully understanding the characteristics of light waves. To thoroughly explore interference, it is necessary to adopt an unequivocal approach that embraces the wave nature of light. From this perspective, fundamental concepts such as wave fronts, wave trains, and rays emerge, playing an indispensable role in our comprehension of luminous phenomena. However, to fully comprehend the complexity of interference, it is essential to also consider the corpuscular nature of light, manifested through the existence of photons.
The propagation of electromagnetic waves is often described in terms of **wave fronts or wave trains**. A wave front represents the locus of a set of points, all in phase, and can be visualized as a concentric ring for a point source of light. When the observation is sufficiently far from the source, wave fronts can also represent planes of light, making the curved surface appear planar. The connection between consecutive wave fronts can be illustrated by connecting maxima, minima, or both. A series of wave fronts, all *in-phase*, are referred to as a **wave train or long wave**, and can also be represented by a series of light rays, which depict the path of photons, thereby referencing the corpuscular nature of light.
This concept of *in-phase waves* is fundamental in interference, a phenomenon where two or more overlapping waves generate a resulting wave with an amplitude determined by the algebraic sum of the individual waves. Constructive interference occurs when the waves are in phase, while destructive interference happens when they are out of phase by 180°. In spectroscopy, these notions are crucial for interpreting diffraction data and designing instruments that involve the dispersion or selection of radiation, such as monochromators and interference filters. Understanding the wave-particle duality of light, manifested in photons, contributes to explaining luminous phenomena and provides a comprehensive approach to describing elec-

tromagnetic waves.

The particulate interaction of light adds an intriguing layer to our scientific narrative. Photons, discrete packets of energy, exhibit wave behavior during their motion through space. This wave-particle duality becomes evident in the equation linking the energy (E) of a photon to the frequency (ν) of the associated wave:

$$E = h \cdot \nu$$

Here, E represents the energy of a photon, h is Planck's constant, and ν is the frequency of the associated wave. This equation underscores a crucial aspect: in monochromatic light—electromagnetic radiation comprising waves with a single frequency and wavelength—all photons possess equivalent energy. Furthermore, similar to how the frequency of a wave remains a constant determined by the radiation source, the energy of associated photons remains invariant.

Delving into the brightness of monochromatic light, its characterization in terms of the particulate nature of light involves the interplay of photon flux and energy per photon. Photon flux refers to the number of photons crossing a unit area perpendicular to the beam per unit time. Consequently, altering the brightness of a monochromatic light beam necessitates a change in photon flux. In spectroscopy, terms such as ”brightness” are typically avoided, and there is a preference for using expressions like radiant power (P) or radiant intensity (I) to denote the quantity of radiant energy impacting a specific area per unit time.

Electromagnetic Spectrum

Spectroscopy, a powerful analytical technique, relies on the fundamental principles governing the interaction between matter and electromagnetic radiation. This interaction occurs within the broader framework of the electromagnetic spectrum, which encompasses a vast range of wavelengths and frequencies.

The electromagnetic spectrum represents the set of all possible frequencies of electromagnetic radiation. The entire spectrum is divided into different regions based on the wavelengths.

Looking at the image 1.1, we can see that the electromagnetic spectrum is characterised by a wide range of wavelengths and frequencies, from high-energy gamma rays to the gentle ripples of radio waves. For my research, **Infrared radiation (IR)** is of particular interest, as well as **UV/Visible** radiation.

Infrared radiation (IR) is characterised by a wavelength longer than the visible range and shorter than radio waves, i.e. a wavelength between 700 nm and 1 mm (*band infrared*).

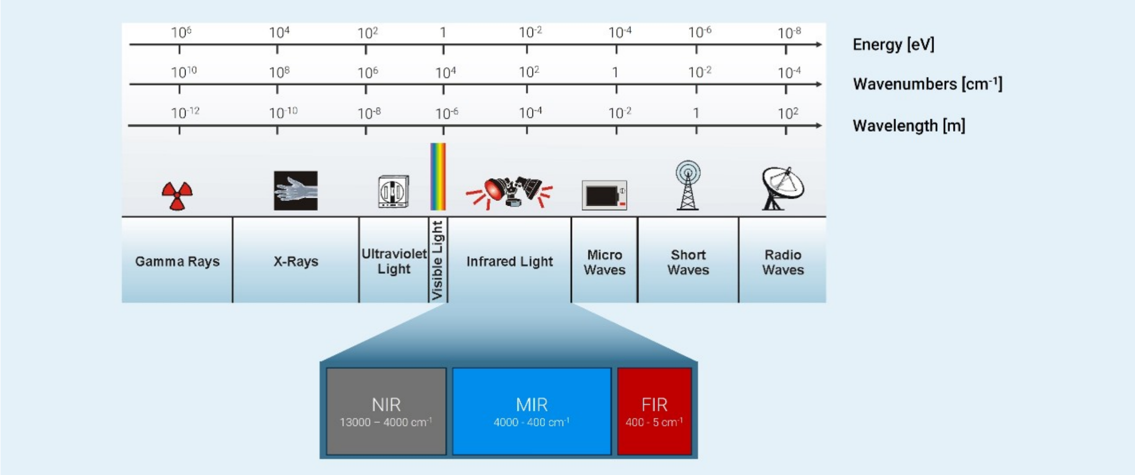


Figure 1.1: Electromagnetic spectrum

Infrared Radiation can be further subdivided into three categories:

- Near infrared (NIR), which includes electromagnetic radiation with wavelengths between $800 - 2500nm$ $(13000 - 4000cm^{-1})$
- Mid infrared (MIR), which includes electromagnetic radiation with wavelengths between $2500nm - 25\mu m$ $(4000 - 400cm^{-1})$
- Far infrared (FIR), which includes electromagnetic radiation with wavelengths between $25 - 500\mu m$ $(400 - 5cm^{-1})$

NIR has the shortest wavelength with higher wave-numbers, while FIR has the longest wavelength with lower wave-numbers.

Quantization of Energy Levels in Spectroscopy

The fundamental comprehension of spectroscopy hinges on the quantization of energy levels within atoms and molecules. This quantization, arising from the wave-particle duality of matter, assigns distinct and discrete energy values to electronic, vibrational, and rotational states. In atoms, electronic energy levels, depicted as orbits and subshells, govern the potential energy of electrons. Only specific energy levels are allowed for atomic electrons, resulting in a series of quantized electronic states. The transitions between these states, triggered by the absorption or emission of photons, form the basis for electronic spectroscopy, providing insights into the composition and structure of substances.

Molecules, with their additional vibrational and rotational degrees of freedom, present a more intricate energy landscape. Vibrational energy levels, associated with atomic motion within a molecule, are quantized, contributing to the complexity of the potential energy diagram. Similarly, rotational energy levels, representing the rotation of molecules around their centers of gravity, introduce another layer of quantization. The energy differences between these levels, analogous to the energy of specific photons, become pivotal in techniques such as microwave spectroscopy.

The quantized nature of energy levels introduces the concept of spectroscopic fingerprints. Each species possesses a unique set of energy spacings, allowing spectroscopists to identify and characterize substances based on their distinct energy signatures. Techniques such as absorption and emission spectroscopy leverage these quantized energy transitions to unravel the composition, structure, and behavior of matter.

Understanding the quantization of energy levels is fundamental for interpreting spectroscopic data, guiding the design of experiments, and extracting meaningful information about the intricate nature of atoms and molecules. As we embark on an in-depth exploration of spectroscopy principles, instrumentation, and applications, the elucidation of UV/Visible, NIR, and MIR spectroscopy awaits, providing a comprehensive understanding of their underlying physical principles and practical utility.

Energy-level transition

Understanding the interaction between electromagnetic radiation and matter entails a detailed exploration of absorption and emission processes. These processes, regulated by the quantized energy levels of atoms and molecules, give rise to distinct energy level transitions forming the basis of spectroscopic analyses.

In the absorption process, atoms or molecules absorb photons, transitioning them from lower energy states to more excited states. This discrete transition is quantitatively matched to the energy content of the incident photons, resulting in characteristic absorption spectra. The absorptivity of a compound, a wavelength-dependent constant, defines the relationship between concentration and measured absorbance.

Molecular transitions induced by absorption encompass electronic, vibrational, and rotational changes, with simultaneous transitions contributing to broader peaks in spectra. The absorption spectrum becomes a fingerprint, revealing the relative absorptivity of photons with different energies. On the other side, emission involves the release of energy in the form of photons when excited molecules return to the ground state. The relaxation processes, often involving

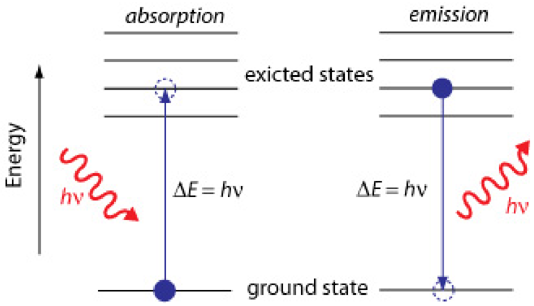


Figure 1.2: The absorption of energy in the form of electromagnetic radiation causes the transition from a lower to a higher energy level. When the molecule returns to the fundamental energy level, it releases an amount equivalent to the difference between the two energy levels.

collisions with other molecules, lead to the dissipation of excess energy as heat. In specific cases, molecules emit photons through fluorescence or phosphorescence, providing additional insights into their dynamics.

Fluorescence, a subset of emission, unfolds when molecules emit lower-energy, longer-wavelength photons during vibrational relaxation. The emitted photons mirror the energy difference between the lowest vibrational level of the excited state and the ground state. This phenomenon contributes to the broader peaks observed in fluorescence spectra.

Understanding these energy level transitions, both in absorption and emission, unravels the dynamic behavior of atoms and molecules during spectroscopic analyses. The rapid relaxation processes, driven by the tendency to minimize internal energy, ensure that, under normal conditions, the population of molecules in the ground state remains essentially unchanged.

Quantitative Absorption Spectroscopy and Lambert-Beer Law Throughout the Spectrum

Quantitative absorption spectroscopy, grounded in the principles of the Lambert-Beer Law, serves as a versatile analytical tool across the electromagnetic spectrum. The primary aim is to determine the concentration of an analyte within a sample solution by quantifying the light absorption during its traversal through the sample.

In practice, the sample solution is confined within an absorption cell and exposed to radiation of specific wavelength. The comparison of the radiation passing through the sample to a reference sample facilitates the estimation of analyte concentration. The fundamental process involves the absorption of photons by the analyte, resulting in a decrease in radiant power as the radiation passes through the solution.

The relationship between incident and exiting beam powers finds expression in terms of transmittance (T) or absorbance (A). While transmittance intuitively represents the fraction of absorbed light, absorbance emerges as a more practical parameter directly proportional to the concentration of the absorbing species. This relationship is articulated through Beer's Law, a cornerstone of quantitative spectroscopy:

$$A = \varepsilon \cdot b \cdot c$$

Here, A is absorbance, ε is the absorptivity (or molar absorption coefficient), b is the pathlength through the solution, and c is the concentration of the absorbing species. The significance of absorbance lies in its direct proportionality to concentration under suitable conditions. The practical measurement of absorbance in a laboratory setting closely aligns with the following formulation:

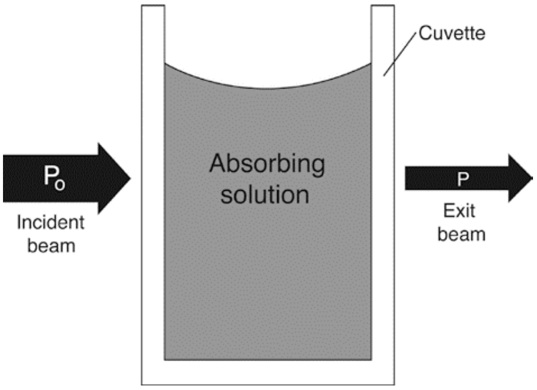


Figure 1.3: Attenuation of Radiation in an Absorbing Solution [35]

$$A = \log\left(\frac{P_{\text{solvent}}}{P_{\text{analyte solution}}}\right) \approx \log\left(\frac{P_0}{P}\right) \tag{1.1}$$

Here, P_{solvent} represents the radiant power of the beam exiting the cell containing the solvent (blank), and $P_{\text{analyte solution}}$ signifies the radiant power of the beam exiting the cell containing the analyte solution. The ratio $\frac{P_0}{P}$ encapsulates the essence of absorbance, emphasizing the impact of the analyte on the transmitted radiation. This formulation enables a practical and quantitative assessment of analyte concentration, providing valuable insights in various spectroscopic applications.

1.3 UV/Visible Spectroscopy

UV/Visible spectroscopy plays a pivotal role in the field of analytical techniques, providing valuable information about the absorption characteristics of atoms and molecules. This robust spectroscopic method focuses on the interaction of matter with ultraviolet (UV) and visible light, elucidating intricate details of molecular composition. The absorption and transmission of light at specific wavelengths offer profound insights into the electronic structure of species, facilitating the identification and quantification of substances. In this section, we examine the fundamental principles and applications of UV/Visible spectroscopy, highlighting its significance as a potent tool across various scientific disciplines.

1.3.1 Spectroscopic Principles and Common Methodologies for NIR, MIR, and UV/Visible: From Calibration to Model Validation

Within the scope of our research, quantifying substance concentrations involves employing chemometric models, emphasizing their application in spectroscopy. These are mathematical models obtained using samples of the same product or class of products. The analysis of the obtained data, particularly the process of finding a possible correlation between the spectrum's characteristics and the substance under study, takes a considerable amount of time. Once the instrument's spectrum has been plotted, a pre-processing procedure is carried out, in which, if necessary, baseline subtractions are performed to reduce the effect of substances present in the sample but not of interest for our investigations. However, it's crucial to handle these procedures carefully, as they can introduce additional noise that deteriorates data quality, requiring a balanced approach.

Furthermore, dealing with outliers is essential, as they can either be potentially harmful to our purpose or very useful. Therefore, it is important to conduct a proper check before proceeding with the creation of a calibration model.

Developing such a model involves calculating a regression equation based on the spectra and all available information about the molecule. For linear calibration, multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) are typically used. In all these methods, parameters such as factors, loadings, and scores are evaluated. The number of factors is relevant because if there are too many, the model has high variability in classification, resulting in an extremely complex model and, very often, prediction errors. Conversely, too few factors can lead to underfitting, resulting in high classification discrepancy and potential omission of important information. Scores are used to check the homogeneity of the samples, while loadings are used to understand how the variables are weighted within the space described by the principal components.

Having completed these steps, the validation phase involves checking if the created model is correct. The samples are divided into two groups: the test set and the training set. A leave-one-out procedure is then applied, where one sample is used for calibration while the others are used to study the model.

This methodology proves essential for our thesis, enabling the measurement of substance concentrations by analyzing the unique spectra they yield.

1.3.2 Instrumental Aspects of UV/Visible Spectroscopy

This section delves into the essential instrumental components of UV/Visible spectroscopy with a focus on the technological foundations that drive this analytical technique. It covers the practical aspects, from detailing light sources to the strategic selection of wavelengths and the intricacies of detection mechanisms. The goal is to provide a comprehensive understanding of the technical intricacies that impact the precision and effectiveness of UV/Visible spectroscopic analyses.

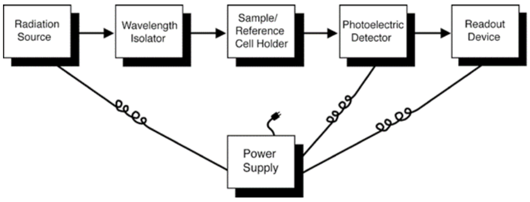


Figure 1.4: Illustration of the Key Components in a Basic Single-Beam UV-Vis Absorption Spectrophotometer Setup [35]

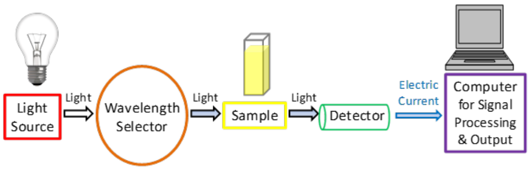


Figure 1.5: Schematic representation of a UV/Visible spectrometer [31]

Light Source

Light possesses energy that is inversely proportional to its wavelength. Shorter wavelengths correspond to higher energy, while longer wavelengths carry less energy. The absorption of light occurs when a specific amount of energy promotes electrons in a substance to a higher energy state, resulting in detectable absorption signals. Different bonding environments within a substance require varying amounts of energy to induce this transition, leading to absorption at distinct wavelengths.

The visible light spectrum, perceivable by humans, ranges from approximately 380 nm (violet) to 780 nm (red). UV light, with wavelengths shorter than visible light, extends to about 100 nm. Describing light in terms of its wavelength is crucial in UV-Vis spectroscopy, enabling the analysis and identification of substances by pinpointing specific wavelengths associated with maximum absorbance. This spectroscopic technique proves invaluable in various applications, as discussed in the subsequent sections.

In UV-Vis spectroscopy, a thorough understanding of its components and processes is essential for accurate and insightful analyses. The choice of a suitable light source, often a xenon lamp emitting across both UV and visible ranges, is pivotal. For visible light, *tungsten filament lamp* are commonly employed. These lamps emit radiation across the visible spectrum, covering wavelengths from 350 to 2,500 nm. Additionally, tungsten filament lamps find utility in near-infrared spectroscopy applications due to their ability to emit radiation in the relevant wavelength range. On the other hand, for UV light applications, *deuterium electrical-discharge lamps* serve as the predominant radiation source. These lamps generate a continuous spectrum of radiation, ranging approximately from 160 nm to 375 nm in the UV range. It is noteworthy that deuterium lamps utilize quartz windows, and to maintain optimal performance, they should be paired with quartz sample holders. This precaution is essential because conventional glass tends to absorb radiation below 350 nm, affecting the accuracy and reliability of UV measurements.

Wavelength Selection

Precise wavelength selection is paramount in UV/Visible spectroscopy, requiring the use of tools like monochromators and filters to specify wavelengths accurately.

A standard monochromator comprises entrance and exit slits, concave mirrors, and a dispersing element, typically a grating. Polychromatic light enters through the entrance slit and is focused by a concave mirror. Subsequently, it undergoes dispersion, separating radiation into different wavelengths. A subsequent concave mirror reflects the dispersed radiation along the focal plane, emitting radiation through the exit slit. The resulting radiation encompasses a narrow wavelength range, approximately centered on the specified wavelength. The term ”bandwidth” defines the size of the wavelength range passing through the exit slit, characterizing the emitted radiation.

Monochromators, particularly those employing diffraction gratings, offer versatility in wavelength selection. Filters are often combined with monochromators to further refine wavelength specificity. This precision is vital for accurate measurements and improving the signal-to-noise ratio in spectroscopic data.

Moving to sample analysis, the selected wavelength passes through the sample, and a reference or ”blank sample” measurement is imperative for accurate results. Understanding the materials used is crucial; for instance, plastic cuvettes are unsuitable for UV absorption studies, necessitating quartz sample holders for UV examination.

Detection

Detection, the next step, involves converting light into an electronic signal using detectors. Photomultiplier tubes (PMTs) and semiconductor-based devices like photodiodes and charge-coupled devices (CCDs) are common detectors. PMTs excel at detecting low light levels, while semiconductor-based detectors operate by allowing an electric current proportional to light intensity to pass through.

Finally, UV-Vis spectroscopy data is presented graphically, typically as absorbance against wavelength. Beer–Lambert’s law is applied for concentration determination. This law establishes a quantitative relationship, especially valuable when a linear correlation exists with standard solutions. In summary, comprehending these components and processes empowers researchers to leverage UV-Vis spectroscopy as a potent analytical tool for characterizing substances based on their absorbance properties.

Application

UV/Visible spectroscopy finds diverse applications across various scientific disciplines due to its versatile capabilities. One primary application is in the field of chemistry, where it is extensively used for quantitative analysis. UV/Vis spectroscopy aids in determining the concentration of a substance in a solution by leveraging Beer-Lambert’s law. This makes it valuable in pharmaceuticals, environmental monitoring, and quality control processes across industries.

In the life sciences, UV/Vis spectroscopy plays a crucial role in biomolecule analysis. Researchers utilize it to study nucleic acids (DNA, RNA) and proteins, as these biomolecules exhibit characteristic absorption bands in the UV region. This is particularly essential in molecular biology, where understanding the structure and concentration of biological macromolecules is fundamental.

In environmental science, UV/Vis spectroscopy is employed for monitoring water quality. The technique helps identify and quantify pollutants in water samples, contributing to the assessment of environmental impact and the development of remediation strategies.

Furthermore, UV/Vis spectroscopy is pivotal in material science, specifically in the characterization of nanomaterials. Researchers use it to study the electronic transitions and properties of nanoparticles, offering insights into their behavior and potential applications in various technologies.

In the pharmaceutical industry, UV/Vis spectroscopy is applied to analyze drug formulations, ensuring the quality and consistency of pharmaceutical products. The technique assists in monitoring the stability of drugs and detecting impurities, supporting the development of safe and effective medications.

Overall, UV/Visible spectroscopy's broad range of applications underscores its significance as a powerful analytical tool with implications spanning chemistry, life sciences, environmental science, material science, and pharmaceuticals.

1.3.3 Advantages and Limitations

UV-Vis spectroscopy offers several advantages that contribute to its widespread use. Firstly, it is non-destructive, allowing for sample reuse or further processing and analysis. The technique provides quick measurements, seamlessly integrating into experimental protocols. Its user-friendly instruments require minimal training, and data analysis is generally straightforward, requiring minimal processing. Moreover, the affordability of acquiring and operating UV-Vis spectrophotometers makes them accessible to many laboratories.

However, like any method, UV-Vis spectroscopy has its limitations. Stray light, arising from imperfections in wavelength selectors or environmental factors, can lead to measurement errors. Light scattering, often induced by suspended solids or bubbles in the sample, may result in irreproducible outcomes. Interference from multiple absorbing species, such as different chlorophylls in a sample, requires careful separation for accurate quantitative analysis. Geometrical considerations, particularly misalignment of components like the sample-holding cuvette, can yield inaccurate and irreproducible results. Thus, proper instrument alignment and basic user training are essential to avoid misuse and ensure reliable outcomes.

1.4 Infrared Spectroscopy

Infrared spectroscopy (IR) is a powerful analytical technique that exploits the interactions between light and matter to provide detailed information about the chemical composition of substances. This methodology relies on the ability of molecules to absorb specific frequencies of infrared radiation, allowing for the identification and analysis of chemical bonds present in samples. The broad application of IR spectroscopy ranges from the characterization of organic and inorganic compounds to the determination of complex molecular structures.

For our research purposes, we will specifically focus on Near-Infrared (NIR) and Mid-Infrared (MIR) spectroscopy. These specialized branches of IR spectroscopy offer unique insights into molecular vibrations, making them invaluable tools for studying biomolecules, nanostructured materials, and various chemical compositions. Our exploration will delve into the operating principles, key applications, and distinctive characteristics of NIR and MIR spectroscopy, highlighting their relevance in our scientific investigations.

1.4.1 Near Infrared (NIR) Spectroscopy

Quantitative analysis of biomedical samples often leverages measurements in the near-infrared (NIR) spectral region, typically ranging from 0.7 to 2.5 μm (700 to 2500 nm). This spectral range is widely favored for such analyses compared to mid-infrared measurements in the biomedical field. Various commercial instruments are available for conducting compositional analyses of biomedical samples using near-IR spectroscopy. An important advantage of near-IR spectroscopy in the biomedical context is its capacity to directly assess the composition of solid biological samples, employing techniques like diffuse reflection.

Physical Principles

In the realm of spectroscopic analysis, Near-Infrared (NIR) spectroscopy stands out for its unique principles governing the interaction between radiation and solid surfaces. When radiation impinges on a sample surface, a portion undergoes specular reflection, which provides limited information about the sample. Specularly reflected radiation is often redirected back toward the energy source. In contrast, another portion penetrates the sample surface, under-going diffuse reflection. This diffusely reflected radiation exits the sample at random angles, covering 180°. During this process, chemical constituents within the sample can absorb radiation, contributing to the information encoded in the diffusely reflected radiation. The absorbed energy at specific wavelengths indicates the chemical composition of the sample.

The size and shape of sample particles influence the amount of radiation that penetrates and exits the sample surface. Therefore, it is necessary to reduce solid materials into fine and uniform particles or apply mathematical corrections during instrument calibration to compensate for these effects.

In the NIR region, absorption bands primarily consist of overtones and combinations, resulting in weak intensity absorptions. This characteristic proves advantageous as the observed absorption bands predominantly arise from functional groups containing hydrogen atoms attached to carbon, nitrogen, or oxygen—common in major biomedical constituents. The broad and overlapping nature of absorption bands in the near-IR region, while complex, is valuable for quantitative analysis.

Notably, the near-IR spectra of various biomedical constituents exhibit broad and overlapping absorption bands. The dominance of water-related -OH groups, evident in the spectra of tissues, is still prominent in samples with varying compositions. Distinct absorption bands arising from biomolecules contribute to the complexity of the spectra. The NIR spectroscopy principles are grounded in the analysis of these absorption bands, providing a robust foundation for the quantitative assessment of sample composition.

The NIR spectrum is without a doubt the most natural and richest resource of information on the anharmonicity of the vibration of molecules. The absorption of the NIR transition gradually decreases progressively towards higher tones and higher order combinations. The coexistence of different bands, such as the first, second, and third tones, is a great advantage. The low intensity of the bands allows molecules to be placed in a solution with a much wider range of possible concentrations, guaranteeing the possibility of more in-depth investigations on intermolecular interactions.

In the NIR region, the bands generated by the vibrations of molecules, such as $\mathrm{C-H}$, $\mathrm{O-H}$, and $\mathrm{N-H}$, are very elaborate. Some types of bands undergo enhancement in their intensity. The bands due to non-associated species are usually much more intense than those belonging to aggregates of molecules. NIR band heights often carry valuable information.

The specificity of the vibrational effects in the NIR, profoundly different from those found in the MIR spectra, creates a large amount of independent spectral information, essential for physical chemistry. However, NIR spectral analysis is subject to ambiguities due to overlaps,

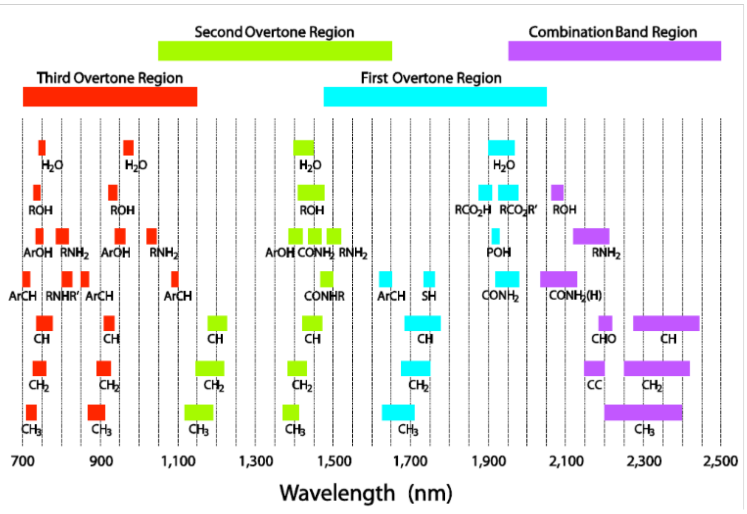


Figure 1.6: Regions of NIR in respect of the functional groups involved

anharmonicity, and the omnipresence of coexisting effects, resulting in convoluted spectral variations. Similar reasons have also been an obstacle in analytical NIRS. The complexity of spectra has necessitated the use of spectral pre-treatment and advanced data analysis methods, such as two-dimensional correlation spectroscopy (2D-COS).

The low absorption of NIR radiation is also added to the advantages described above. This feature allows for application for hyperspectral imaging purposes. The deep penetration of NIR light is crucial, allowing for effective in-depth investigation of the sample, a fact that finds wide application in biophysical and biomedical studies.

To perform quantitative and qualitative measurements, the system must have access to different chemometric models, representing the analyte being tested. These are mathematical models obtained using samples of the same product or class of products. The analysis of the obtained data, particularly the process of finding a possible correlation between the spectrum's characteristics and the substance under study, takes a considerable amount of time. Once the instrument's spectrum has been plotted, a pre-processing procedure is carried out, in which, if necessary, baseline subtractions are performed to reduce the effect of substances present in the sample but not of interest for our investigations. However, it's crucial to handle these procedures carefully, as they can introduce additional noise that deteriorates data quality, requiring a balanced approach.

Furthermore, dealing with outliers is essential, as they can either be potentially harmful to our purpose or very useful. Therefore, it is important to conduct a proper check before proceeding with the creation of a calibration model.

The creation of such a model involves calculating a regression equation based on the NIR spectra and all the information already available about the molecule to be analyzed. For linear calibration, multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) are typically used. In all these methods, parameters such as factors, loadings, and scores are evaluated. The number of factors is relevant because if there are too many, the model has high variability in classification, resulting in an extremely complex model and, very often, prediction errors. Conversely, too few factors can lead to underfitting, resulting in high classification discrepancy and potential omission of important information. Scores are used to check the homogeneity of the samples, while loadings are used to understand how the variables are weighted within the space described by the principal components. Having completed these steps, the validation phase involves checking if the created model is correct. The samples are divided into two groups: the test set and the training set. A leave-one-out procedure is then applied, where one sample is used for calibration while the others are used to study the model.

Instrumentation

The instrumentation we use to perform NIR spectroscopy consists of several components as illustrated in the figure.

Specifically, we can find:

- Light Source: A single polychromatic source is generally used for NIR spectroscopy. It has an inert solid, usually tungsten (a halogen lamp or a deuterium lambada), which radiates uniformly in the IR spectral range. In the instrument, we are considering we have as a light source a Halogen Lamp of the AvaLight series that composes a wavelength range from 400 to $2500nm$.

- Monochromator or Spectrometer: A monochromator or spectrometer is used to disperse the incoming NIR light into its various wavelengths (or frequencies). This component allows for the selection of specific wavelengths or bands of interest. The spectroscope features a specify NIR grating. Unlike discrete wavelength spectroscopes, this instrument is definitely much more flexible in terms of application.

- Detector: The detector measures the intensity of the NIR light after it has passed through the sample. Common detectors include photodiodes or charge-coupled devices (CCDs).

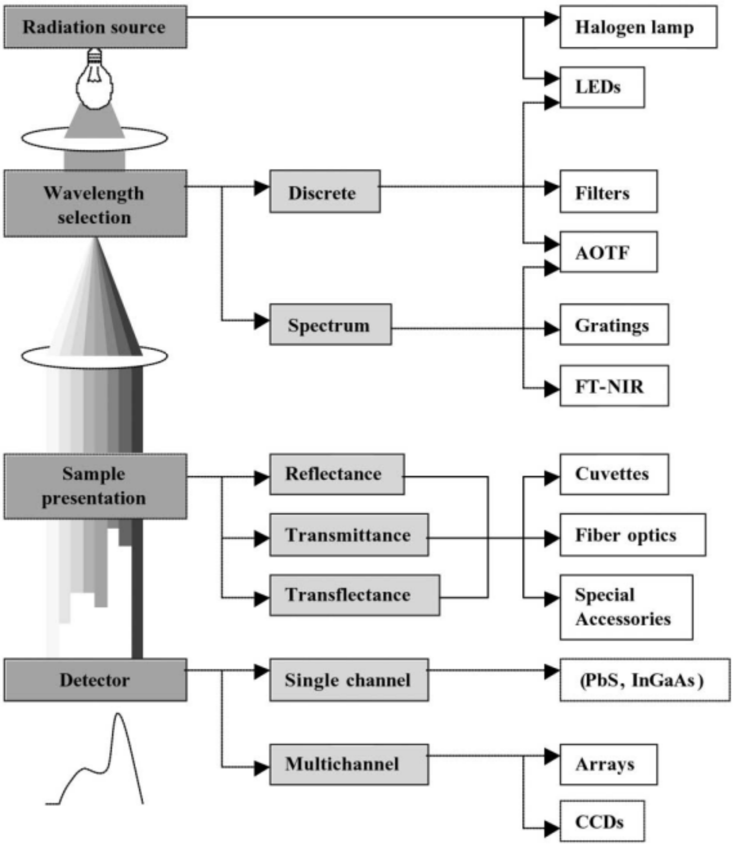


Figure 1.7: Principal features of NIR spectroscopy equipment [6]

The detector converts the NIR signal into an electrical signal for further analysis. In our instrument, we have InGaAs detector with 2 stage TEC, which consists of InGaAs is a semiconductor material commonly used for infrared photodetectors. It is sensitive to infrared light in the range of about 0.9 to 2.6 micrometers (µm). A TEC is a solid-state cooling device that uses the Peltier effect to transfer heat from one side (the cold side) of the device to the other side (the hot side). In the context of an InGaAs detector, a two-stage TEC is used to maintain the detector at a stable, low temperature, which is crucial for reducing noise and improving sensitivity. The two-stage TEC typically consists of two thermoelectric modules stacked together to achieve greater cooling efficiency.

For my research, a crucial step involved evaluating various devices to select the one most suitable for the intended application. After careful consideration, we opted for the **AvaSpec-NIR-2.5-HSC-EVO NIRLine Near-infrared Fiber Optic Spectrometer**.
Two configurations are available which differ in the number of pixels of the detector array. In

fact, there is a version with 256 pixels and one with 512, the difference is essentially in terms of resolution, which is better with a higher number of pixels.

NIRLine provides leading-edge performance for dispersive NIR instruments with toroidal focusing mirrors and dynamic dark correction (DDC) for enhanced stability. The NIRLine is comprised of both thermoelectrically cooled and uncooled instruments.They present thermoelectric, Peltier-cooled InGaAs detectors which support cooling down to-25°C against ambient. The InGaAs array are optimal for measurements with wavelengths in the $900 - 2500nm$ range, suitable for different applications. The detector consists of a charge amplifier array with CMOS transistors, a shift register, and a timing generator. For InGaAs detectors, the dynamic range is limited by dark noise. In our case, since we will be using a device with an extended range between 2.0 and 2.5 μm, all are equipped with 2-stage thermoelectric cooling (TEC) to reduce dark noise.

There are two different modes are available: high-sensitivity (HS) and low-noise (LN). The default setting is the HS mode, which provides a better signal at a shorter integration time. The other mode of operation, the LN mode, provides a better S/N (signal-to-noise) performance. Sensitivity, S/N, dark noise and dynamic range are given as HS and LN values. To make the choice of the device we had to make considerations at the Signal Noise Ratio (SNR) level, in particular for our application we will take into consideration the data relating to the Low Noise (LN) configuration.

Figure 1.8: AvaSpec-NIR-2.5-HSC-EVO NIRLine Near-infrared Fiber Optic Spectrometer

A further characteristic that was requested from us is that the measurement is carried out with

a full scale of 0 - 10 mmol/L and with an uncertainty of ± 0.1.
Initially, however, we want to do some tests with an uncertainty of ± 1 and see if we can obtain results already with these values.
In order to understand which of the tools proposed to us was the most suitable, we made some evaluations taking as reference some data obtained through a previous study presented by the article *"In-silico investigation towards the non-invasive optical detection of blood lactate"* [11].
From the graphs present in the article, we were able to see that the instrument must be able to have a sensitivity in the order of magnitude of 0.0005 in terms of assorbance.
The SNR which is reported in the spectrometer's technical data sheet is equal to $5000 : 1$, this means that we are able to detect variations in the order of magnitude of 0.0002, compliant with the specifications that we considered during the testing phase.
Particularly interesting is the fact that the information reported on the instrument's technical data sheet relating to the SNR was obtained considering the smallest integration time equal to 20 μs. Theoretically, by increasing the integration time I also increase the quantity of information acquired by the instrument, therefore allowing us to obtain good results, consistent with the specifications, even with a smaller uncertainty.

Advantages and Limitations

In general, IR spectroscopy has several advantages including:

- The analysed samples do not require any specific pretreatment, such as the addition of a radioactive dye, so that analysis is practically in real-time; in addition, data obtained by the NIRS method have greater reproducibility;
- Absence of thermal noise;
- Non-invasive and non-destructive technique;
- The results are comparable in terms of accuracy to those obtain through other analytical techniques.
- It has a high scanning speed. In just a few seconds, it is possible to obtain all the necessary information for the entire frequency range.
- The infrared spectrometer has a very high resolution, especially the Fourier transform spectrometer (FTIR).

- It has a wide range of applications in both qualitative and quantitative analysis, so it can be used to analyse almost any organic compound.

Also, with NIR analysis the most useful features are contained in the overtones of the spectrum, or combinations of them, which are much more significant than the fundamental components found in the other two infrared ranges.
The absorbance obtained in the NIR region is very small, consequently I can consider that it increases as the concentration of the substance in the sample increases, and it is possible to evaluate it without having to dilute or add additional reagents in the tube.
NIRS is not a standalone technology, so for each substance and each component you want to analyze, you need to undergo the entire calibration process. This is all to ensure greater reliability of the predictions that will be made.
Despite this, through sophisticated chemometric procedures, we can obtain truly meaningful results for substance quantification and qualitative analysis.

1.4.2 Mid Infrared (MIR) Spectroscopy

Mid Infrared (MIR) Spectroscopy, operating within the wavelength range of approximately $2.5–15, \mu m$ (4000–650 cm^{-1}), serves as a pivotal analytical technique. In this spectral region, fundamental absorptions play a crucial role, especially in the study of organic compounds. The absorption bands in Mid-IR spectroscopy are intricately linked to the vibrational modes of specific functional groups. The precise positioning and intensity of these bands offer valuable insights into the energy of the bonds, their local environment, and their concentration within the matrix. Such characteristics make Mid Infrared spectroscopy an ideal choice for a wide range of applications, spanning both qualitative and quantitative analyses.

Instrumentation

MIR spectroscopy is a commonly used device in laboratory for quantitative and qualitative measurements of samples in all states of aggregation, and, additionally, it doesn't cause any damage to specimens.
The MIR spectral band ranges from 2500 nm to 25000 nm, allowing the study of the insight of the structure of molecules and enabling quantification of concentration.
Since the 1880s, it has been possible to record the MIR spectra of a number of simple organic compounds through the use of a bolometer as a detector. The analysis and investigation of the relationship between MIR spectra and molecular structures began in the 20^{th} century, but it

was not until after World War II that MIR spectroscopy rapidly developed. This was due to the introduction of new technologies such as highly sensitive thermocouples and high-quality prisms.

A further step forward was taken in the 1970s when Fourier Transform (FT) spectroscopy was introduced in the field of MIR, leading to a real revolution. This technology allows for high accuracy in determining the wavelength and the corresponding intensity peak, as well as showing high spectral resolution.

Major technological advancements in moder MIR spectroscopy are attributed to MIR light source technology. The most common MIR spectrometers use **thermal emitters**, which are infrared light sources that emit a broad spectrum of IR light from 1000 to 20000 nm depending upon the operating temperature of the device. Constructed from materials exhibiting high electrical resistivity, these devices proficiently transform electrical energy into thermal energy. The resulting radiation possesses distinctive characteristics influenced by factors such as the optical emissivity of the emitter material, the supplied electrical current, and the emitter's surface area and temperature. Achieving enhanced performance necessitates a substantial elevation in temperature.

The thermal emitter, a venerable light source, has experienced rapid evolution in the past century, particularly driven by advancements in detectors and methodological approaches. Recognized for its blackbody-like emission, the thermal source offers a spectral coverage well-suited for a wide range of MIR spectroscopy applications. Despite its historical significance, thermal emitters do come with limitations, notably their inadequate brightness, also known as spectral radiance. This limitation directly impacts the spectral power incident on the surface unit of the sample being analyzed. Consequently, the reduced path length that light must traverse between the source and the matter complicates the study of a larger number of molecules.

To overcome the major limitations of thermal emitters, lasers are introduced as an advanced, high-brightness light source. In particular, we have a brightness that is several orders of magnitude higher than that of the thermal emitter. The main problem is that lasers are almost monochromatic, so they need a wide adaptation to be used for MIR spectroscopy.

A pivotal advancement in the realm of Mid-Infrared (MIR) spectroscopy lies in the adoption of Quantum Cascade Lasers (QCL) as light sources. This innovation boasts a multitude of advantages, including robustness, operational stability at ambient temperatures, compact dimensions, and the unique capability of extending spectrum coverage. The latter is achieved by seamlessly integrating multiple chips to create a unified emitter, effectively spanning the entire MIR range. The widespread adoption of QCLs has revolutionized MIR spectroscopy systems, enabling highly sensitive measurements that were previously unattainable. This technology unlocks access to fundamental and potent absorption bands, marking a significant leap forward compared to other laser systems currently in use.

Nowadays, new technologies are being developed to increase the brightness of the source, the spectral resolution and accuracy. Among these we have those based on the use of frequency-comb laser in dual-comb and those based on the use of **supercontinuum laser sources**.

The latter are special laser that present a wide and continuous spectrum. This very broad spectrum is due to the presence of a process called **supercontinuum generation process** which converts laser light into light with a very wide spectral bandwidth, i.e. an extremely broad continuous optical spectrum. Spectral broadening is usually achieved by propagating light pulses through a strongly nonlinear device.[40]

MID supercontinuum laser sources represent a valid element that can fill the gap between QLCs and standard thermal emitters.

The fundamental components for creating a spectroscopy system include detectors. The main requirements for them are high quantum efficiency and ideally tailorable absorption band gaps.[23] Of the various types of detectors available, our attention focused on those made from lead selinide (PbSe).

PbSe detectors extend into the MIR, covering a wavelength range from 1 to 5 μm. It uses the photoconductive effect, which is the basis of how photoresistors work. These are radiation-sensitive transducers made of semiconductor materials, such as, in our case, lead selenide.

When the sensitive surface of the photoresistor is exposed to radiation, the absorbed energy causes covalent bonds to break and gap-electron pairs to increase. Consequently, exposure to infrared radiation causes the resistance in the active area to decrease as a function of radiation intensity.

Photoresistors cannot operate in photovoltaic mode as they do not generate a photocurrent themselves and, in addition, always require a bias voltage for detector operations. Another key feature is that they behave like a resistor and not a diode, so they have no p-n junction, no junction capacitance and no polarity.

PbSe detectors are subject to *pink noise*, which results from the presence of a series of charge carrier traps. This tends to decrease as the modulation frequency increases. This type of photoresistor can operate at a high modulation frequency due to a very low time constant of 4 μs. As a result, the associated pink noise can be attenuated. In general, PbSe detectors are also subject to 1/f noise, i.e. *flicker noise*.

In the construction of our MIR spectroscope system, the choice of a suitable detector is pivotal. After a thorough analysis of commercially available detectors, the lead selinide (PbSe) detectors emerged as viable options. Laser Components offers various models, including the PB45, PB50, and PB55 series detectors. These detectors serve diverse applications, spanning medical gas analysis (e.g., CO_2 measurement), industrial and automotive emission measurement, as well as moisture and hydrocarbon measurement.

Among the available options, the PB45 detector aligns most closely with our requirements based on the specifications outlined in the datasheet. With a spectral range of 1 to 4.7 μm, the PB45 operates at room temperature, featuring a 20$\%$ cut-off at 4.7 μm and an impressive detectivity (D*) exceeding 10^{10}. This makes it an optimal choice for our application, especially considering our need to operate at room temperature.

While the PB50 and PB55 series detectors are cooled versions designed for operation at low temperatures (-20°C to -35°C and -45°C to -55°C, respectively), offering enhanced performance and a wavelength shift to 5.2 μm, these advantages are not necessary for our specific application. Given that our spectroscope operates at room temperature, the PB45 stands out as the most suitable choice. Furthermore, its spectral range aligns well with the characteristic peaks of the substance of interest—lactate. The figure below 1.9 illustrates the PB45 series, showcasing its uncooled design.

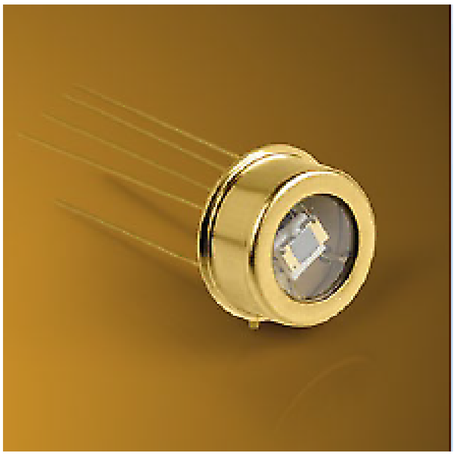


Figure 1.9: Uncooled Lead-Selinide detector PB45 Series

Applications

Mid-Infrared (MIR) spectroscopy, recognized for its diverse applications, plays a pivotal role in biomedical research. This spectroscopic technique monitors the interaction of functional groups in chemical molecules with infrared light, resulting in predictable vibrations that provide a unique ”fingerprint” characteristic of the chemical substances present in biological samples. In the mid-IR region, spectra exhibit well-resolved bands assigned to functional groups within biomolecules, enabling structural characterization. The intensity of these bands correlates with the concentration of specific biomolecular components, allowing for both qualitative and quantitative analyses.

In biomedical applications, mid-IR spectroscopy proves valuable for gaining insights into the composition of biological samples. The unique spectral profiles generated by biomolecules, such as lipids, proteins, and carbohydrates, enable the identification of specific functional groups in unknown substances. Comparison of mid-IR spectra to standard spectra in reference databases facilitates the identification of chemical compounds in complex biological samples.

Mid-IR spectroscopy in biomedical research benefits from multivariate statistical analysis techniques, commonly known as chemometrics, to extract meaningful information from spectra. These techniques are crucial for the classification and quantitative analysis of multiple components in biological samples. Instrument calibration is essential for ensuring accurate results, and chemometric techniques play a key role in this calibration process.

The application of mid-IR spectroscopy in biomedical research extends to various areas, including the analysis of biological tissues, fluids, and other complex matrices. The ability to simultaneously analyze multiple components makes mid-IR spectroscopy a powerful tool for researchers seeking comprehensive insights into the molecular composition of biological samples. This versatility positions mid-IR spectroscopy as a valuable asset in advancing our understanding of biomolecular interactions and contributing to advancements in medical research and diagnostics.

Advantages and Limitations

One of the primary strengths of mid-infrared (MIR) spectroscopy lies in its capacity to furnish intricate details regarding the molecular composition of a sample. Operating within the mid-infrared segment of the electromagnetic spectrum, this technique delves into the fundamental molecular vibrations of atoms within the sample, facilitating the precise identification of chemical bonds and molecular functionalities. Such precision renders MIR spectroscopy particularly advantageous in diverse fields, including chemistry, pharmaceuticals, and biological research.

Another noteworthy benefit of MIR spectroscopy is the expeditious nature of its analyses. The technique enables the swift acquisition of spectra, making it conducive to seamless integration into experimental protocols and the handling of a substantial volume of samples.

However, MIR spectroscopy is not without its limitations. One significant challenge is its restricted ability to penetrate opaque or excessively thick samples. The absorption or scattering of mid-infrared light by dense materials diminishes the efficacy of the technique in specific experimental scenarios.

In addition to the discussed limitations, one prominent drawback emphasized in the literature is the incapability of MIR spectroscopy to be utilized for concentration analyses of substances within the bloodstream. This is attributed to the fact that mid-infrared waves are unable to penetrate the skin, let alone the walls of blood vessels. As a result, the application of MIR spectroscopy is restricted when attempting to conduct analyses related to the concentration of substances circulating in the blood.

Additionally, the necessity to prepare samples in the form of powder or thin films for certain analyses poses operational challenges. This preparatory step demands additional time and may impact the representativeness of the sample compared to its real-world conditions.

In conclusion, despite these specific drawbacks, mid-infrared spectroscopy stands as a valuable and advanced analytical technique for molecular analysis, leveraging its strengths in delivering precise chemical information and facilitating rapid analyses.

1.5 Conclusion

In conclusion, this first chapter has provided a comprehensive introduction to the fundamental principles and applications of spectroscopy, spanning the domains of UV/Visible, Near Infrared (NIR), and Mid Infrared (MIR). We embarked on this journey by laying the groundwork for understanding the physical principles that govern these spectroscopic techniques.

The exploration of UV/Visible spectroscopy elucidated its spectroscopic principles, common methodologies, instrumental aspects, and the inherent advantages and limitations. This foundational knowledge sets the stage for the subsequent chapters, where we will delve deeper into the specific applications and advanced techniques within UV/Visible spectroscopy.

Transitioning into the infrared spectrum, the chapter unfolded the unique capabilities of both Near Infrared (NIR) and Mid Infrared (MIR) spectroscopies. The versatility of NIR spectroscopy, especially in biomedical applications, and the molecular insights provided by MIR spectroscopy were highlighted. These insights will serve as a springboard for our detailed investigations in the following chapters.

Chapter 2

Substances Detectable in Blood

The spectroscopic analysis of substances present in blood plays a crucial role in our research aimed at lactate detection. This chapter explores the fundamental interactions between light and key blood components, focusing specifically on carbon dioxide, oxygen, and hemoglobin. The choice to investigate the behavior of these substances is motivated by their physiological relevance and their impact on spectroscopy in the infrared range, which has proven particularly promising for our purposes. Through a detailed analysis of the spectral responses of each substance, we aim to understand their peculiarities and interconnections, laying the groundwork for an accurate and specific assessment of lactate in the context of blood.

2.1 Overview of Analyzable Substances

In the complex network of the human circulatory system, the composition of blood serves as a profound reflection of the body's physiological intricacies. This chapter embarks on an exploration of three pivotal substances coursing through our veins – Carbon Dioxide (CO_2), Oxygen, and Hemoglobin.
Beyond their status as mere components, these substances embody critical indicators of respiratory function, metabolic activity, and oxygen transport, casting a unique light on the body's overall health. However, our journey doesn't stop at mere understanding; it extends into the realm of cutting-edge instrumentation that fuels the precise detection of these substances in blood. The subsequent sections meticulously dissect the state of the art in instrumentation, unraveling the methodologies that underpin the detection of CO_2, oxygen, and Hemoglobin.
Characteristic absorptions of functional groups are cataloged in tables known as **correlation tables**. These tables furnish a list of characteristic infrared absorptions for different types of

bonds and functional groups. The absorption intensity is categorized into three levels: strong (s), medium (m), and weak (w). The infrared spectrum comprises two regions:

- The functional groups of the molecule are found in the region between 2500 nm and 6667 nm of the IR spectrum.
- The fingerprint region, above 6667nm, contains absorptions of complicated vibrations, unique to each molecule. A database with recorded infrared spectra of known organic molecules facilitates the comparison of spectra produced for unknown compounds.

In our study, we focus on the area of Extracorporeal Circulation, leveraging detailed documentation on the company's technology park. This documentation outlines measured parameters, their ranges, and the methods employed, particularly those measured with infrared technology: saturation, hemoglobin, and temperature. Of additional clinical interest are parameters like *lactate concentration*, which our research seeks to identify and detect using suitable technology. Moreover, our investigation deepens in the forthcoming chapter, centering specifically on lactate—an element of paramount significance serving as the central theme of this master thesis. The subsequent section delves into the complexities of lactate, unveiling its importance in physiological processes and examining specialized techniques devoted to its accurate detection. This cohesive journey ensures a comprehensive grasp of both analyzable substances and the instrumental advancements propelling our comprehension of the circulatory system.

2.1.1 Carbon Dioxide (CO_2)

Metabolic reactions, such as cellular respiration, take place within the cells that constitute the tissues of internal organs. This intricate process involves the combustion of nutrients obtained from digestion, breaking them down into simple molecules to yield energy available to the cell in the form of ATP—the primary energy currency of cells. From these vital reactions, essential for the correct functioning of the organism, carbon dioxide is produced and must be efficiently eliminated from the body.

At the muscular level, the production of carbon dioxide is not only very high and variable but also requires constant elimination through pulmonary ventilation. This respiratory process is vital for efficient gas exchange, ensuring the removal of carbon dioxide and replenishment of oxygen.

Once produced by cellular metabolism, CO_2 rapidly diffuses from the cells into the arterial blood, which perfuses the tissues. Notably, carbon dioxide lacks a specific carrier and is transported in the form of dissolved gas, carbamino compounds, and bicarbonate ions in both plasma and red blood cells. These three modes of transport are reversible and intricately balanced, although they do not handle the same quantities.

The regulation of carbon dioxide levels is crucial for maintaining homeostasis within the body. Disruptions in this balance can have clinical implications and may be associated with certain medical conditions. Therefore, understanding the dynamic processes involved in carbon dioxide production, transport, and elimination is not only fundamental to respiratory physiology but also holds significance in the broader context of maintaining overall health and well-being.

State of the Art in CO_2 Detection Techniques

The state of the art in technologies for detecting the concentration of carbon dioxide (CO_2) in blood is predominantly characterized by the use of optical sensors, with a prevalent focus on infrared-based methodologies. These techniques leverage Lambert Beer's law to calculate analyte concentrations, with incident light falling within the mid-infrared (MIR) range, typically between 3000 and $8000nm$. Notably, the primary absorption peak for carbon dioxide occurs at $4250nm$.

The instrumental setup for these detection systems includes an **emitter**, available in two configurations: one emitting broadband light, necessitating additional components like a filter for radiation selection, and the other utilizing a Light Emitting Diode (LED) with a narrow spectrum. The **photodetector** , typically made with an epitaxial layer of InAsPb on an InAs substrate, completes the system. The coupling of the LED and photodetector eliminates the need for mechanical modulators and interference filters, resulting in a less complex sensor design.

In the biomedical field, two configurations of optical sensors are prominent for CO_2 detection: Main-Stream and Side-Stream approaches. The Main-Stream approach directly measures CO_2 concentration in the gas flow duct, while the Side-Stream approach diverts the gas flow into a specialized chamber for measurement.

A noteworthy advancement in this domain is the sensor developed in 2020, as detailed in the article *"Development of CO_2 Sensor for Extracorporeal Life Support Application"* created by Bellancini et al. is also based on a similar principle.

This Main-Stream type sensor adheres to European legislation's accuracy specifications and is designed to measure carbon dioxide concentration in the exhaust gases of an oxygenator membrane during extracorporeal procedures.

The sensor comprises two main portions: one measuring the gas flow applied to the mem-

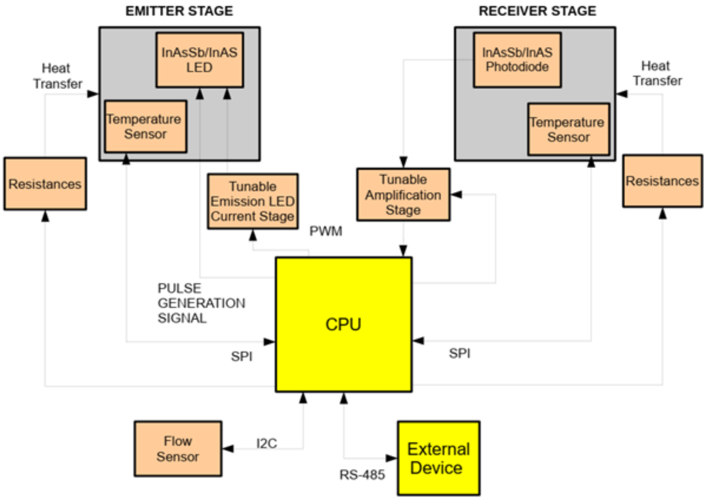


Figure 2.1: Schematic representation of the sensor [5]

brane and the other quantifying the concentration of carbon dioxide extracted from the patient's body. To mitigate the condensation of water vapor and preserve data quality, a heating module is incorporated. Despite the known influence of temperature on sensor operation, an algorithm is employed to establish a relationship between signal and temperature, enhancing sensor sensitivity.

The sensor comprises two main portions: one measuring the gas flow applied to the membrane and the other quantifying the concentration of carbon dioxide extracted from the patient's body. To mitigate the condensation of water vapor and preserve data quality, a heating module is incorporated. Despite the known influence of temperature on sensor operation, an algorithm is employed to establish a relationship between signal and temperature, enhancing sensor sensitivity.

Results obtained from this sensor align with European regulations, particularly ISO standards, making it suitable for the examined clinical applications. The schematic representation of the sensor is depicted in 2.1, showcasing its design and functionality.

2.1.2 Oxygen and Hemoglobin

Oxygen plays a crucial role in our bodies, being consumed along with glucose in cells for the oxidation process that generates the energy necessary for cellular functioning. The metabolic process shares similarities with combustion, resulting in waste products such as carbon dioxide.

Through the respiratory system, we ensure a continuous and proper supply of oxygen, simultaneously expelling carbon dioxide through exhalation. To reach cells, oxygen is transported within the bloodstream by erythrocytes, or red blood cells. These cells uptake oxygen in the pulmonary alveoli, transport it to various anatomical districts, exchange it for carbon dioxide, and return to the heart and lungs through the veins for the release of carbon dioxide. Red blood cells contain hemoglobin, a globular protein capable of transporting four oxygen molecules at once. Oxygen saturation is a vital blood index, expressing the percentage of saturated hemoglobin compared to the total hemoglobin. This parameter allows the detection of hypoxemic conditions. When an oxygen molecule binds to one of hemoglobin's four binding sites, the affinity of the remaining three binding sites significantly increases, generally causing oxygen to preferentially bind to hemoglobin with at least one site already occupied.

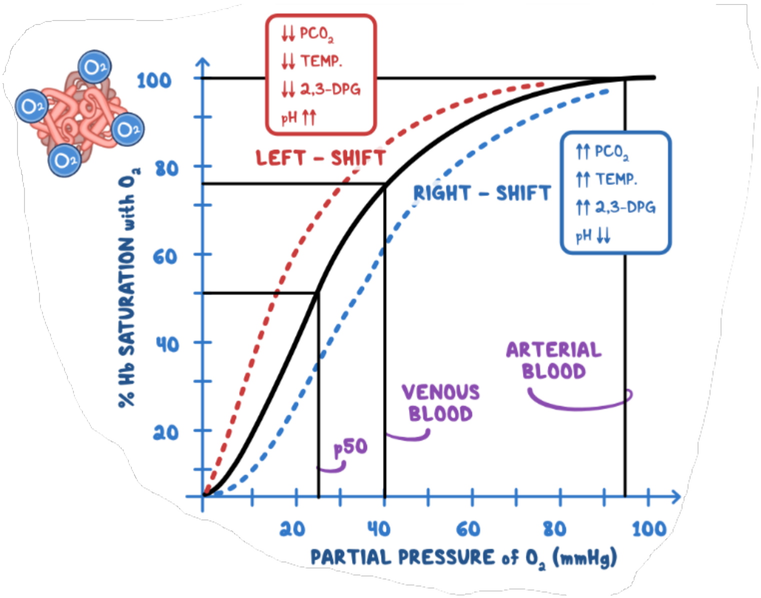


Figure 2.2: Oxygen Dissociative Curve

This leads to the creation of the Oxygen Dissociative Curve (Figure 2.2), which exhibits a sigmoidal pattern. The curve illustrates the intricate relationship between hemoglobin saturation with oxygen (SO_2) and the partial pressure of oxygen (PO_2). Its sigmoidal shape signifies positive cooperativity in the binding of oxygen molecules to hemoglobin's heme group, allowing successive binding until a saturation plateau is reached.

Key features of the curve include the Bohr effect, where factors like PCO_2 and pH influence its position. A left shift, indicating higher oxygen affinity, occurs under conditions such as decreased PCO_2, increased pH, and lowered temperature. Conversely, a right shift, reflecting lower oxygen affinity, is associated with increased PCO_2, decreased pH, and elevated temperature. These shifts play a pivotal role in oxygen unloading in metabolically active tissues, highlighting the adaptive nature of hemoglobin.

Additionally, the influence of 2,3-diphosphoglycerate (2,3-DPG) and carbon monoxide (CO) on the curve is noteworthy. Under hypoxic conditions, an increase in 2,3-DPG concentration promotes oxygen unloading, while CO binding causes a left shift, indicating higher oxygen affinity but reducing the blood's absolute oxygen-carrying capacity. The oxygen-hemoglobin dissociation curve proves fundamental in comprehending the intricacies of oxygen transport in blood, emphasizing hemoglobin's adaptability under diverse physiological and pathological conditions. This adaptability results in faster loading of oxygen molecules in environments with high concentrations, such as lung alveoli, and more readily dissociates in deficient environments, like in metabolically active tissues.

Finally, the structure of hemoglobin consists of 4 polypeptide subunits, namely 2 α chains and 2 β chains, and on each of them is a heme group, which contains an iron ion responsible for interaction with oxygen. This bond is a reversible bond, since the interaction is very weak. They exist in two distinct configurations:

- Taut (T): deoxygenated, which is a configuration with low affinity for oxygen that promotes the release of those present;
- Relaxed (R): oxygenated, which is a configuration that has high affinity for oxygen and promotes binding to other oxygen molecules if there are binding sites still available.

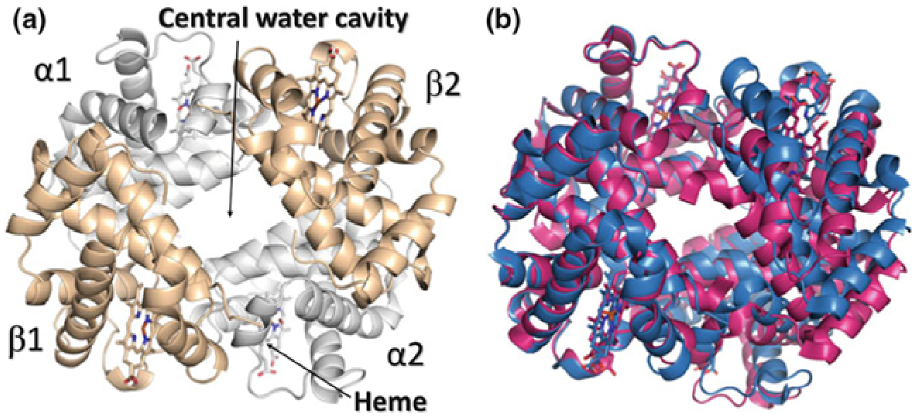


Figure 2.3: **a)** Quaternary structure of Hemoglobin. **b)** Structure of oxygenated Hemoglobin (magenta) superimposed on the structure of deoxygenated Hemoglobin (blue). [1]

These two configurations have the characteristic of exhibiting different electromagnetic absorption and consequently different light emission.

State of the Art in Detection Techniques

Pulse Oximeter

The measurement of saturation is based precisely on this characteristic, and to measure it, one uses the oximeter or pulse oximeter, which is a transcutaneous electromedical device.
There are two possible configurations: *transmittance device* and *reflectance device*. The latter are little used especially in clinical settings because they show low reliability, while transmittance devices are the most widely used.
The transmittance device configuration consists of an electronic processor and a pair of light-emitting diodes (LEDs) placed in front of a photodiode. Between these two components is the translucent part of the patient's body, which is usually a finger or earlobe. One LED is red with a wavelength of 660 nm and the other is infrared with a wavelength of 940 nm. The need for two different LEDs is related to the absorption properties of the two different configurations of hemoglobin, since:

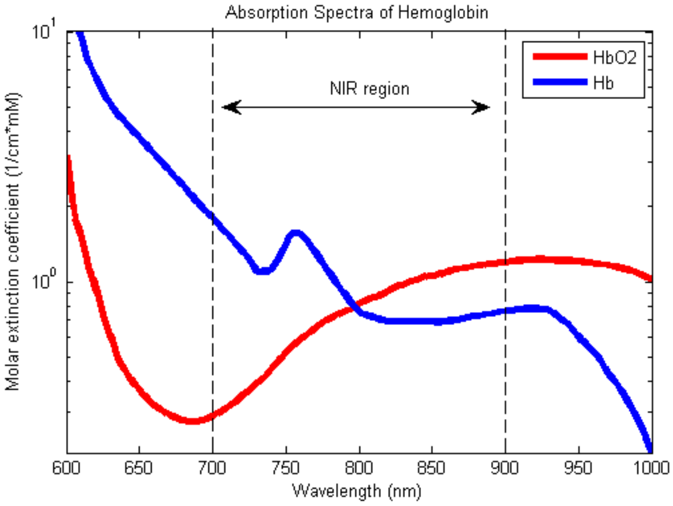


Figure 2.4: Absorption spectrum of the two configurations of hemoglobin

- Oxygenated hemoglobin absorbs more infrared light and allows more red light to pass through;

- Deoxygenated hemoglobin absorbs more red light and allows more infrared light to pass through.

LEDs are not always on, but rather follow their own on-off cycle. The amount of light that is not absorbed is measured. These signals tend to fluctuate because of the very nature of arterial blood flow.
All this happens because blood is fed at high velocity into the aorta and pulmonary artery only for a short interval of time, namely that which coincides with the opening of the semilunar valves during ventricular systole; therefore, the input is intermittent and, as a result, the flow is accelerated in the first part of the cardiac cycle and decelerated in the last phase.
At first glance, the intermittency might cast some doubt on the effective blood supply to all internal organs, however, the high pressure to which the arterial walls are subjected stretches the elastic fibers of the blood vessel, accumulating elastic energy that will be returned by compressing the blood and ensuring flow.

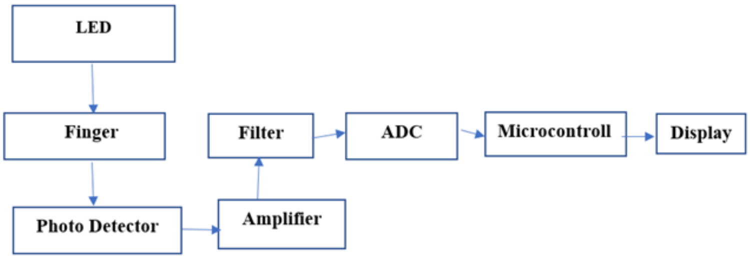


Figure 2.5: Block diagram of the Pulse Oximeter[26]

To calculate saturation, the processor will evaluate the ratio of red light measurement to infrared light measurement, that is, it measures the ratio of oxygenated to deoxygenated hemoglobin. The ratio will then be converted to saturation using a table that is based on Lambert-Beer's law.
In recent years, the use of this device has increased exponentially due to the COVID-19 pandemic.

Optical sensor

The concentration of haemoglobin in blood is another very important parameter that is currently measured using various methods. The most common non-invasive methods are based on using spectrophotometry to analyse the absorption of light and subsequently calculate the concentration of the molecule by applying the Lambert-Beer law. One such system is proposed in the article *'Optical Sensor System for Hemoglobin Measurement'* by Doshi et al [17]. The system is based on the pulse photometric measurement method, the principle behind which is very similar to that of the pulse oximeter.

The device consists of a portion, which is placed on the tip of the finger, that emits infrared signals through an LED, which pass through the skin and blood vessel walls. As with the pulse oximeter, it is considered that oxygenated and deoxygenated haemoglobin absorb different wavelengths (960nm and 660nm). The transmitted light in the area of interest is measured through the use of a transimpedence amplifier photodiode. The output of this detector increases linearly with increasing light intensity. The signal obtained is normalised and the ratio between the pulsating and non-pulsating component of the red and IR signal is calculated to determine the haemoglobin concentration using the Lambert-Beer law.

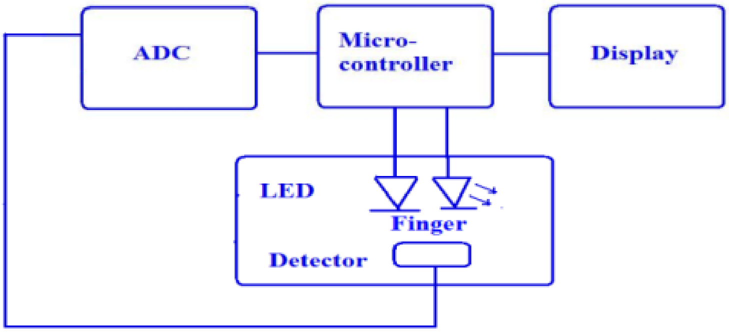


Figure 2.6: Block diagram of hemoblobin sensor [17]

Signal acquisition by this method is totally non-invasive. The sensors assembled in this research are fully integrated in wearable finger clips. It is a continuous process that may, however, show some fluctuations related to the heartbeat and the configuration of the blood circulation.

Chapter 3

In-Depth Examination of Lactate

This chapter initiates a comprehensive exploration of lactate, delving into its chemical structure and fundamental physiological properties. As a monocarboxylic acid, lactate holds a pivotal position in cellular metabolism. Its chiral carbon center introduces nuances that significantly impact its behavior within biological systems. Understanding these structural intricacies lays the groundwork for unraveling lactate's diverse roles in vital biological processes. At its core, lactate plays a central role in energy metabolism, actively participating in glycolysis and the Cori cycle. This chapter navigates through the involvement of lactate in these fundamental pathways, shedding light on its contributions to cellular respiration. Beyond its metabolic functions, the examination extends to the physiological effects of lactate on tissues and cells. It serves as a signaling molecule, influencing cellular pH and redox balance, thereby affecting broader cellular functions.

The metabolic interplay of lactate unfolds further, emphasizing its interactions with other key components in cellular pathways. This exploration sets the stage for subsequent discussions on lactate's involvement in gluconeogenesis and its role in maintaining metabolic equilibrium. The chapter concludes by underscoring the significance of accurate lactate measurement, positioning it as a valuable biomarker in assessing cellular stress and overall metabolic health.

3.1 Chemical Structure of Lactate

3.1.1 Molecular Composition

Lactate, also known as lactic acid, exhibits a molecular composition represented by the chemical formula $C_3H_6O_3$. This formula signifies the presence of three carbon atoms, six hydrogen

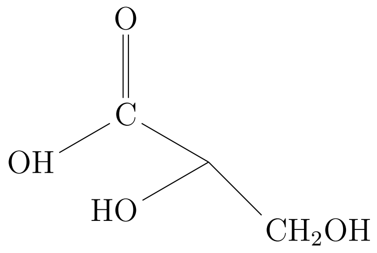


Figure 3.1: Structural formula of L-lactate

atoms, and three oxygen atoms in each molecule of lactate. It is classified as a carboxylic acid due to its carboxyl group, which consists of a carbon atom double-bonded to an oxygen atom and single-bonded to a hydroxyl group.

3.1.2 Isomeric Configuration

Lactate exists in two primary stereoisomeric configurations: L-lactate and D-lactate. The prefix ”L” denotes the levorotatory or left-handed form, while ”D” represents the dextrorotatory or right-handed form. In biological systems, L-lactate is the predominant isomer, and it plays a crucial role in various metabolic processes, particularly during anaerobic metabolism.

3.1.3 Structural and Functional Properties

The structural formula of L-lactate, $\mathrm{CH_3CHOHCOOH}$, intricately portrays its carboxylic acid and hydroxyl functional groups. This configuration underscores the compound's significance in biological systems, particularly within muscle cells during anaerobic metabolism. Lactate assumes a central role in essential metabolic pathways such as the Cori cycle, gluconeogenesis, and energy metabolism. Functionally, lactate serves as a crucial intermediate in the conversion of pyruvate during anaerobic respiration, functioning both as a metabolic byproduct and an energy substrate. Its dynamic equilibrium with pyruvate is indispensable for maintaining cellular homeostasis and contributing to overall energy balance.

Comprehending the molecular composition, isomeric configuration, and structural and functional properties of lactate is imperative for unraveling its intricate roles in various physiological processes, offering valuable insights into its implications in health and disease. The accompanying structural formula visually represents the molecular arrangement, enhancing our understanding of L-lactate's pivotal contributions to cellular metabolism and homeostasis.

3.2 Physiological Properties and Significance

3.2.1 Role of Lactate in Biological Processes

The role of lactate in biological processes is multifaceted, particularly under conditions of oxygen deficiency, high-intensity exercise, or recruitment of non-oxidative muscle fibers. In these scenarios, a substantial and continuous accumulation of lactic acid is observed both in the muscles engaged in exercise and in the bloodstream.

Clinically, lactate serves as a crucial prognostic indicator due to its proportional relationship with the presence of oxygen in tissues. Lactic acid, with a chemical formula of $C_3H_6O_3$, is a chiral molecule with two enantiomers, and the L-(+)-lactic acid is predominant in metabolic cycles. At physiological pH, lactic acid readily dissociates into H^+ and lactate ions. The figure 3.2 illustrates the relationship between lactic acid and lactate, emphasizing that lactate is the conjugate base of lactic acid. Lactic acid plays a vital role in human energy processes, particularly within the lactacid anaerobic system. This system is activated during activities requiring strength and endurance for around one minute.

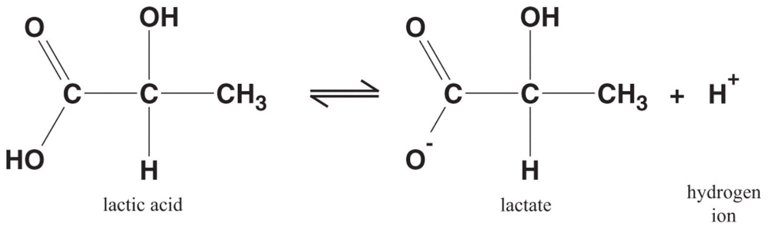


Figure 3.2: Lactate is the conjugate base of Lactic acid

In this system, ATP is produced anaerobically from glycogen, leading to the production of pyruvic acid, which is further converted into lactic acid. During prolonged exertion, lactate accumulates, lowering the pH and causing acidosis. The body eliminates lactate through various mechanisms, including conversion to glycogen, conversion to protein, and oxidation to carbon dioxide and water.

In a clinical context, lactate concentrations provide valuable insights. Under resting conditions, a healthy subject typically ranges between 0.5 and 1.5 mmol/L. During physical activity, this range extends to 12-25 mmol/L.

In critical care, an increase in lactate concentration is a symptom of an imbalance in lactate production and removal. A lactate concentration of less than 2 mmol/L is considered normal and should remain around a value of 0.7 mmol/L, whereas in a healthy individual, under physical exertion, the range shifts between 11 and 25 mmol/L. In the case of an individual registering a value above 4 mmol/L at rest, this could be a red flag, as it could underlie health problems such as myocardial infarction, heart attack and collapse of the blood circulation. Understanding these measures is crucial for diagnosis, and they should be carried out with robust and very rapid methods.

There are three distinct ways to eliminate lactic acid from our bodies:

- Through sweat and urine;
- Conversion to glycogen through the Cori cycle, which converts lactic acid to glycogen or glucose in the liver and kidneys, or to pure glycogen in muscles;
- Conversion to protein;
- Oxidation to carbon dioxide and water.

3.2.2 Interaction with Energetic Metabolism

Lactic acid plays a pivotal role in human energy processes, particularly within the lactacid anaerobic system. This system is activated under conditions of oxygen deficiency, sustained high-intensity exercise, or recruitment of non-oxidative muscle fibers. During activities demanding strength and endurance for approximately one minute, the body relies on the lactacid anaerobic system.

In this system, ATP is produced anaerobically from glycogen, stored in skeletal muscle and the liver. The subsequent hydrolysis of glycogen to glucose facilitates intense muscle activity but for a limited period. The process of ATP production is oxygen-independent, leading to the production of pyruvic acid, which is further converted into lactic acid.

Lactate, the conjugate base of lactic acid, becomes a central player in energy metabolism, contributing to essential metabolic pathways such as the Cori cycle, gluconeogenesis, and overall energy balance. Its dynamic equilibrium with pyruvate is crucial for cellular homeostasis.

During prolonged exertion, lactate accumulates in muscles and blood, influencing the pH and causing acidosis. The body employs various mechanisms, such as conversion to glycogen, conversion to protein, and oxidation to carbon dioxide and water, to eliminate lactate.

3.2.3 Physiological Effects in Tissues and Cells

Lactate emerges as a multifaceted player, extending its influence beyond the conventional perception as a metabolic byproduct. This chapter delves into the profound physiological effects of lactate on tissues and cells, illuminating its dynamic role in cellular function and adaptation. Lactate, often relegated to the status of a metabolic waste, reveals itself as a vital metabolic signaling molecule, orchestrating adaptive responses to shifts in energy demands, oxygen availability, and metabolic states. Beyond its conventional role, lactate serves as an alternative energy substrate during intense cellular activities, contributing to the delicate balance of cellular energy homeostasis. The nuanced contributions of lactate encompass pH regulation, redox balance maintenance, and intricate cellular signaling, impacting gene expression, proliferation, and differentiation.

Furthermore, recent insights indicate that lactate plays a role in immunomodulation, influencing inflammatory responses and the function of immune cells. Cells and tissues demonstrate adaptive responses to varying lactate levels, leading to changes in gene expression and cellular processes. The intricate physiological effects of lactate are becoming increasingly apparent, revealing its crucial involvement in cellular homeostasis, metabolic adaptation, and the nuanced orchestration of cellular responses to diverse environmental stimuli.

This aligns with findings from the article *”Lactate metabolism in human health and disease”* by Li et al. (2022), published in Signal Transduction and Targeted Therapy. The article likely delves into the comprehensive understanding of lactate metabolism, providing valuable insights into its multifaceted roles, particularly in the context of inflammatory responses and overall human health. The inflammatory response, as discussed in the article, encompasses various acute and chronic diseases affecting almost all organs. In addition to participating in inflammatory injury and immune energy metabolism, lactate accumulation triggers the activation of cellular signaling pathways regulating inflammatory progression and tumor immune tolerance. Importantly, these regulatory effects are distinct from lactate’s ability to acidify the cellular environment. While acute inflammation is considered a host defense mechanism, unrestrained activation can lead to tissue necrosis and prolonged disease. Recent studies have affirmed that lactate exerts an inhibitory effect on acute inflammation, as illustrated in the figure 3.3.

3.3 Metabolic Cycles Associated with Lactate

Metabolic cycles associated with lactate encompass a paradigm shift from the conventional belief that the L-enantiomer of the lactate anion, once thought to be linked to oxygen deficiency in contracting skeletal muscle, is now recognized to form under fully aerobic conditions. This utilization of L-lactate persists across various cells, tissues, organs, and at the

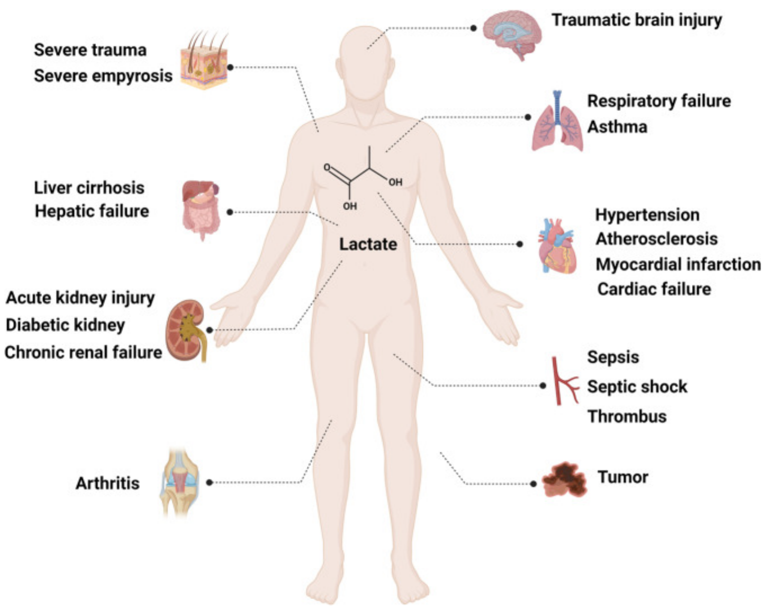


Figure 3.3: Lactate is implicated in the pathogenesis of diverse diseases, exerting regulatory effects on the cardiovascular system, respiratory system, digestive system, urinary system, and various other health conditions. Its significance extends to clinical applications, where lactate serves a pivotal role in the diagnosis and prognosis of different diseases. [28]

whole-body level, while the atypical D-enantiomer in mammalian metabolism carries documented adverse effects. Serving as both an inevitable byproduct and a substrate for mitochondrial respiration in mammalian systems, L-lactate establishes a vital link between glycolytic and aerobic pathways, challenging previous notions of lactate as a mere metabolic waste product and contributor to fatigue.
Instead, lactate emerges as the primary messenger in a sophisticated feedback loop system, as proposed by the Lactate Shuttle Hypothesis. This hypothesis posits that the connection between cells generating lactate and those utilizing or signaling with lactate can transcend compartmental barriers, occurring within and between cells, tissues, and organs. Challenges to adenosine triphosphate (ATP) supply prompt lactate production, initiating immediate, short-term, and long-term cellular adaptations to maintain ATP homeostasis. While recent reviews have covered the physiology and biochemistry of this topic, it is essential to consider newly emerging information, particularly regarding the role of lactate shuttling in metabolic signaling, to fully grasp the intricacies of metabolic cycles associated with lactate.

In various mammalian model systems, including humans, lactate metabolism plays three crucial roles at the whole-body level. Firstly, it serves as a major energy source. Secondly, lactate

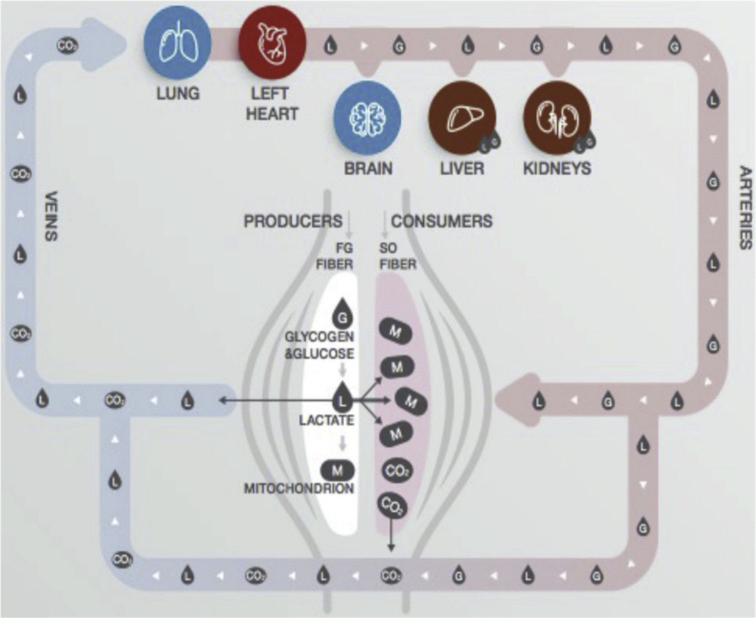


Figure 3.4: Illustration of the Lactate Shuttle, portraying the various functions of lactate in transporting oxidative and gluconeogenic substrates, along with its involvement in cellular signaling. [8]

is a primary precursor for gluconeogenesis. Thirdly, lactate acts as a signaling molecule, functioning in autocrine, paracrine, and endocrine-like ways, earning the designation of a ”lactormone.” Lactate exchanges within and among cells are categorized as ”Intracellular” and ”Cell-Cell” lactate shuttles, highlighting its roles in delivering oxidative and gluconeogenic substrates and serving as a signaling molecule. Examples include exchanges between cytosol and mitochondria and between cytosol and peroxisomes. Cell-Cell Lactate Shuttles involve exchanges between different muscle fibers, as well as between working skeletal muscle and various organs such as the heart, brain, liver, and kidneys, and between astrocytes and neurons. These shuttles are often driven by concentration or pH gradients, or redox states. Importantly, various body compartments and systems, such as the interstitial space, vasculature, and circulation, contribute to lactate shuttling in vivo. [8]

3.3.1 Glycolysis

The glycolysis represents the primary metabolic pathway through which most organisms break down glucose molecules to produce energy. This process was fundamental for early life on Earth as it occurred in the absence of oxygen, an element not available during that time. Glycolysis, indeed, does not require oxygen and plays a crucial role in anaerobic metabolic processes.

From a chemical perspective, glycolysis is a partial oxidation of glucose occurring in the cell's cytoplasm. Starting with six carbon atoms, two molecules of pyruvic acid, characterized by three carbon atoms, are obtained. Compounds such as adenosine triphosphate (ATP), adenosine diphosphate (ADP), and adenosine monophosphate (AMP) are involved in the process, responsible for transferring and storing phosphate groups. Coenzymes like nicotinamide adenine dinucleotide ($NAD^+/NADH$) and flavin adenine dinucleotide ($FAD/FADH_2$) participate in the redox reactions of metabolic intermediates. Additionally, coenzyme A is responsible for storing and transferring acetyl groups.

Glycolysis takes place in the cytoplasm of all cells, but its energy yield is limited, representing only 6% of the energy a cell can derive from a glucose molecule. Under aerobic conditions, the pyruvate and NADH generated during glycolysis are transferred to the mitochondria, where cellular respiration is completed with a significant energy gain.

However, under anaerobic conditions, some cells and unicellular organisms convert glycolysis products through a fermentation process, as additional ATP molecules are not generated. Glycolysis proceeds through 10 reactions catalyzed by specific enzymes, producing metabolic intermediates in the form of a phosphoryl ester, organic molecules preventing the products from crossing the cell membrane due to the negative charge provided by the phosphate group. The main oxidizing agent is NAD^+, and for each glucose molecule oxidized to pyruvate, two ATP molecules are produced.

$$\text{Glucose} + 2\ NAD^+ + 2\ \text{ADP} + 2\ HPO_4^{-2} \xrightarrow{\text{Glycolysis}} 2\ \text{Pyruvate} + 2\ \text{ATP} + 2\ \text{NADH} + 2\ \text{H}_2\text{O}$$

The resulting pyruvate does not accumulate but undergoes one of three possible enzyme-catalyzed reactions, depending on oxygenation status and cell type. These reactions include reduction to lactate (lactic acid fermentation), reduction to ethanol (alcoholic fermentation), and oxidation and decarboxylation to Acetyl-CoA.

A key to understanding the biochemical logic of two of the three possible fates of pyruvate is to consider that it is produced from glucose oxidation through glycolysis. As glycolysis constantly requires NADH replenishment, two metabolic pathways utilize pyruvate to regenerate NAD+ under anaerobic conditions.

Lactic acid fermentation, of particular interest in research, is the main pathway for NAD+ regeneration under anaerobic conditions and is catalyzed by the enzyme lactate dehydrogenase.

$$\text{Pyruvate} + \text{NADH} + H_3O^+ + 2\ HPO_4^{-2} \xrightleftharpoons[\text{Lactate dehydrogenase}]{} \text{Lactate} + NAD^+ + H_2O$$

Under conditions of oxygen deficiency in muscles or during intense physical activity, there is a substantial and continuous accumulation of lactate in muscle tissues and bloodstream. Exhaustion occurs when the blood lactate concentration reaches critical levels, around 8-15 mmol/L. During the dissociation of lactic acid into lactate, the ion H^+ is formed, managed by the body's buffering systems such as hemoglobin and bicarbonate. These responses, however, lead to physiological consequences, such as shifting the oxygen-hemoglobin dissociation curve and influencing lung gas exchange due to bicarbonate buffering of the H^+ ion.

It is essential to note that these buffering systems have limits, both temporally and quantitatively, and persisting under these conditions can lead to a decrease in pH in muscle tissues and blood. When muscle pH reaches acidic levels, the activity of key glycolysis enzymes is inhibited, providing a self-protective mechanism against excessive acidification of body fluids. Under physiological conditions, muscle pH is around 7.05, and during intense physical activity, it can decrease to 6.5.

3.3.2 Cori Cycle

Lactate, produced in the muscles through anaerobic glycolysis during physical exertion beyond the anaerobic threshold, plays a crucial role in the recovery process. During this phase, lactate diffuses from the muscles into the bloodstream and is transported to the liver.

The key enzyme involved in this process is lactate dehydrogenase, which catalyzes the conversion of lactate to pyruvate in the liver. Subsequently, pyruvate is converted into glucose through hepatic gluconeogenesis. This process enables the liver to generate new glucose from the lactate derived from the muscles.

The glucose produced in the liver can be released into the bloodstream through the action of the hepatic enzyme known as glucose-6-phosphatase. Once in the bloodstream, glucose can be recaptured by the muscles, where it is used to replenish glycogen stores. The ability to convert lactate into glucose in the liver is crucial for maintaining a continuous supply of glucose during and after intense physical exercise.

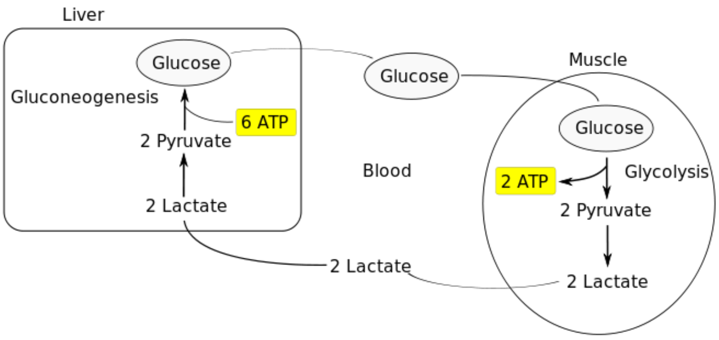


Figure 3.5: Cori Cycle schematic representation

3.4 Importance of Lactate Measurement

The measurement of L-lactate levels in blood is a crucial parameter in various fields, providing valuable insights into physiological and pathological processes. Normally ranging from 0.5 to 2.2 mmol/L, L-lactate levels can significantly increase, reaching 12-25 mmol/L during intense physical activity. However, the body efficiently restores these elevated levels to normal within 5-10 minutes through hepatic metabolism and lactate-to-pyruvate conversion. Prolonged hyperlactatemia, indicative of tissue lactate hyperproduction or utilization system pathology, is associated with a negative prognosis for patient outcomes. In intensive care units, dynamic measurements of lactate levels serve as crucial indicators, assessing the severity of patient conditions and predicting the likelihood of shock, collapse, and mortality.

Beyond medical applications, lactate measurements play a pivotal role in diverse industries. In the dairy sector, monitoring lactic fermentation processes and assessing the quality of finished products are common practices. Distinguishing between D- and L-lactate aids in identifying fermentation processes in food products, with D-lactate acting as an indicator of bacterial contamination in packaged meat, fish, and fruit juices. Additionally, D-lactate detection is relevant in dentistry for assessing bacterial contributions to carious cavities. In the wine industry, lactate monitoring informs the evaluation of malate-lactate fermentation processes. Several biosensors have been developed for lactate determination, contributing to the assessment of water contamination levels in silo sewage. This multifaceted importance of lactate measurement underscores its significance across diverse sectors. [27]

3.4.1 State of Art of the technology

The biggest problem we have in measuring lactate concentration at present is that we do not have noninvasive devices available that allow continuous measurement in clinical practice. The state of the art consists exclusively of devices that can perform intermittent measurements through the Arterial Blood Gas Analyser (ABG), which is a gold standard for these measurements. This method of analysis is also very expensive, invasive, as they require a large amount of blood (100-200 μL), and complex to perform, so much so that skilled nurses are required. Consequently, there is a need for studies and research to find a possible solution to this unmet clinical need.

Research in recent years has moved more toward ex vivo and in vivo electrochemical sensing systems. The ex vivo monitoring systems are used in conjunction with subcutaneous or ultrafiltration probes, while the in vivo ones involve subcutaneous implantation and directly measure blood lactate concentration. The problem with this category of biosensors is essentially that they are invasive, since they are implanted, do not allow for reproducible measurements and require sample preparation.

Some studies conducted in recent years are trying to create a wearable, noninvasive device to go and detect lactate in sweat, which is one of the means by which the body eliminates the excess metabolite. In the article *" Wearable Sensor System for Detection of Lactate in Sweat"* [15] a demonstration case is reported in which an organic electrochemical transistor is used, which, however, shows a limited range of sensitivity at concentrations below 1mM.

As a result, research has been directed toward new horizons, specifically, the use of spectroscopy in conjunction with multivariate analysis. In particular, we want to focus on the prospects that the use of NIRS in the determination of lactate concentration opens up.

As early as the late 1990s, several studies had shown that NIRS could detect lactate in human plasma and amniotic fluid. The problem is that if our goal is to detect lactate in blood, we must also consider the effects erythrocytes may have on absorption and scattering. One of the most recent approaches proposed is that of 2020, reported in the article *'Identification and Quantitative Determination of Lactate Using Optical Spectroscopy-Towards a Noninvasive Tool for Early Recognition of Sepsis'* by Budidha et al. [10]. The experiment proposes, first of all, to explore the optical properties of lactate in a solute very similar in constitution and properties to blood, but with a smaller amount of adsorbents and a wider physiological range, i.e. between 0 and 20 mmol/L. In this way, we are able to try to isolate the variations in the spectrum as a function of the lactate concentration in the fluid from the other chromophores. The solute proposed in several studies is phosphate-buffered saline (PBS).

Thirty-seven samples with different sodium lactate (NaLac) concentrations in the range of 0 -20 mmol/L are examined. Temperature and pH are kept constant in order to avoid artifacts in the spectrum.

To validate the procedure, samples containing very high concentrations of NaLac $(100-200-300-400-500-600 mmol/L)$ are also taken, so that the peaks of interest are enhanced. Two spectrometers are used:

1. One for Visible/UV (regions between 300 - 860 nm) and for NIR (800-1800 nm);
2. The other for MIR (1800-2600 nm).

Using this instrumentation, they carried out spectral measurements on all thirty-seven samples.

A pre-processing procedure is then performed in which the spectrum obtained from the sample with a concentration of 0 mmol/L is subtracted from all the others. This procedure ensures that all the effects that the PBS components generate on the spectrum are eliminated, leaving only the information relating to lactate. In this way, they isolate the peaks due to NaLac.

To enhance the small variations in the spectrum caused by the lactate alteration, they apply the Robust Linear Multiplicative Scatter Correction algorithm. The spectrum is then filtered and processed to remove the effects of high-frequency instrument noise and all outliers are removed. The same procedure is also used for high concentration samples used for validation.

PLS is used to try to extract lactate data from the samples and separate them from water, sodium and potassium. To calculate the most suitable number of latent variables I use PRESS. The PLS model is created with a leave-one-out procedure.

In the UV/Visible spectrum, the peaks are mainly due to absorption of the PBS solute molecules, not the lactate. These are all peaks due to O-H bonds, not characteristic of the molecule.

In contrast, NIR is mainly absorbed by O-H and C-H bonds, the latter being very present in organic molecules such as lactate. After pre-processing the source spectrum, i.e., subtracting the spectrum obtained from the sample with 0 $mmol/L$ NaLac concentration, bands related to C-H bonding occur at $1215nm$, $1730nm$, $1684nm$, $2299nm$ and $2259nm$. But changes in absorbance of these bands in the spectrum, caused by changes in NaLac concentration, are hardly visible. Hence, it becomes essential to conduct a ”leave-one-out” PLS analysis. These observations show that the best absorbance peak occurs between 2100-2400 nm. Furthermore, to create the PLS model they chose to use eight latent variables (obtained by evaluating the PRESS parameter). From this model we obtained the results that $R^2 = 0.976$ and $RMSECV = 0.89mmol/L$. These values are very good and in particular the RMSECV is

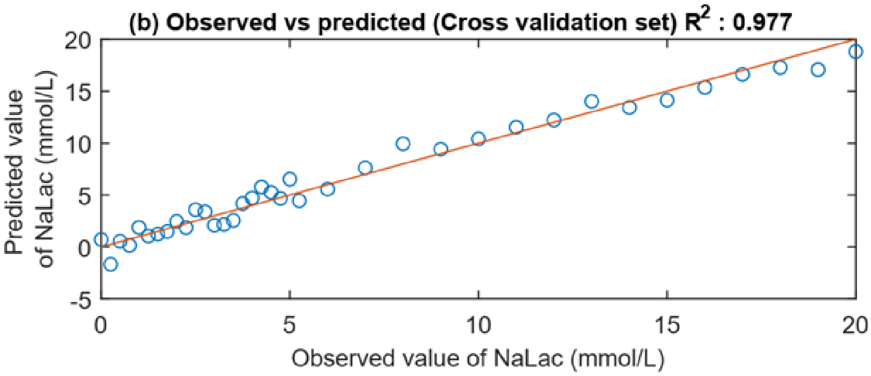


Figure 3.6: Relationship between the NaLac concentration predicted by the PLS model with NIR radiation and the reference NaLac concentration of thirty-seven samples. [10]

lower and therefore better than the one measured for the UV/Visible range.

This indicates that although the close correlation of data at very low lactate concentrations is not evident, it is easy to distinguish changes approaching 1 mmol/L lactate in PBS solutions (as per specification).

Absorption in the MIR is linked to the typical bonds of organic compounds. There are two portions of the spectrum in the IR, namely the region from 2500 nm to 6667 nm called the functional group region, or diagnostic region, which provides information on the presence of certain functional groups, and the region from 6667 nm to 25000 nm it is called ”fingerprint” and is characteristic of each individual compound. In the latter, in our case we see several overlapping absorbances, so although this band is highly specific for the compounds, these overlaps create quite a few difficulties and it is preferable not to use it. The diagnostic region, however, is more interesting. We have peaks at $3166nm, 3347nm, 3413nm$ and $3506nm$, associated with the $C = O$, $-OH$ and CH_3 bonds typical of the lactic acid molecule. Due to the complexity of the MIR spectrum, the linearity between absorbance and concentration is no longer visible, as can also be observed from measurements on samples with high lactate concentrations. However, the PLS model with 8 latent variables works very well and leads to results such as: $R^2 = 0.992$ and $RMSECV = 0.495mmol/L$. The accuracy here is comparable to that of the gold standard of these measurements, the arterial blood gas analyzer, which shows an accuracy of $< 0.5mmol/L$ over a physiological range of $1.0 - 10mmol/L$. Therefore, the MIR as a radiation range is certainly better, however it has a significant problem. MIR cannot be used if you want to measure blood lactate concentration through a wearable device, since the radiation cannot penetrate the skin and blood vessel walls, while NIR can. In our case, therefore, given that the device that must measure the radiation must be installed

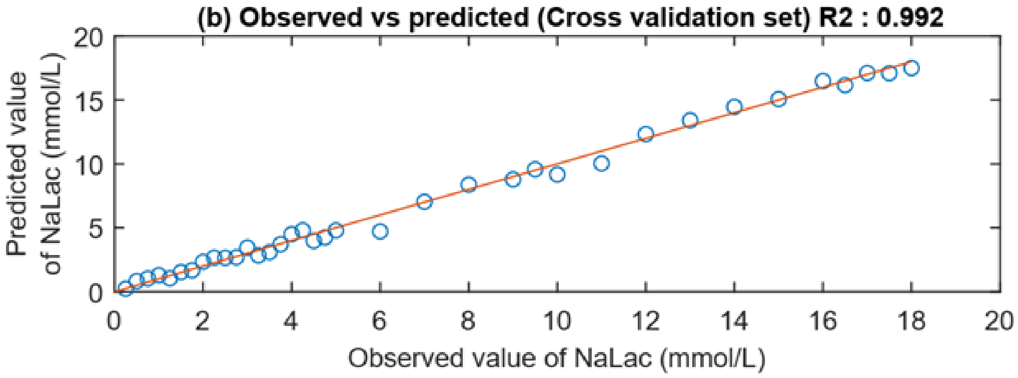


Figure 3.7: Relationship between the NaLac concentration predicted by the PLS model with MIR radiation and the reference NaLac concentration of thirty-seven samples.[10]

in a position where it is directly in contact with the blood, both could be used.

Despite of the peaks that we are able to observe in NIR range, by consulting further documentation, we have been able to establish that there were interesting peaks also in MIR range. As a result, we decided to take a closer look at technology used to conduct spectroscopic analyses in the MIR range. We want to analyse their main differences.

Through a thorough literature review, we identified research that combines spectroscopy with machine learning algorithms to develop diagnostic devices for assessing lactate concentration. In particular, the article *’Machine Learning-Assisted Raman Spectroscopy for pH and Lactate Sensing in Body Fluids’* by Olaetxea et al. ([32]) explored the application of Raman spectroscopy, demonstrating its promising utility in the medical context.

This methodology is ideal for such analyses, as it has the ability to provide high molecular selectivity despite being a non-invasive method. Furthermore, thanks to numerous studies previously conducted, Raman spectroscopy is considered to be adequately developed to perform measurements of physiological parameters for both rapid ex vivo biomedical analyses and continuous in vivo monitoring.

Despite the evident challenges in interpreting spectra, attributed to weak signals, biomolecule fluorescence, and overlapping spectral features of different sample components, it has been observed that integrating Raman spectroscopy with appropriate data preprocessing methods and Machine Learning algorithms allows for extracting crucial information to create predictive models applicable to new datasets. Unfortunately, we are not yet able to meet clinical requirements for the quantification of complex mediums like body fluids.

The proposed approach holds significant potential for improving the diagnosis of conditions related to pH and lactate, such as hypoxia and sepsis, contributing to the understanding and accurate monitoring of physiological parameters in complex biological samples.

Throughout this study, analyses were conducted both in vitro and ex vivo using Raman spectroscopy. Initial measurements were performed on aqueous solutions, demonstrating the method's validity with pH variations (from 6.80 to 7.60) and lactate concentrations. Subsequently, blood and plasma samples from domestic pigs were analyzed using a customized Raman spectroscopy system.

A crucial step in analyzing Raman spectra from biological samples is preprocessing, given the heterogeneous nature of the samples. In this study, spectra were limited between 5882 nm (MIR region) and 33333 nm (FIR region) and preprocessed in MATLAB. Raw spectra underwent sixth-order extended multiplicative signal correction to eliminate additive and multiplicative noise, along with a normalization procedure to remove interferences. Occasional noises were treated with the median of consecutive spectra, and a Savitzky-Golay filter was applied for smoothing. Finally, asymmetric least squares were used to subtract the smoothed background created by biomolecule fluorescence.

Subsequently, multivariate analysis was applied. An issue with Raman spectroscopy is the generation of a high number of variables, potentially leading to overfitting and rendering the model unusable with new datasets. Therefore, dimensionality reduction techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) were employed, as detailed in the next chapter of this thesis. PCA facilitates the visualization and interpretation of spectral data, while PLS is used to formulate predictive regression models correlating spectral variability with pH and lactate concentration in samples.

Performance analysis of predictive models is divided into two phases: the first involves constructing and calibrating the model, and the second entails validation. To mitigate overfitting and optimize the model, the leave-one-out cross-validation method is adopted. Two parameters, the Root Mean Squared Error of Prediction (RMSEP) and the coefficient of determination R^2, are employed to measure the quality of the predictive model. This entire procedure is executed in MATLAB.

For our purpose, we are particularly interested in analyzing the results obtained in lactate analysis. Three different experiments were conducted: the first with lactate in an aqueous solution to exclude complex medium effects, the second in blood, and the third in plasma.

In the first experiment, a PLS regression predictive model was constructed on preprocessed spectra of 60 samples, divided into 48 for the calibration set and 12 for the validation set. The model was calibrated through RMSECV, and the optimal number of latent vectors was determined according to the Wold criterion or ”Q² Wold criterion.” The Wold Q² is based on the difference between the explained variance of the model and the unexplained variance. A higher value of this parameter indicates a better model, as it explains more variance in the response data. The model was then tested on unknown samples, providing a very low RMSEP of 0.32 mM and an R^2 of 0.99, confirming the excellent performance of the model, as illustrated in Figure 3.8. In the second experiment, the model was developed using 32 samples of domestic pig blood, divided into 24 for calibration and 8 for validation. The predictive capacity of the model when applied to unknown data was measured once again through the parameters RMSEP and R^2, assuming values of 1.25 mM and 0.96, respectively. (Figure 3.9) 3.8. In this case, an analysis of the regression coefficients (Figure 3.10) was also performed, revealing two crucial concepts: spectral variation is primarily correlated with lactate concentrations, identifying its characteristic Raman bands; however, other spectral bands not related to lactate influence the model. To understand the origin of these spectral bands, an electrochemical analysis was conducted on four blood samples with varying lactate concentrations over two time periods. The results show that changes in clinical parameters over time can contribute to spectral effects. Additionally, the effects of prolonged use of high-power lasers causing hemoglobin denaturation were considered. All these considerations are essential for the accurate interpretation of results, emphasizing the importance of experimental condition control. 3.8.

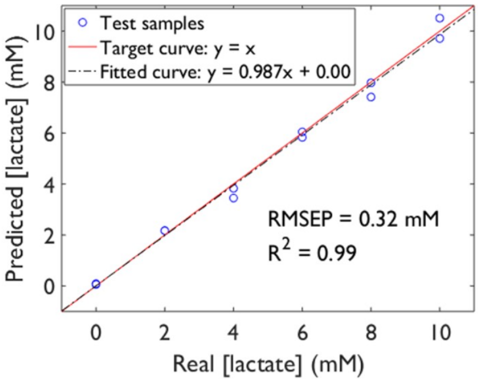


Figure 3.8: Results obtained with the PLS regression model for the detection of lactate concentration in Aqueous Solution [32]

Finally, the measurement of lactate concentration in blood plasma was performed. (Figure 3.11) The obtained model is much more robust, presenting an RMSEP of 0.51 mM. Through these various experiments, it was observed that blood, especially hemoglobin, is

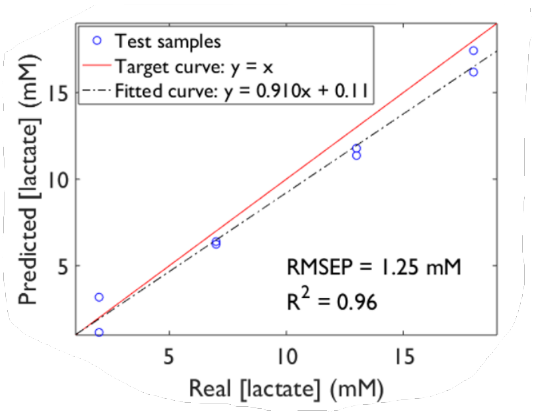


Figure 3.9: Results obtained with the PLS regression model for the detection of lactate concentration in domestic pig blood [32]

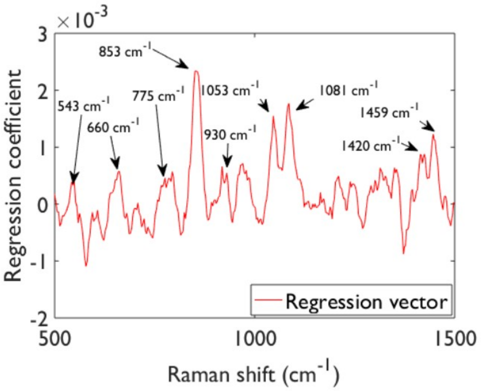


Figure 3.10: Model regression coefficients [32]

extremely sensitive to measurement conditions, adding uncontrollable variability to the predictive model. In conclusion, the articles ”Identification and Quantitative Determination of Lactate Using Optical Spectroscopy towards a Non-invasive Tool for Early Recognition of Sepsis” and ”Machine Learning-Assisted Raman Spectroscopy for pH and Lactate Sensing in Body Fluids” have further strengthened our belief that applying PLS as a Machine Learning algorithm to create a model capable of predicting lactate concentration through spectrum acquisition represents the optimal solution to this challenge.

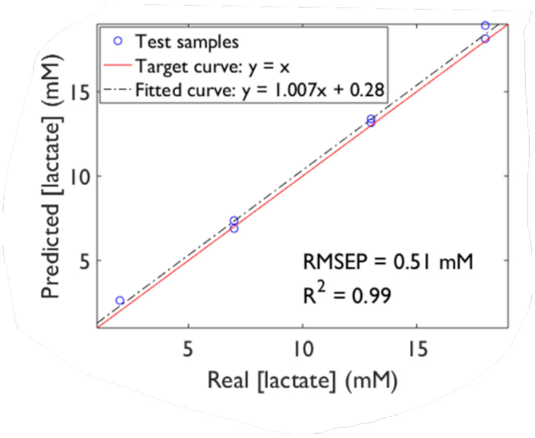


Figure 3.11: Results obtained with the PLS regression model for the detection of lactate concentration in domestic pig blood [32]

Chapter 4

Mathematical Models for the Prediction of Biochemical Substance Concentrations

This chapter delves into the pivotal role of latent variable modeling in predicting concentrations of biochemical substances from spectra. ”Latent variable modeling” proves indispensable in overcoming challenges posed by complex spectral data and decoding intricate molecular interactions.

Latent variables serve as a crucial link between direct spectral measurements and the sought-after information on biochemical concentrations. Their ability to reveal concealed features in data becomes especially pertinent when dealing with the intricacies of mixed samples, where spectral overlap complicates direct differentiation.

Through latent variable modeling, we extract meaningful insights and unveil concealed patterns in spectral data, providing a clearer representation of the relationships between spectra and biochemical concentrations. This goes beyond the visible surface of spectra, expanding latent dimensions to heighten the accuracy and sensitivity of our predictive model.

The chapter underscores the significance of latent variable modeling in predicting biochemical concentrations, with a specific focus on its application in our research. The analysis emphasizes how the inclusion of latent variables has deepened interpretative insights and improved predictive accuracy, thus significantly contributing to the success of our scientific inquiry.

Within the same context, the chapter concentrates on a comprehensive analysis of mathematical models for predicting concentrations of biochemical substances. While briefly exploring *Multiple Linear Regression (MLR), Principal Component Analysis (PCA)*, and *Partial Least Squares Regression (PCR)*, our primary focus lies on **Partial Least Squares Regression (PLS)**. This choice stems from the demonstrated effectiveness of PLS in our practical applications, particularly in the precise prediction of substance concentrations in mixed samples based on their spectra. Through a focused analysis of PLS, we will explore its fundamental aspects, highlighting its distinct advantages over alternative models in our specific biochemical application.

4.1 Exploring Dataset Configurations: From Classic Tools to Contemporary Complexities

During the development period of manufacturing and chemical engineering industries between the '20s and '50s, data collected from processes were limited to a few columns, obtained through manual measurements incurring considerable costs. "Classic" tools used included scatter plots, time-series plots, and multiple linear regression (MLR) models based on least squares.

The representation of each dataset occurs through a matrix, where each row contains an observation or sample, and each column represents a specific variable or attribute associated with each observation or sample.

During that era, datasets often had more rows than columns, as adding new columns involved high expenses and time. The selection of columns to measure was carefully considered to avoid unnecessary duplications.

However, while these datasets met the requirements for using "classic" tools, modern datasets exhibit more varied configurations, with a significant number of columns for each observation. If we consider a dataset with N rows and K columns, we can identify different types of datasets:

1. **Both N and K small:** This case is typical of complex and expensive measurements, generally analyzed using "classic" methods.

2. N **small but** K **large:** This situation is common in laboratory instrumentation, especially in spectroscopy. Instruments provide spectral responses at numerous wavelengths, represented in a matrix. The main challenge is managing $K > N$, making the application of ordinary linear regression models difficult. Selecting relevant columns requires an advanced approach, such as latent structure projection, to effectively address data non-independence.

3. N **large and** K **small:** Modern chemical refineries generate vast amounts of data, with

numerous observations per second on a high number of variables. Analyzing such data requires advanced approaches.

4. N **and** K **approximately equal:** In situations where the number of variables approximately matches the number of observations, we get quadrangular matrices. These peculiar situations can influence data analysis.

5. **Matrices** X **and** Y**:** This configuration occurs when we want to predict one or more variables from a group of other variables, common in linear regressions.

6. **Three-dimensional and higher-dimensional datasets:** With the increased use of new technologies, three-dimensional and higher-dimensional datasets have become common. They require sophisticated analytical approaches to handle the additional complexity of extra dimensions.

7. **Batch datasets:** In batch systems, typical in high-value industries, we combine data describing batch preparation (matrix Z) with constant data during the batch duration (matrix X) and the final product properties (matrix Y).

8. **Data Fusion:** Data fusion involves collecting and utilizing data from various sources, such as near-infrared probes, offering integrated analysis.

These various dataset configurations highlight the importance of considering the size of N and K, with particular emphasis on the case where N is small and K is large, a situation addressed in the current study.

4.2 Critical Challenges with Engineering Data: Size, Independence, Noise, and Other Considerations

Data analysis in engineering encounters several challenges, primarily related to the size of the data, lack of independence, low signal-to-noise ratio, the non-causal nature of data, and measurement errors, not to mention the presence of missing data.
Data Size: A distinctive feature of the datasets in question is their vast size in terms of rows and columns, primarily due to the increasing affordability of data acquisition and storage. Managing a high number of columns, especially when exceeding 10, becomes complex, requiring specialized tools to simplify the analysis.

Lack of Independence: The modern complexity of datasets is reflected in the lack of data independence, complicating the application of models such as multiple linear regression (MLR), where the $(X'X)$ matrix becomes singular with strongly dependent data. Selecting a reduced number of columns may make the data more independent, but this process is often cumbersome and risky.

Low Signal-to-Noise Ratio: Engineering systems, aiming for stability, produce data with minimal signals and high noise. Even though the recording is high-frequency, much information is discarded by computer systems. Finding meaningful signals in this sea of data is a considerable challenge.

Non-Causal Data: The nature of happenstance data limits the ability to establish cause-and-effect relationships. However, correlation models can provide valuable insights, and causality verification can be subsequently performed through randomly designed experiments.

Errors in Data: Conventional tools often operate under the assumption of error-free data, while many measurements in engineering systems have errors, often significant. A more flexible approach is needed to handle this reality.

Missing Data: The frequent presence of missing data requires methodologies that go beyond simply eliminating rows or columns, avoiding the loss of crucial information.

In conclusion, addressing these challenges requires advanced methodologies like latent variable methods, capable of rapidly handling large amounts of data, addressing missing data, considering higher dimensions, and ensuring flexibility compared to traditional assumptions.

4.3 Multiple Linear Regression (MLR) Model

4.3.1 Mathematical Foundations

Linear Regression serves as a fundamental tool for elucidating the relationship between the dependent variable y and the independent variable x through the regression equation:

$$\hat{y}_i = \beta_0 + \beta_1 X_i, \quad i = 1, 2, 3, ..., n \tag{4.1}$$

Expanding upon this, Multiple Linear Regression (MLR) extends the paradigm to scenarios involving two or more independent variables. Multiple linear regression shares the same assumptions as simple linear regression, including:

1. **Uniformity of Variance (Homoscedasticity):** The magnitude of prediction errors remains relatively constant across different values of the independent variable.

2. **Independence of Observations:** Data points are collected using statistically sound sampling methods, ensuring no concealed relationships among variables.

3. **Multicollinearity Check:** In multiple linear regression, it's crucial to examine potential correlations among independent variables. If two variables display a high correlation ($r^2 >\sim 0.6$), only one should be included in the regression model.

4. **Normal Distribution:** The data adheres to a normal distribution.

5. **Linearity:** The relationship between the independent and dependent variables can be adequately represented by a straight line, avoiding curves or groupings.

For a system with k variables, the MLR equation takes the form:

$$\hat{y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}, \quad i = 1, 2, 3, ..., n \tag{4.2}$$

Herein, Y embodies the dependent variable, $X_1, X_2, ..., X_n$ denote independent variables, β_0 represents the y-intercept, and $\beta_1, \beta_2, ..., \beta_n$ signify regression coefficients denoting the change in Y for a one-unit alteration in the corresponding independent variable while keeping others constant.

Expressing equation (4.2) in matrix form, we have:

$$\hat{\mathbf{Y}} = \beta_0 + \mathbf{X}\beta \tag{4.3}$$

In practical applications, MLR is employed to determine the plane that best fits the data. Although models with more than two independent variables exhibit increased structural complexity, MLR techniques remain applicable. In the preprocessing phase, mean-centering of the x and y vectors is undertaken, enabling the omission of the model's intercept term β_0.

MLR analysis leverages the method of least squares to estimate the regression coefficients β_i. The objective is to minimize the discrepancy between the model's output $\hat{Y_i}$ and the actual value of Y, with the disparities incorporated into the error vector e. The least square objective function is formulated as:

$$f(\beta) = e^T e = (y - X\beta)^T (y - X\beta) = y^T y - 2y^T X\beta + \beta X^T X\beta \tag{4.4}$$

By taking the partial derivative with respect to the entries β and equating the result to a vector of zeros, we establish that:

$$\beta = (X^T X)^{-1} X^T y \tag{4.5}$$

Despite its efficacy, MLR is not without limitations. Assumptions such as the absence of measurement errors in the variables and the inability to handle missing data pose challenges. The method encounters issues with strongly correlated columns in X, assumes noise-free X, and lacks the capacity to address missing values in X. Moreover, the requirement for $N > K$ can be impractical in certain scenarios, leading to sub-optimal predictions and necessitating variable selection.

4.3.2 Model Operation and Application Domains

The operational foundation of the Multiple Linear Regression (MLR) model lies in its ability to capture the intricate relationships between a dependent variable and multiple independent variables. This model extends the principles of simple linear regression to accommodate the complexities of real-world scenarios where two or more predictors influence the outcome.

In practical terms, MLR excels in determining the best-fitting plane for a given dataset, making it a valuable tool in fields such as statistics, economics, and various scientific disciplines. The application domains of MLR span diverse industries, where it is employed for tasks ranging from predicting stock prices and economic trends to analyzing experimental data in scientific research.

4.3.3 Pros and Cons

While Multiple Linear Regression (MLR) offers valuable insights and predictive capabilities, it is crucial to acknowledge its strengths and limitations. The method's strengths lie in its simplicity, interpretability, and efficiency in capturing linear relationships between multiple variables. MLR is particularly adept at revealing the impact of individual predictors on the dependent variable, providing a comprehensive understanding of the system under study. However, the limitations of MLR are notable.

Assumptions such as the absence of measurement errors and the necessity for noise-free data can limit its applicability in real-world scenarios. Challenges arise when dealing with strongly correlated independent variables, and the method's inability to handle missing data poses constraints. Additionally, the requirement for a higher number of observations ($N > K$) may be impractical in certain situations, necessitating careful consideration and potentially leading to suboptimal predictions. Despite these limitations, MLR remains a foundational tool in statistical modeling, offering valuable insights when applied judiciously in appropriate contexts.

4.4 Latent Variable Modeling

4.4.1 Concept of Latent Variables

The concept of a latent variable refers to a quantity that is not directly observable or measurable but can be deduced or inferred through the observation of other measurable variables. In other words, it is a variable that influences observed data but cannot be directly observed or measured independently.

Latent variables find extensive use in various fields, including statistics, psychology, machine learning, and data modeling. In the context of statistical modeling, latent variables are often introduced to represent abstract or hidden concepts that cannot be directly measured. In machine learning models, latent variables are frequently employed to represent hidden or abstract features in the data, enabling the capture of complex relationships between input and output variables.

Conceptually, a latent variable is an abstract construct that cannot be directly measured, such as an individual's overall health. In this context, physical measurements like blood pressure, cholesterol level, weight, and other quantities are used to assess overall health. The latent variable, therefore, represents the underlying phenomenon that cannot be directly observed but is correlated with the measurements taken.

Mathematically, a latent variable can be expressed as a linear combination of observable variables, weighted by appropriate coefficients. If we consider a set of observable variables $X_1, X_2, \ldots, X_K$, a latent variable T can be represented as $T = \sum_{i=1}^{K} X_i \cdot p_i$, where p_i denotes the weights associated with each observable variable. This formulation allows the synthesis of information contained in observable variables by constructing a latent variable that captures the common essence of the data.

From a geometric perspective, the analysis of latent variables can be visualized in a multidimensional space, where observable variables are represented as vectors. The latent variable can be interpreted as a combination of these dimensions, highlighting the correlation and coherence in the data.

4.5 Principal Component Analysis (PCA) Model

4.5.1 Mathematical Foundations

Principal Component Analysis (PCA) is a fundamental technique in multivariate statistics, primarily employed to streamline the description of a dataset by reducing the number of variables to a more manageable set of *latent variables*. The objective is to minimize potential information loss in this process.

The methodology involves a linear transformation of variables, projecting the original data onto a new Cartesian coordinate system. Visualized as a point cloud, this representation illustrates how variables co-vary. The initial step, termed *mean-centering*, shifts these points to the central position of the coordinate system, eliminating undesired biases from measurements. Subsequently, scaling the data, often to unit variance, ensures uniform units of measurement. Establishing new coordinates for the reference system involves identifying principal components, each comprising a *direction vector* p_i of size $K \times 1$and a *vector of scores* t_i, representing distances measured by projecting data onto the new Cartesian coordinate system. When employing a singular component, the latent variable model takes the form of a line; introducing a second component extends the model to a plane. Subsequently, employing three or more components defines the model within a hyperplane. This hyperplane serves as the optimal approximation of the original data, and the perpendicular distance from each point on the plane is termed *residual distance* or *residual error*.

PCA effectively discerns raw data into a latent variable model and a residual error.

The objective post identifying the optimal line is to minimize the error, i.e., maximize the variance among the scores in the score vector t. Considering x_i as a row of raw data, the score for an observation $(t_{i,1})$ is defined as the distance from the origin of the vector p_1 to the point where the perpendicular for that data falls.

These mathematical transformations and the underlying principles emphasize the efficacy of PCA in simplifying complex datasets while retaining essential information.

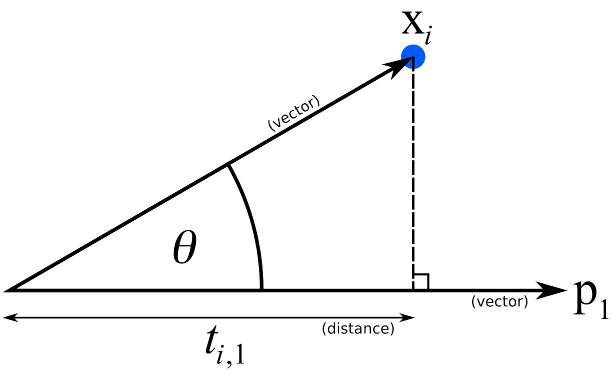


Figure 4.1: Component along a vector [18]

Recalling trigoniometric relations, we observe that:

$$cos(\theta) = \frac{adjacentlength}{hypothenuse} = \frac{t_{i,1}}{\|x_i\|}$$
$$cos(\theta) = \frac{x_1^{'} p_1}{\|x_1^{'}\| \|p_1\|}$$
$$\frac{t_{i,1}}{\|x_i\|} = \frac{x_1^{'} p_1}{\|x_1^{'}\| \|p_1\|}$$
$$t_{i,1} = x_1^{'} p_1 \tag{4.6}$$
$$(1 \times 1) = (1 \times K)(K \times 1)$$

where $\| \cdot \|$ indicates the length of the contained vector. Given that the vector p_1 has length 1, we can derive equation 4.6. The $t_{i,1} = x_1^{'} p_1$ represents a linear combination:

$$t_{i,1} = x_{i,1} p_{1,1} + x_{i,2} p_{2,1} + ... + x_{i,k} p_{k,1} + ... + x_{i,K} p_{K,1}$$

Similarly, the expression for the second score value for the i^{th} observation is given by:

$$t_{i,2} = x_{i,1} p_{1,2} + x_{i,2} p_{2,2} + ... + x_{i,k} p_{k,2} + ... + x_{i,K} p_{K,2}$$

and so on for all the components needed.
This entire process can also be expressed in a concise matrix form:

$$t_i^{'} = x_i^{'} P \tag{4.7}$$
$$(1 \times A) = (1 \times K)(K \times A) \tag{4.8}$$

This approach enables the simultaneous calculation of all A-score values for a given observation and, consequently, for the entire data matrix:

$$T = XP \tag{4.9}$$

$$(N \times A) = (N \times K)(K \times A) \tag{4.10}$$

4.5.2 Model Operation and Application Domains

PCA involves three fundamental steps:

1. Preprocessing of the raw data;

2. Eigenvalue decomposition;

3. Testing by, for example, Squared Prediction Error (SPE) and Hotelling's T^2.

Pre-processing can be done the different ways such as transformations, expanding the initial matrix, handling outliers, centering data in the chosen reference system, and scaling to ensure uniform units.

Once the preprocessing procedure is completed, we move on to the eigenvalue decomposition procedure. It can be implemented in different ways explained step by step below.

Eigenvalue decomposition

The initial step of the method involves formulating an optimization problem, recognizing that the direction of latent variables (or the loading vector) is aligned to maximize the variance of scores while being orthogonal to each other. For the first component, the objective function is defined as:

$$\begin{gathered} max\phi = t_1^{'}t_1 = p_1^{'}X^{'}Xp_1 \\ s.t.p_1^{'}p_1 = 1 \end{gathered} \tag{4.11}$$

The conventional approach for computing principal components employs Lagrange multipliers, transforming the constraint into an objective function. The maximum value is obtained when the partial derivatives with respect to p_1 are set to 0:

$$\begin{aligned} \frac{\partial \phi}{\partial p_1} = 0 = p_1^{'} X^{'} X p_1 - \lambda_1 (p_1^{'} p_1 - 1) \\ 0 = 2X^{'} X p_1 - 2\lambda_1 p_1 \\ 0 = p_1 (X^{'} X - \lambda_1 I_{K \times K}) \\ X^{'} X p_1 = \lambda_1 p_1 \end{aligned} \tag{4.12}$$

Equation 4.12ncludes both the eigenvectors $X^{`}X$ and the eigenvalues λ_1.

The procedure for the second principal component is similar, with the addition of the constraint representing the orthogonality of the direction vectors ($p_1 \perp p_2$). This can be mathematically translated with the following function: $X^{'} X p_2 = \lambda_2 p_2$.

To apply eigenvalue decomposition, once the raw data has been preprocessed, the correlation matrix $X^{'}X$, is calculated. Subsequently, the eigenvalues and eigenvectors are determined and sorted in decreasing order. From a computational standpoint, calculating all eigenvalues can be resource-intensive. The maximum number of eigenvalues to calculate is $A_{max} = min(N, K)$.

Singular Value Decomposition (SVD)

A second technique for performing eigenvalue decomposition is Singular Value Decomposition (SVD). This method involves decomposing the matrix X into three matrices:

$$X = U \Sigma V^{'} \tag{4.13}$$

Here, U and V are orthonormal matrices, and Σ is a diagonal matrix. The relationship linking this decomposition to PCA is given by:

$$X = T P^{'} \tag{4.14}$$

In this equation, P is also an orthonormal matrix. After applying SVD to the source matrix X, it can be considered that $P = V$ and $T = U\Sigma$. Moreover, the terms on the diagonal of Σ are associated with the variance of each principal component. While this method enables the calculation of all possible main components, it defaults in handling missing data.

Non-Linear Iterative Partial Least-Squares (NIPALS)

The NIPALS Algorithm was initially developed for PCA and later adapted for Projection to Latent Structures (PLS). It is a widely used method for calculating the principal components of a dataset, offering more numerically accurate results compared to the SVD of the covariance matrix, albeit being slower to calculate. The method can be summarized by outlining its various steps:

1. Generate an initial column for t_a, which can either be a column with random values or a column selected at random from those constituting the matrix X;

2. Conduct a Least Square Regression (LSR) on the initial column of t_a for each column of X. Once the regression coefficients are computed, store them in p_a, as expressed by the equation:

$$p_{a,k} = \frac{t_a^{'} X_k}{t_a^{'} t_a}$$

3. Normalize the vector p_a to have a magnitude of 1.0.

4. For each row of X, perform a Least Square Regression onto p_a and let the regression coefficients become the values of the scores for the i^{th} row, stored in the t_a vector:

$$t_{i,a} = \frac{x_i^{'} p_a}{p_a^{'} p_a};$$

5. Iterate through steps 2, 3, 4 until the changes between one iteration and the next are very small, approximately on the order of 10^{-6} or 10^{-9};

6. Store the vectors t_a and p_a in the a^{th} column of the respective matrices, namely T and P. After computing the vectors of variable weights (scores) and principal component coefficients (loadings) for the first principal component, deflate the data. This involves removing the information explained by the first principal component from the original data, resulting in a residual data set. Mathematically, this is accomplished as follows:

$$\begin{aligned} E_a &= X_a - t_a p_a^{'} \\ X_{a+1} &= E_a \end{aligned} \tag{4.15}$$

7. Return to step 1 and repeat the entire procedure for the subsequent components.

The final step in PCA involves model testing, where the goodness of fit of the model is evaluated using precise indices such as Hotelling's T^2 and SPE.

Squared prediction error (SPE)

After obtaining the PCA model, let's assume, for simplicity, a single component. The best estimate of the observation x_i is the point along p_1 where the original observation is projected. Additionally, $t_{i,1}$ is the distance from the origin of the direction vector to the projection point, and the actual point along p_1 is a vector representing our best estimate of the original observation, denoted as $\hat{x}_{i,1}$.

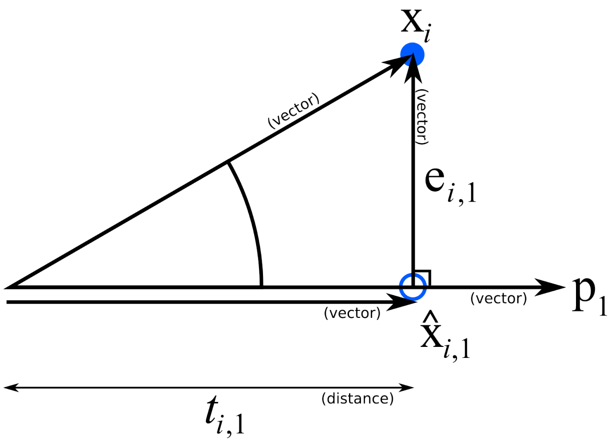


Figure 4.2: Prediction along a vector [18]

Mathematically, I can express $\hat{x}_{i,1}$ as follows:

$$\begin{gathered} \hat{x}_{i,1}^{'} = t_{i,1}p_1^{'} \\ (1 \times K) = (1 \times 1)(1 \times K) \end{gathered} \tag{4.16}$$

If we consider adding a second component, the expression becomes:

$$\begin{gathered} \hat{x}_{i,2}^{'} = t_{i,1}p_1^{'} + t_{i,2}p_2^{'} \\ (1 \times K) = (1 \times K) + (1 \times K) \end{gathered} \tag{4.17}$$

In the case of multiple variables, the matrix form is more suitable:

$$\hat{X} = TP^{'} \tag{4.18}$$
$$(N \times K) = (N \times A)(A \times K)$$

Once the estimated value is obtained, calculating the vector of residuals is useful. This vector contains the differences between the actual value and the one calculated by the model:

$$e^{'}_{i,A} = x^{'}_{i} - \hat{x}^{'}_{i,A} \tag{4.19}$$
$$(1 \times K) = (1 \times K) - (1 \times K)$$

The residual distance is the sum of the squares of the residuals, and to calculate the distance, we take the square root. The Squared Prediction Error (SPE) is defined as:

$$SPE_i = \sqrt{e^{'}_{i,A} e_{i,A}} \tag{4.20}$$
$$(1 \times 1) = (1 \times K)(K \times 1)$$

where $e_{i,A}$ is the residual vector of the i^{th} observation using A components.

Hotelling's T^2

Another critical metric for evaluating the PCA model is Hotelling's T^2. Hotelling's T^2 distribution is an extension of Student's t-distribution used in multivariate hypothesis testing. The value of T^2 for the i^{th} observation is defined as:

$$T^2 = \sum_{a=1}^{a=A} \left(\frac{t_{i,a}}{s_a} \right)^2 \tag{4.21}$$

where s_a^2 values are constant and represent the variances of each component. T^2 is a scalar value that summarizes all score values. It is a non-negative number, and it indicates the distance from the center of the hyperplane to the projection of the observation onto it.
The calculated T^2 value is a measure of the deviation between the multivariate means of the considered groups. A higher T^2 value indicates a greater difference between the groups.

4.5.3 Pros and Cons

Principal Component Analysis (PCA) is a widely employed technique in data analysis, offering both advantages and limitations.

One of its primary strengths lies in its ability to effectively reduce the dimensionality of datasets by transforming them into a set of uncorrelated variables known as principal components. This facilitates simplified data representation and aids in identifying the most influential features, making it particularly useful for large and complex datasets. PCA also serves as a valuable tool for feature extraction, capturing underlying patterns and structures within the data. Its application is advantageous in scenarios where multicollinearity is prevalent, as it mitigates issues associated with highly correlated variables. Moreover, PCA aids in visualization by projecting high-dimensional data into a reduced-dimensional space, enabling the identification of trends, clusters, and outliers.

Despite these benefits, it is essential to acknowledge certain drawbacks, such as the potential loss of interpretability due to the abstract nature of principal components. PCA assumes linear relationships between variables, which may not always hold true, and it is sensitive to the scale of variables, making careful preprocessing necessary. Additionally, its sensitivity to outliers and the orthogonality assumption of principal components are aspects that should be considered when applying PCA in a Master's thesis research context. Understanding these pros and cons is crucial for making informed decisions about the appropriateness of PCA for a specific dataset and research objectives.

4.6 Principal Component Regression (PCR) Model

PCR is a statistical analysis technique that is based on PCA. In particular, it is used to estimate unknown regression coefficients in a standard linear regression model. Furthermore, it represents one of the methods by which I collect and process data from a spectrometer. PCR also represents an alternative to MLR and has several advantages over it. In PCR we have that the principal components of the explanatory variable are used as regressors. The main idea of the method is to replace the K columns in the raw data matrix and to do this, we are going to replace the $N \times K$ raw data matrix with a smaller one, $N \times A$, which summarizes the source matrix.

At this point we are going to identify a relationship that connects the A scores to the y variable and these two steps can be represented mathematically as follows:

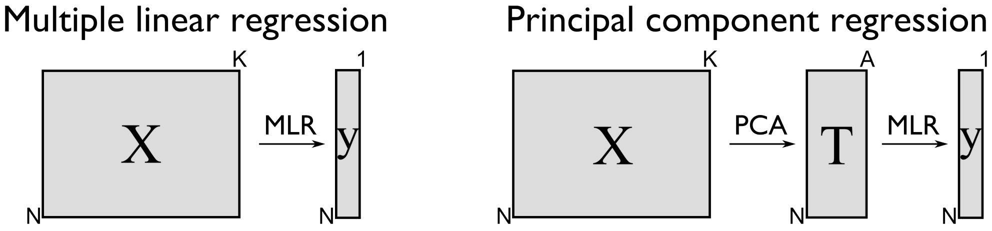


Figure 4.3: PCR data structure compared to MLR [18]

1. $T = XP$ from the PCA model
2. $\hat{y} = Tb$

The second equation can be solved as $b = (T'T)^{-1}T'y$
In the context of constructing and using the Principal Component Regression (PCR) model, it is crucial to follow a specific set of steps:

1. **Observation of New Data:** Collect the (X) and (y) data required for the model.
2. **Building the PCA Model:** Use the data in X to develop a PCA model, determining the number of components (A) through cross-validation. Evaluate the need to modify (A) after the initial model.
3. **Analysis of PCA Graphs:** Examine the Squared Prediction Error (SPE) and T^2 plots generated by the PCA model to ensure that the model is not influenced by outliers.
4. **Utilizing Columns in (T):** Use the columns in (T) from the PCA model as (X) variables in the normal Multiple Linear Regression (MLR) model.
5. **Estimation of MLR Parameters:** Solve for the parameters of the MLR model, represented by a vector b, through the equation $b = (T'T)^{-1}T'y$, where each coefficient in b corresponds to a score.
6. **Application of the Model to New Observations:** For a new observation represented by the vector $x'_{\text{new, raw}}$:
   - Preprocess the vector as done during the construction of the PCA model.
   - Calculate the scores t'_{new} for the new observation.
   - Obtain the predicted value $\hat{x}'_{\text{new}}$ and the residual vector e'_{new}.

- Calculate the residual distance from the model plane (SPE_{new}) and the value of Hotelling's T^2 (T^2_{new}).
- Before calculating the final prediction, check if SPE_{new} and T^2_{new} are below the 95% or 99% limits. If yes, proceed with the calculation of the prediction $\hat{y}'_{\text{new}} = t'_{\text{new}}b$; otherwise, investigate the reasons for the unusual behavior of the new observation.

4.6.1 Application Domains

Principal Component Regression (PCR) is a versatile technique widely employed in various applications. In the fields of chemistry and spectroscopy, PCR contributes to the analysis of complex spectral data, such as that derived from nuclear magnetic resonance (NMR) or infrared spectra. In process engineering, PCR models and optimizes industrial processes by identifying key variables. Environmental applications involve using PCR to analyze complex data from environmental monitoring, pinpointing major contributors to variations. In biology and genetics, PCR analyzes molecular data, including gene expression, revealing relationships between biological variables. In the financial sector, PCR is applied to the analysis of complex economic data. In medicine, it aids in the analysis of clinical data to identify key factors. In industrial quality control, PCR monitors and improves product quality by identifying critical variables. Overall, PCR adapts to contexts with complex and multicollinear data, making it valuable in various scientific and industrial disciplines.

4.6.2 Pros and Cons

Principal Component Regression (PCR) is a modeling technique that integrates Principal Component Analysis (PCA) with Multiple Linear Regression (MLR). It offers several advantages, including effective handling of multicollinearity by reducing variables into uncorrelated principal components. PCR also allows for the reduction of the original variable set, particularly beneficial when dealing with numerous variables and aiming to simplify the model. Moreover, it helps prevent overfitting through dimensionality reduction, crucial in scenarios with limited observations. However, PCR comes with challenges. The interpretation becomes more complex as it transforms original variables into principal components, lacking intuitive significance. Additionally, it relies on the linear assumption between independent and dependent variables, which might not hold for all data types. There is a risk of overfitting, especially if an excessive number of principal components is used relative to observations. PCR may be sensitive to outliers, influencing both principal components and subsequent regression. Furthermore, it necessitates scaling, making it advisable to standardize or normalize variables before application. In conclusion, the decision to employ PCR should be based on a careful evaluation of the specific problem requirements and the characteristics of the available data.

4.7 Partial Least Squares (PLS) Regression Model

The origin of PLSR lies in the field of chemistry. The development of near-infrared spectroscopy (NIR) would have been challenging without a method to quantitatively analyze spectra, given their highly overlapping lines and challenging overtones.

In PLSR, there is typically a minor component in X highly correlated with Y. The inclusion of this component in the first latent variable distinguishes PLSR from other methods. Despite similarities with PCR, PLSR often requires fewer latent variables, covering more variation in Y and achieving comparable prediction accuracies.

PCR and PLSR act as shrinkage methods in algorithms, with PLSR occasionally increasing the variance of individual regression coefficients. This phenomenon may explain why PLSR is not consistently superior to PCR.

PLS aims to find latent variables that best explain X_a and Y_a, facilitating the creation of a robust relationship between X_a and Y_a. PLS involves three simultaneous operations.

4.7.1 Model Operation and Application Domains

PLS simultaneously extracts the score vectors for X and for Y:

$$\begin{aligned} t_a &= X_a w_a \\ u_a &= Y_a c_a \end{aligned}$$

The goal of PLS is to extract these scores in such a way as to obtain maximum covariance. Since a high covariance implies a strong correlation between the two vectors, it is more useful to speak of correlation and variance. Accordingly, we can express it as follows:

$$\begin{aligned} Cov(t_a, u_a) &= Correlation(t_a, u_a) \times \sqrt{}(Var(t_a)) \times \sqrt{}(Var(u_a)) \\ Cov(t_a, u_a) &= Correlation(t_a, u_a) \times \sqrt{}(t_a^{'} t_a) \times \sqrt{}(u_a u_a^{'}) \end{aligned} \tag{4.22}$$

The objective is to maximise the three components simultaneously which allows us to obtain both the best explanation of the space X_a and that of the space Y_a, which are given by the two variances, plus I also obtain the best relationship between X_a and Y_a, which, on the other hand, is derived from the correlation.
All scores are subject to the same constraint, i.e. $w_a {w_a}' = 1.0$ and $c_a {c_a}' = 1.0$.

4.7.2 Construction of the PLS model

To create the PLS model, apply the NIPALS algorithm to the preprocessed X_a and Y_a matrices when $a = 1$.
Also in this case, we will illustrate the various steps:

1. First, we will randomly take a column of the matrix Y_a as the initial estimate of u_a;

2. Next, I regress each column from X_a onto the vector u_a. The regression coefficients are stored as entries of w_a as follows

$$w_a = \frac{X_a^{'} u_a}{u_a u_a^{'}}$$

3. Then, we proceed with the normalization of the weight vector, following the formula

$$w_a = \frac{w_a}{\sqrt{w_a w_a^{'}}}$$

4. At this point I am going to regress each row of X_a on the weight vector, t_a. This means that the rows in X_a that have characteristics similar to those contained in the weight vector will have a higher value in t_a, while the observations that are totally different from what is reported in w_a are recorded with value close to zero. The values stored in t_a follow the following formula:

$$t_a = \frac{X_a w_a}{w_a w_a^{'}}$$

5. Now, I regress each column from Y_a onto the score vector t_a. The regression coefficients are stored as entries of c_a as follows

$$c_a = \frac{Y_a^{'} t_a}{t_a t_a^{'}}$$

6. Then, we proceed with the normalization of the weight vector, following the formula

$$c_a = \frac{c_a}{\sqrt{c_a c_a^{'}}}$$

7. Again, I am going to regress each row of Y_a on the weight vector, c_a. The values stored in c_a follow the following formula:

$$c_a = \frac{Y_a c_a}{c_a c_a^{'}}$$

After all these steps, we proceed with the deflation of the matrix X, in order to remove the already explained variability from X_a and Y_a. To do this, using the vector w_a, I remove from X_a its best prediction, $\hat{X}_a$. Mathematically, this involves the following steps

$$\hat{X}_a = t_a w_a^{'}$$
$$E_a = X_a - \hat{X}_a = X_a - t_a w_a^{'}$$
$$X_{a+1} = E_a$$

The exact same procedure is also applied to Y_a, that is

$$\hat{Y}_a = t_a c_a^{'}$$
$$E_a = Y_a - \hat{Y}_a = Y_a - t_a c_a^{'}$$
$$Y_{a+1} = E_a$$

Note that we never used the score vector u_a. This is because if we apply the PLS model to a new data set in the future, we will not know the values of y and consequently not even the values contained in u_a.

Very often when using PLS models, rather than using the W vector, it is preferred to use the R vector for the weights. Specifically, R is a matrix whose columns can be found by doing:

$$R = W(P^{'}W)^{-1} \quad where \quad r_1 = w_1 \tag{4.23}$$

4.7.3 Pros and Cons

While Partial Least Squares Regression (PLS) finds extensive applications in fields such as chemometrics and spectroscopy, it is less commonly employed or understood in certain domains. For instance, traditional statistical fields, like classical linear regression or analysis of variance, may be more prevalent in disciplines such as statistics or economics. In areas where practitioners are more accustomed to classical statistical methods and less inclined towards advanced machine learning techniques, PLS might face lower adoption rates. Moreover, some branches of machine learning, particularly those favoring transparent and easily interpretable models, might opt for simpler techniques like linear regression or decision trees. PLS has the advantage of handling multicollinearity and is suitable for datasets with complex relationships, making it valuable in specific contexts. However, its sensitivity to outliers and less widespread use in fields emphasizing model interpretability and underlying statistical assumptions could limit its application. Additionally, in disciplines where the focus is primarily on hypothesis testing or causal inference, PLS may not be leveraged as extensively, as its primary strength lies in predictive modeling and capturing complex relationships. The less common use of PLS in certain fields is not indicative of its inefficiency but rather reflects the diversity of statistical and machine learning tools available, with each field often having its preferred methods influenced by historical practices, disciplinary norms, and the nature of the data typically encountered.

4.8 Comparison and Model Selection

4.8.1 Comparative Analysis of Models

In the realm of predictive modeling for your specific aim of predicting substance concentration from a spectrum, three prominent techniques stand out: Multiple Linear Regression (MLR), Principal Component Regression (PCR), and Partial Least Squares (PLS). MLR, a classical approach, models the relationship between multiple independent variables and a dependent variable through a linear equation. It assumes independence of observations, normal distribution, and linearity, making it a foundational method. However, MLR has limitations, including sensitivity to multicollinearity and the need for a high number of observations relative to variables.

PCR, an extension of PCA combined with MLR, addresses multicollinearity by transforming variables into uncorrelated principal components. It simplifies complex datasets and allows for efficient dimensionality reduction. However, PCR has drawbacks, such as the challenge of interpreting principal components and a reliance on the linear assumption between variables.

PLS, specifically chosen for your case, excels in predicting substance concentration from a spectrum. It integrates features of PCA and MLR, extracting latent variables that maximize covariance between spectral data and concentration. PLS is particularly effective with high-dimensional and collinear data, offering a robust solution for capturing complex relationships. It addresses the limitations of MLR and PCR, providing a powerful tool for your analytical objectives.

In summary, while MLR provides a straightforward approach, PCR focuses on dimensionality reduction, and PLS emerges as a versatile solution for predicting substance concentration from spectral data. The choice among these models should consider the specific characteristics of your dataset and the trade-offs between simplicity, interpretability, and predictive performance.

4.8.2 Selected Methods for Concentration Prediction

For the specific aim of predicting substance concentration from a spectrum, the selected method is Partial Least Squares (PLS). This choice is based on several factors that align with the unique characteristics and requirements of your analytical objectives.

Firstly, PLS demonstrates exceptional efficacy in handling high-dimensional and collinear spectral data, which is inherent in scenarios involving substance concentration prediction from spectra. By extracting latent variables that maximize the covariance between the spectral features and the concentration levels, PLS captures complex relationships effectively.

Moreover, PLS integrates the strengths of both Principal Component Analysis (PCA) and Multiple Linear Regression (MLR). While PCA alone focuses on dimensionality reduction, PLS goes further by considering the relationship with the response variable (substance concentration). This dual focus allows PLS to provide a more comprehensive solution compared to MLR or PCR.

The ability of PLS to handle collinearity among variables, a common challenge in spectral data, is a significant advantage. Collinearity often leads to instability in traditional regression models, making PLS a more robust choice. Additionally, PLS is less sensitive to outliers compared to PCR, contributing to improved model reliability.

In summary, the selection of PLS is driven by its capability to effectively model the intricate relationships between spectral data and substance concentration. Its versatility in handling collinear and high-dimensional data, along with its robustness to outliers, positions PLS as the most suitable method for achieving accurate and reliable predictions in your specific analytical context.

Chapter 5

Laboratory Experiments and Results Analysis: Aqueous Pigmented Solutions and Culture Medium Samples

This chapter focuses on fundamental preliminary experiments conducted using aqueous solutions containing colorants with significant spectra in the visible range. These experiments serve as an essential prelude to understanding and interpreting subsequent laboratory analyses, which center on evaluating the concentrations of glucose and lactate produced by cells in culture.

The use of colorants provides an intriguing window into the dynamics of aqueous solutions, allowing for the tracking of qualitative and quantitative changes through visible color alterations. These carefully selected colorants exhibit absorption spectra that extend into the visible range, providing crucial information about the progression of the involved chemical reactions. This approach proves particularly useful for monitoring changes in the chemical properties of solutions and highlighting the presence or formation of specific compounds.

Additionally, the chapter will delve into the detailed execution of a laboratory experiment aimed at determining the concentrations of glucose and lactate within a cultured cellular system. The choice of these two metabolically crucial molecules reflects their significance in fundamental biochemical processes and cellular energy pathways. Analyzing the concentrations of glucose and lactate offers a comprehensive overview of the metabolic activity of the examined cells, allowing for the assessment of their physiological state and behavior in response to specific stimuli or environmental conditions.

Through the combination of experiments with colorants and biochemical analyses of cultured cells, this chapter aims to provide a complete and detailed picture of the methodologies used to explore chemical and metabolic dynamics in aqueous environments. The results obtained from these experiments will form the foundation for subsequent research phases, contributing to an in-depth understanding of the biological and chemical processes under investigation.

5.1 Instrumentation (Tecan Spectrometer)

During this initial phase of our research, the Infinite® 200 PRO microplate reader has played a pivotal role in conducting a meticulous analysis of absorbance, offering precise assessments of the concentrations of substances under investigation.

Recognized as a cornerstone instrument in our study, the Tecan Infinite 200 PRO operates with exceptional precision during absorbance measurements. In this mode, a tungsten lamp serves as the light source, emitting light across a specified range of wavelengths. This emitted light permeates through the samples housed in the microplate wells. Absorption measurements rely on the interaction between incident light and molecules within the sample, resulting in the absorption of specific wavelengths dictated by the unique characteristics of the molecules. The resulting absorbance spectra provide both quantitative and qualitative insights into the concentration and nature of the substances. The instrument's sophisticated optical system, featuring filters or monochromators, facilitates the precise selection of wavelengths for excitation and emission. Absorbance values are then detected by a sensitive photodetector, ensuring accurate quantification.

The Tecan Infinite 200 PRO's advanced temperature control system guarantees experiment stability, a critical factor for reactions sensitive to temperature fluctuations. Seamless integration of data acquisition and analysis into the user-friendly software enhances the efficiency of interpreting and visualizing absorbance data. This comprehensive approach to absorbance measurements underscores the Tecan Infinite 200 PRO's significance in elucidating the chemical composition and concentration of target molecules, thereby substantially contributing to the success and reliability of experimental outcomes in our research.

The Tecan Infinite 200 PRO microplate reader boasts an intuitive and user-friendly interface that significantly contributes to the efficiency and accessibility of our experimental endeavors. This interface serves as a central hub for experiment design, allowing for the easy specification of parameters such as measurement type, wavelength settings, and microplate layout. Real-time monitoring capabilities provide invaluable insights into the ongoing experiments, enabling prompt assessments of data quality. The temperature control feature, seamlessly

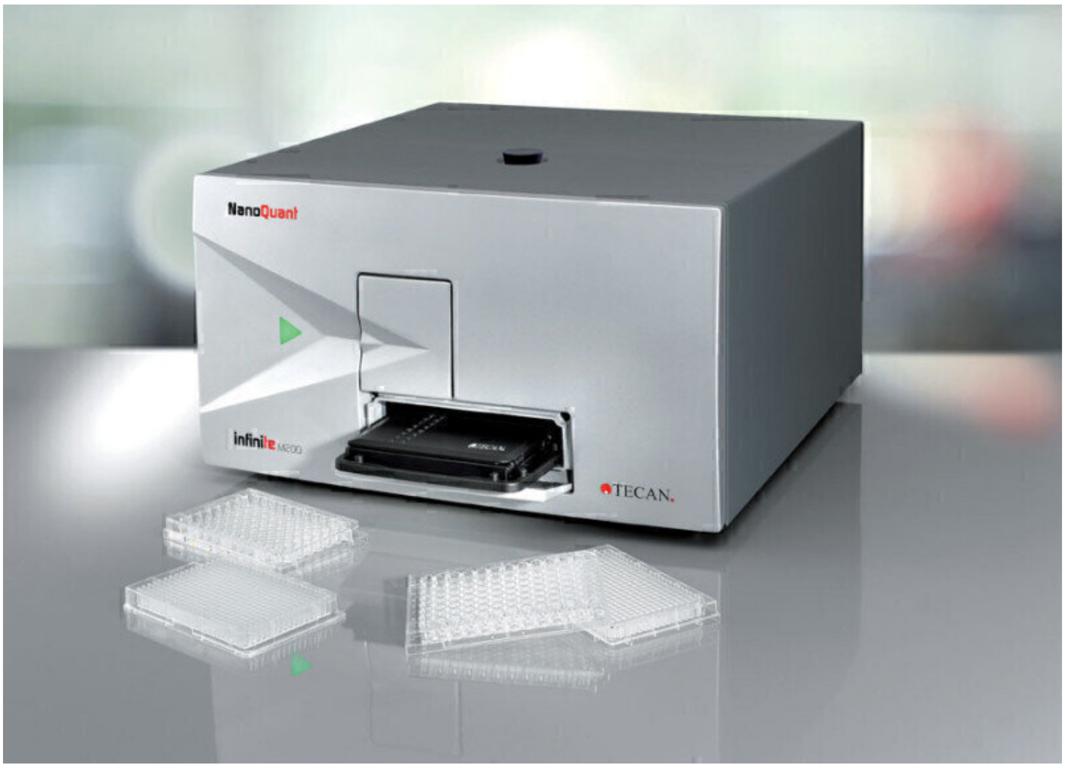


Figure 5.1: Tecan Infinite 200 PRO

integrated into the interface, allows for precise regulation and monitoring of experimental conditions. Ultimately, the Tecan Infinite 200 PRO's user interface plays a pivotal role in facilitating experiment execution, data analysis, and ensuring a seamless user experience for researchers across different proficiency levels. The figure 5.2 shows the graphical user interface of the tool to set up and start the analysis.

5.2 Preliminary Experiments

5.2.1 Materials Used

Used Dyes: Alizarin and PrestoBlue

The first dye used for laboratory testing is **Alizarin Red S**. It is a water-soluble sodium salt of Alizarin sulphonic acid with a chemical formula of $C_{14}H_7NaO_7S$. In the field of histology it is used to mark calcium deposits in tissue, while in geology it is used to identify minerals containing carbonate ions (CO_3^{2-}). The structure of alizarin presented in the figure 5.3 shows

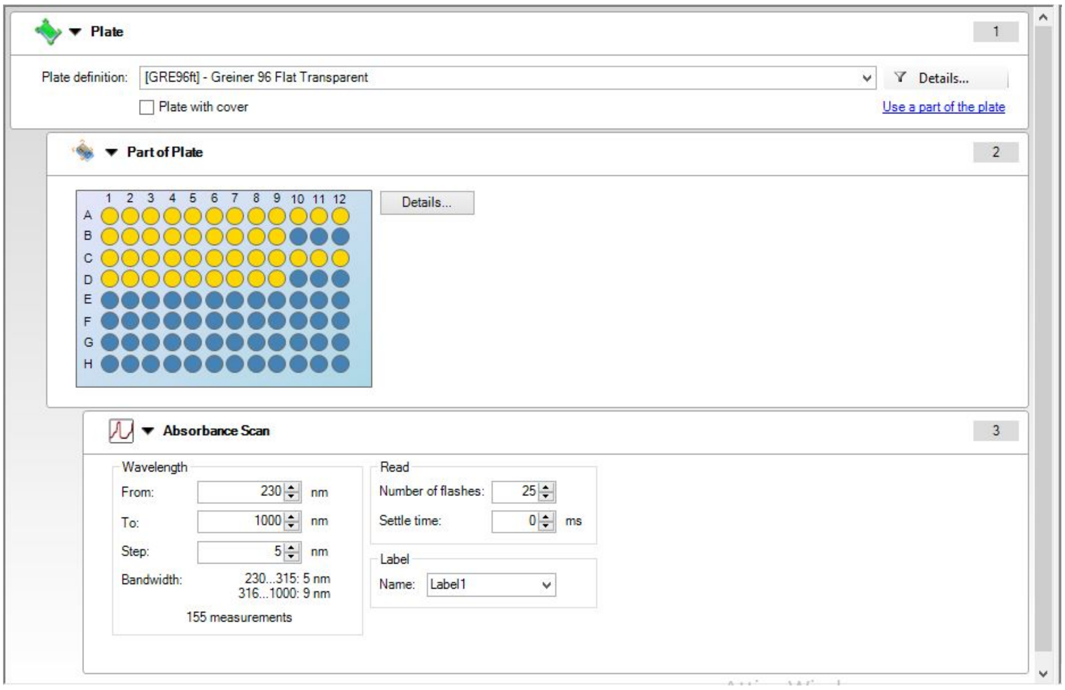


Plate 1

Plate definition: [GRE96ft] - Greiner 96 Flat Transparent

Details...

☐ Plate with cover

Use a part of the plate

Part of Plate 2

	1	2	3	4	5	6	7	8	9	10	11	12
A	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
B	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🔵	🔵	🔵
C	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡
D	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🟡	🔵	🔵	🔵
E	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵
F	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵
G	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵
H	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵	🔵

Details...

Absorbance Scan 3

Wavelength
From: 230 nm
To: 1000 nm
Step: 5 nm
Bandwidth: 230...315: 5 nm
316...1000: 9 nm
155 measurements

Read
Number of flashes: 25
Settle time: 0 ms

Label
Name: Label1

Figure 5.2: Graphic interface to set up and start the analysis

a 9,10-anthraquinone structure with two hydroxyl substituents at positions 1 and 2. They shift the energy absorption to the visible region and are responsible for the indirect binding of the molecule to a support. The groups responsible for the red colouring of the molecule are the two hydroxyls and the carbonyls, i.e. its chromophore groups. UV/visible spectroscopy of the molecule makes it possible to describe the microscopic properties responsible for the macroscopic properties, such as which transitions are responsible for the molecule's colouring. The spectrum shows two bands: the first is due to $\pi - \pi$ transitions in the 200-300 nm range and, being in the ultraviolet, makes no significant contribution to the colouration; the second, on the other hand, although also due to $\pi - \pi$ transitions, but being in the visible range at around 450 - 500 nm determines the colouration of the molecule in a sometimes peculiar manner. It is precisely the position of the latter band that leads to a significant variation in the molecule's colouration. The second dye we used was **PrestoBlue**. It has resazurin as an active ingredient, which is a non-fluorescent blue dye that can be reduced by metabolically active cells to a fluorescent pink product called resorufin. Specifically, the dye contains a redox indicator that changes colour in response to the metabolic activity of living cells. Viable cells with active metabolism reduce the PrestoBlue dye, resulting in a colour change or fluorescence. For this reason, PrestoBlue is used in the measurement of cell viability. The extent

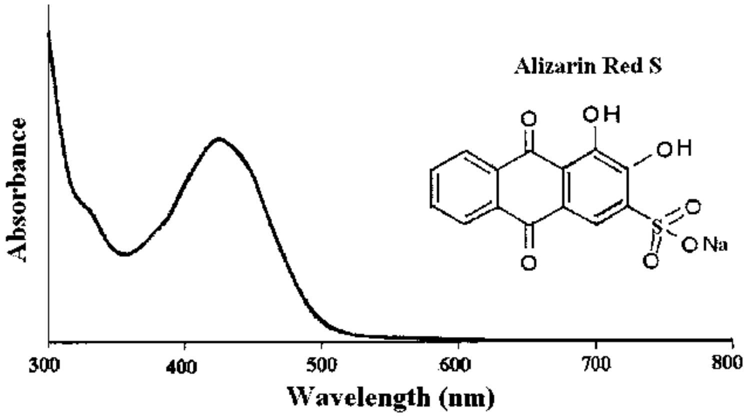


Figure 5.3: Chemical structure and Visible/UV absorption spectrum of ALizarin Red S dye [38]

of the colour change or fluorescence can be measured by spectrophotometry or fluorimetry, providing a quantitative assessment of cell viability. Researchers commonly use PrestoBlue in assays where they need to determine the effects of various treatments (such as drugs or experimental conditions) on cell viability. The test is relatively quick and convenient, making it a popular choice for high-throughput screening.

The chromophore properties of resazurin and its transformation to resorufin are key to its application in cell viability assays. Resazurin in its unaltered state has a maximum absorbance around 600-605 nm, resulting in a blue color. In the figure 5.4, we can observe the chemical structure and absorption spectrum of both resazaurin and resorufin.

5.2.2 First Experiment: Alizarin Solution

The first experiment we chose to conduct involved samples containing different concentrations of Alizarin. We chose to make 21 samples with different dilutions of alizarin and used laboratory micropipettes to dose them.

Specifically, we started with the mother substance, which we will call sample (1), placed inside a 1.5 mL laboratory tube with a solution consisting of 850 μL of water and 150 μL of Alizarin.

From mother tube (1), I take 750 μL of solution previously mixed with a shaker and place it in another 1.5 mL test tube with 250 μL of water inside. The new test tube will be number (2) and I mix the contents with the shaker. Now from test tube number two I take 750 μL and place it in test tube number (3), which contained 250 μL of water and so on.

This procedure will be repeated another 17 times until I have obtained the 21 predetermined

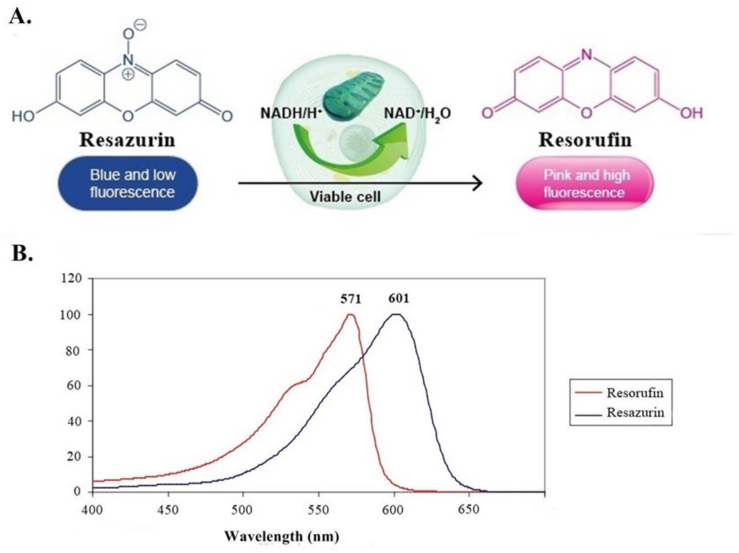


Figure 5.4: Chemical structure and Visible/UV absorption spectrum of PrestoBlue dye [14]

samples at different concentrations.

At this point, I take a 96-well plate. Inside the wells I put 100 μL of the solutions in the tubes. I need 100 μL otherwise I cannot cover the entire bottom of the well, compromising the measurement with the spectrometer that I am going to make.

Once I have loaded all my samples with diluted solutions, I go on to add an additional sample containing only water.

Now, the 96-well plate is ready and I insert it into the spectrometer and start the acquisitions. Specifically, acquisitions will be made at wavelengths between 200 nm and 1000 nm with a step size of 5 nm.

Our spectrometer will report all the results obtained in an Excel file from which the values that will make up our spectra will be extracted.

Once I have the file with all the data, I go and import it into MATLAB.

To do this, I use the following code:

```
% Load Excel file and specify sheet name
fileExcel = 'Spectra_20.xlsx';
sheet = 'Sheet0';
% Define cell ranges for spectrum data in Excel
gammacells = ["S30:EZ30", "S38:EZ38", "S46:EZ46", "S54:EZ54", "S62:EZ62",
    "S70:EZ70", "S78:EZ78", "S86:EZ86" ,...
               "S31:EZ31", "S39:EZ39", "S47:EZ47", "S55:EZ55", "S63:EZ63",
                   "S71:EZ71", "S79:EZ79", "S87:EZ87" ,...
```

```
                  ”S32:EZ32”, ”S40:EZ40”, ”S48:EZ48”, ”S56:EZ56”, ”S64:EZ64”];
% Initialize a matrix to store spectrum data
data = zeros(21, 138);
% Define wavelengths for plotting
wavelengths = [315:5:1000];
% Create a figure for plotting
figure
% Loop through each sample, load spectrum data, and plot
for i = 1:21
    % Load spectrum data from Excel into the data matrix
    data(i, :) = xlsread(fileExcel, sheet, gammacells(i))';
    % Plot the spectrum
    plot(wavelengths, data(i, :));
    % Add title and axis labels to the plot
    title('Spectra of the 21 samples');
    xlabel('Wavelengths (nm)');
    ylabel('Absorbance');
    % Hold the plot to overlay multiple spectra
    hold on
end
```

Listing 5.1: Code for importing and plotting data

In this way, I obtain the plot of all 21 spectra as shown in the image 5.5. At this point, to create the predictive model, I need to calculate the known concentrations. To do this, I need

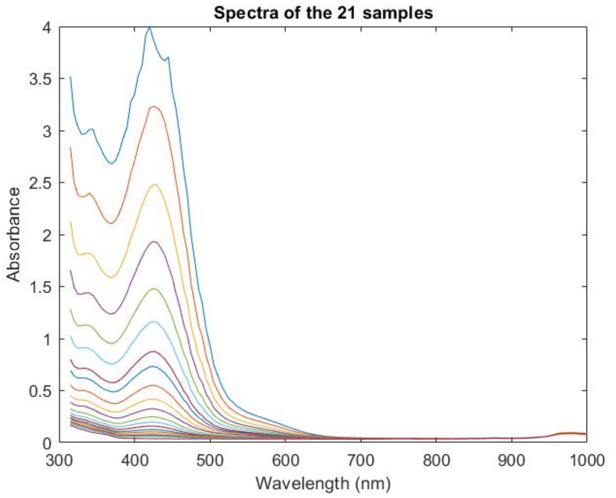


Figure 5.5: Spectrum of the 21 samples

to:

1. I calculate the concentration by doing:

$$Concentration_1 = \frac{150\mu L}{150\mu L + 850\mu L} \tag{5.1}$$

2. To calculate the concentration of following test tubes, I perform the following operation:

$$Concentration_i = \frac{Concentration_{i-1} Volume_{i-1}}{Volume_{water} + Volume_{i-1}} \tag{5.2}$$

   where $Concentration_i$ is the concentration of the test tube I am currently considering, $Concentration_{i-1}$ is the concentration of the solution I have taken from the 'previous' test tube and placed in the i-th tube which a priori contains a certain volume of water (250 μL), $Volume_{i-1}$ is the volume of the solution I have taken from the $i - 1$ test tube and added to the i-th tube.

This procedure is performed by the code 5.2

```
% Initial volume of Alizarin and water in microliters
Alizarin = 150; %[mcrL]
Water = 850; %[mcrL]
% Initialize an array to store known concentrations
Concentration_known = zeros(21, 1);
% Calculate concentration for the first sample
Concentration_known(1) = Alizarin / (Water + Alizarin);
% Calculate concentrations for the subsequent samples using a recursive formula
for i = 2:20
    % Concentration formula based on the volume of the previous solution added
    Concentration_known(i) = Concentration_known(i - 1) * 750 / 1000;
end
```

Listing 5.2: Code for calculating concentrations

Once all these steps have been performed, we are ready to perform PLS with MATLAB's PLS Toolbox.

The PLS Toolbox is a computational tool used to perform multivariate data analysis, in particular to deal with regression problems, dimensionality reduction and modelling complex relationships between variables.

PLS, as is already known, is a technique that can be particularly useful when working with complex datasets in which many variables are correlated or when there is a risk of multicollinearity.

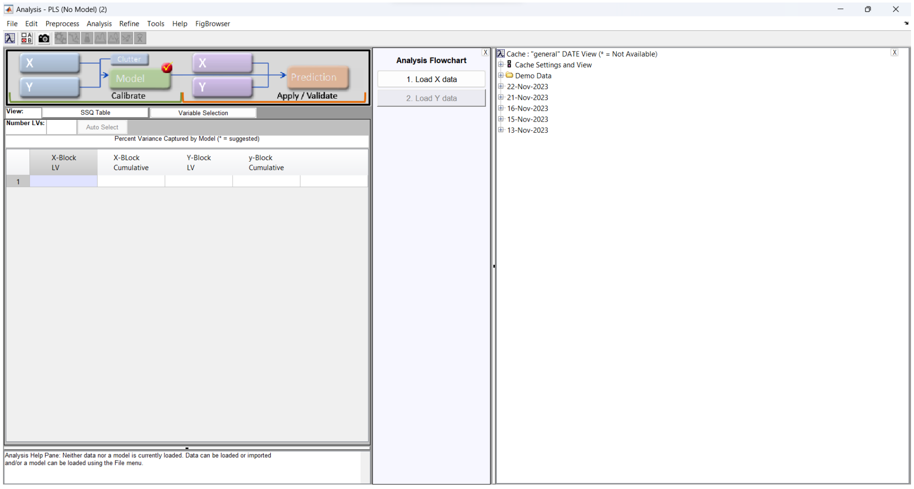
PLS works through a combination of dimensionality reduction and regression. The main objective is to establish a relationship between a set of independent variables X and a set of dependent variables Y.

In particular, PLS seeks to identify latent components (or factors) that are linear combinations of both sets of variables.

To start the Toolbox, we have to type in the MATLAB Command Window the ”*browse*” command, which, once executed, will open a new window.

Within the Toolbox it is possible to follow different types of data analysis and consequently it is necessary to select the one we are interested in, namely PLS - Partial Least Square. Once I have selected the analysis that is right for us, a window like the one shown in the figure 5.6 will open.

Now, first I have to load the data correctly. Specifically, the matrix containing all the spectra will be placed in the calibration X's, while in the calibration Y's I will place the matrix with the values of the known concentrations. Once this is done, the preprocessing is set to ”*Autoscale*” mode by default and we leave it exactly that way.



Percent Variance Captured by Model (* = suggested)

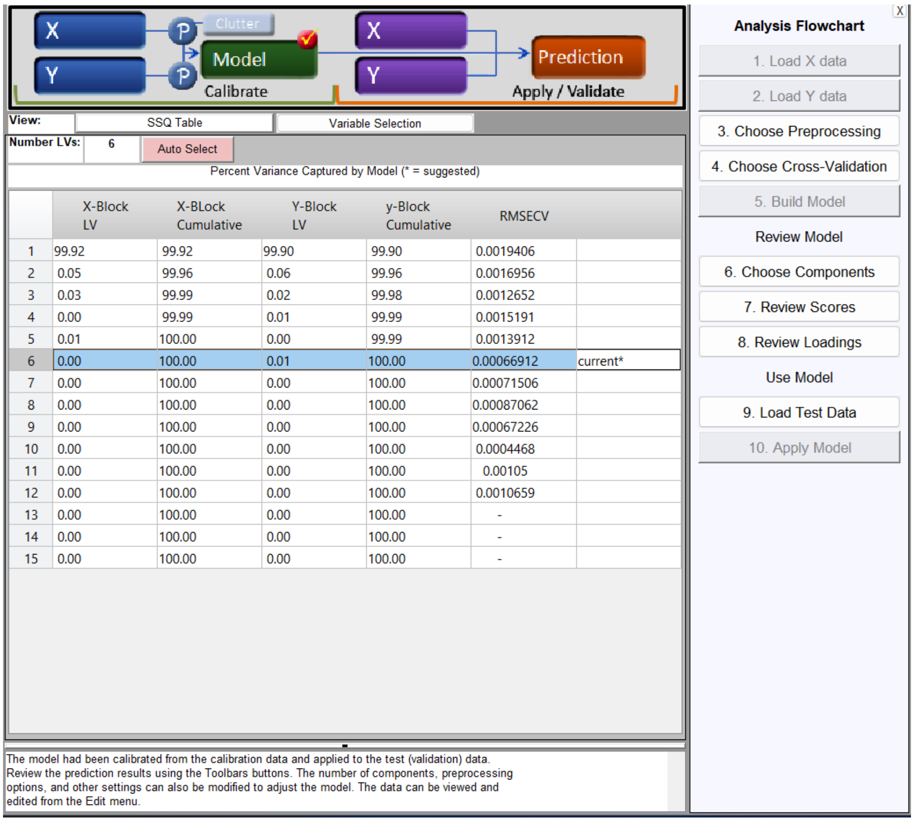
	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	
1					

Analysis Help Pane: Neither data nor a model is currently loaded. Data can be loaded or imported and/or a model can be loaded using the File menu.

Figure 5.6: Toolbox screen view

At this point, I go to select from the spectra provided those that will constitute the spectra to be used for the claibration of the model and those that I am going to use for the validation of the model. To do this, I right-click on the validation X box and select the command ”Split into

Calibration / Validation”. I randomly select 6 spectra that we will use for validation, while the remaining ones will be the spectra for calibration.
I run the programme to obtain the model. The next step consists of a series of procedures



Percent Variance Captured by Model (* = suggested)

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV	
1	99.92	99.92	99.90	99.90	0.0019406	
2	0.05	99.96	0.06	99.96	0.0016956	
3	0.03	99.99	0.02	99.98	0.0012652	
4	0.00	99.99	0.01	99.99	0.0015191	
5	0.01	100.00	0.00	99.99	0.0013912	
6	0.00	100.00	0.01	100.00	0.00066912	current*
7	0.00	100.00	0.00	100.00	0.00071506	
8	0.00	100.00	0.00	100.00	0.00087062	
9	0.00	100.00	0.00	100.00	0.00067226	
10	0.00	100.00	0.00	100.00	0.0004468	
11	0.00	100.00	0.00	100.00	0.00105	
12	0.00	100.00	0.00	100.00	0.0010659	
13	0.00	100.00	0.00	100.00	-	
14	0.00	100.00	0.00	100.00	-	
15	0.00	100.00	0.00	100.00	-	

The model had been calibrated from the calibration data and applied to the test (validation) data. Review the prediction results using the Toolbars buttons. The number of components, preprocessing options, and other settings can be modified to adjust the model. The data can be viewed and edited from the Edit menu.

Figure 5.7: Screenshot of results obtained at the end of the analysis

that are implemented in order to improve the predictive capability of the model we are considering. First, I go to the ”Variable selection” section and click on ”Execute”. By doing this, the programme searches within the spectra that I have provided to it, for the portions with the highest variability. Once the analysis is finished, I click on ”Use” so that to calculate the model, the toolbox uses those portions of the spectrum more predominantly.

PLS Analysis Results for Alizarin Solution

When evaluating the effectiveness of a predictive model, a key parameter often considered is the RMSE, which stands for Root Mean Square Error. The RMSE is a crucial metric in assessing the accuracy of a predictive model compared to observed data. Its expression in dimensionless values makes it versatile, as it is not influenced by the original unit of measurement of the data.

When one wishes to express the error in percentage, a significant approach involves comparing the RMSE with the maximum value present in the data. This calculation provides an indication of the relative error compared to the maximum observed value, offering a more intuitive understanding of the overall accuracy of the model.

In this way, the analysis of percentage error takes into account the maximum scale of measurements, allowing for an evaluation of the impact of the error proportionally to the maximum range of the data. Such an approach contributes to a more comprehensive assessment of the model's performance, providing valuable insights for improvement and optimization of predictions.

Now I have to run the programme again in order to obtain the model and I get the screen shown in the figure 5.7. I select the number of latent variables that shows a lower RMSECV, which in our case is 6.

At this point, it is very interesting to visualise the graph that relates the measured Ys reported on the X-axis and the predicted Ys reported on the Y-axis. This relationship is shown in the figure 5.8

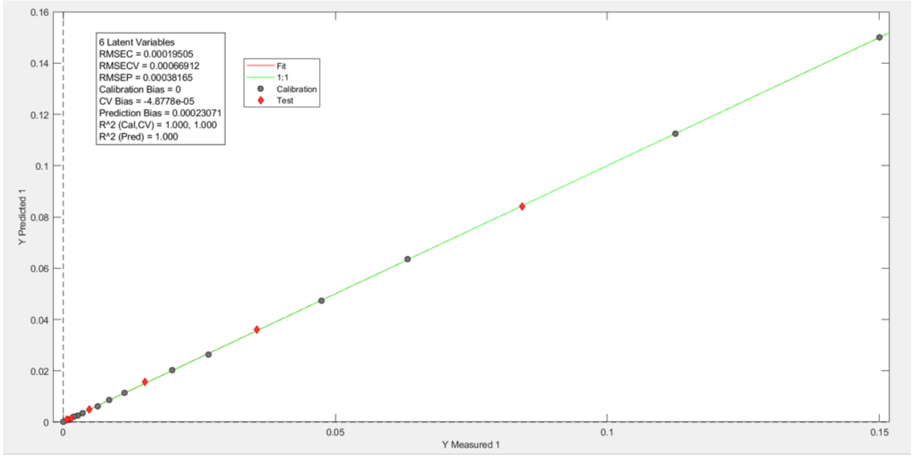


Figure 5.8: ”Screenshot depicting the results obtained at the end of the analysis through PLS application for estimating concentrations of Red Alizarin in the samples. The graph illustrates the correlation between predicted concentrations (y-axis) and actual measured concentrations (x-axis).”

The graph also shows RMSEC (Root Mean Squared Error of Calibration), RMSECV (Root Mean Squared Error of Cross-Validation) and RMSEP (Root Mean Squared Error of Prediction).

The RMSEC represents the root mean squared error between observed and predicted values using the model on the calibration data, i.e. it measures the accuracy of the model in reproducing the data used to build it. Being very low, this indicates a good fit of the model to the calibration data.

Next comes the RMSECV (Root Mean Squared Error of Cross-Validation). This value represents the root mean squared error between the observed values and the values predicted using the model during the cross-validation process. Again, if this value is rather low, it indicates a greater capacity for generalisation of the model, which is certainly a positive aspect.

We also have the RMSEP (Root Mean Squared Error of Prediction), which represents the root mean square error between the observed values and the values predicted using the model on an independent prediction dataset. It therefore measures the model's ability to make accurate predictions on new data not used in the calibration. It too is rather low, indicating a higher accuracy in predicting new data.

Looking at all three parameters, we see that they are all low, so it can be affirmed that the model has a good ability to fit the calibration data, generalise to new data and make accurate predictions on independent data.

However, it is also important to consider other performance indicators, such as R^2, to obtain a complete view of the model's performance. Values of R^2 (coefficient of determination) are statistical measures that provide information on the quality and fit of a regression model High R^2_{Cal}, R^2_{CV} and R^2_{Pred} values suggest a very well-fitted model, capable of explaining and generalising the data well.

5.2.3 Second Experiment: PrestoBlue Solution

In the same way, tests are carried out with the PrestoBlue. So, we prepared 20 samples with distinct dilutions of PrestoBlue, employing laboratory micropipettes for precise dosing. The procedure commenced with the creation of the primary solution, denoted as sample (1), housed in a 1.5 mL laboratory tube. This solution comprised 850 µL of water and 150 µL of PrestoBlue. Subsequently, from the mother tube (1), 750 µL of the solution, previously mixed using a shaker, was transferred to another 1.5 mL test tube labeled as number (2), containing 250 µL of water. The contents were thoroughly mixed. This process was iteratively repeated 17 times, with each subsequent test tube receiving 750 µL from the preceding tube and 250 µL of water. This sequence yielded the 20 predetermined samples with varying concentrations.

Next, a 96-well plate was employed, and each well was filled with 100 µL of the solutions from the respective tubes. The 100 µL volume was crucial to ensure complete coverage of the well bottom, essential for accurate spectrometer measurements. Following the loading of all samples, an additional well containing only water was added. The 96-well plate was then inserted into the spectrometer, and acquisitions commenced. The measurements were specifically conducted at wavelengths ranging from 200 nm to 1000 nm, with a step size of 5 nm. The resulting spectrometer data were recorded in an Excel file, from which the values constituting our spectra were extracted. Subsequently, the data were imported into MATLAB using a custom code. The concentrations of PrestoBlue in the samples were calculated using specific formulae, providing known concentrations for subsequent multivariate data analysis.

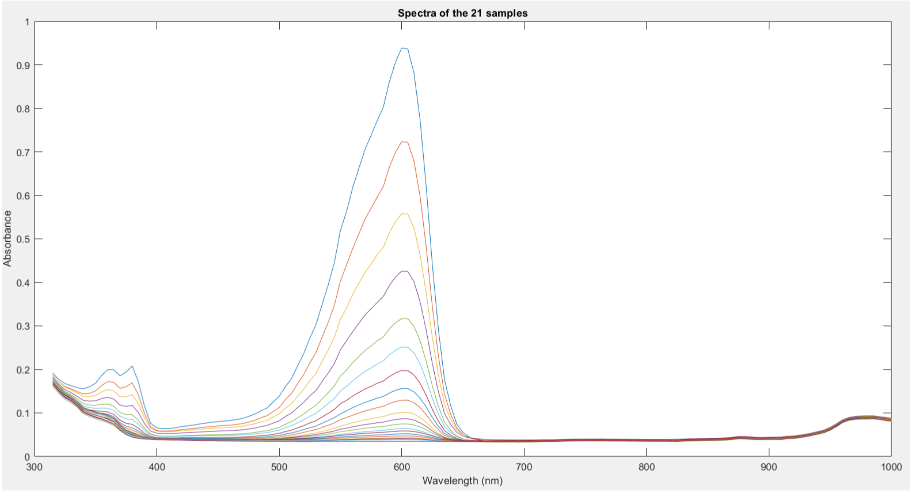


Figure 5.9: Spectrum of the 21 samples of solutions with different concentration of PrestoBlue

PLS Analysis Results for PrestoBlue Solution

For the predictive modeling phase, we employed the MATLAB PLS Toolbox. The Toolbox facilitated the exploration of complex relationships between the spectra and the known concentrations of PrestoBlue. Utilizing partial least squares regression, we aimed to establish a robust model capable of predicting concentrations accurately. The resulting model underwent optimization steps, including variable selection to enhance its predictive capability. The evaluation of the model's performance, as depicted in figures and statistical metrics, indicated its efficacy in fitting the calibration data, generalizing to new data, and making accurate predictions on independent datasets. The high values of R-squared further affirmed the model's quality and suitability for the analysis of PrestoBlue concentrations.

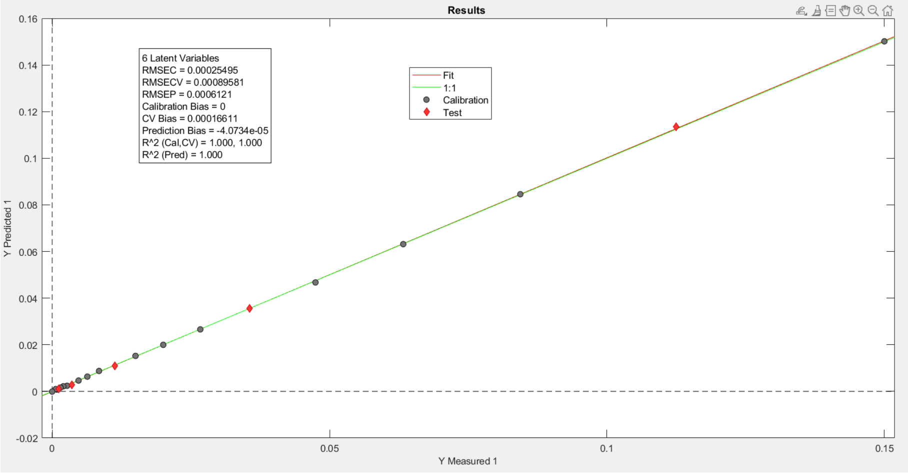


Figure 5.10: Screenshot depicting the results obtained at the end of the analysis through PLS application for estimating concentrations of Prestoblue in the samples. The graph illustrates the correlation between predicted concentrations (y-axis) and actual measured concentrations (x-axis).

5.2.4 Third Experiment: Alizarin and PrestoBlue Solution

The second experiment involved the preparation of samples using a mother solution composed of 150 μL of Alizarin and 150 μL of PrestoBlue. Similar to the first experiment, we aimed to create 20 samples with varying concentrations by diluting the mother solution. The procedure began with sample (1) placed in a 1.5 mL laboratory tube with a solution of 850 μL of water and 150 μL of the Alizarin-PrestoBlue mixture.

Sequentially, 750 μL of the solution from each tube was transferred to the next tube containing 250 μL of water. This process was repeated 17 more times to obtain the 20 samples with different concentrations. Subsequently, 100 μL of each sample solution was dispensed into wells of a 96-well plate, ensuring full coverage of the well bottom for accurate spectrometer measurements.

After loading all samples, a pure water sample was added to the plate. The plate was then inserted into the spectrometer, and acquisitions were carried out across wavelengths ranging from 200 nm to 1000 nm with a step size of 5 nm. The resulting data were exported to an Excel file for further processing.

Once I have the file with all the data, I go and import it into MATLAB.

In the provided MATLAB code (5.3), we initiate the process of importing and visualizing spectral data for a solution containing PrestoBlue and Alizarin. The Excel file and sheet containing the spectral data are specified, and cell ranges for each sample are defined. Notably, due to saturation in the peak of the first two samples, only 19 spectra are utilized. Additionally, to mitigate saturation issues, the wavelength range starts from 365 nm rather than 200 nm. The script then proceeds to plot individual spectra for each sample, overlaying them for a comprehensive view. Concentrations of Alizarin and PrestoBlue in the mother solution are calculated iteratively using prescribed formulas, considering the dilution process. The resulting concentrations are stored in arrays for further analysis.

```
% Specifying the Excel file and sheet containing spectral data
fileExcel = 'Spettri_PrestoBlue_Soluzione_per_PLSII.xlsx';
sheet = 'Sheet0';
% Defining the cell ranges containing spectral data for each sample
gammacells = [”AC53:EZ53”, ”AC54:EZ54”, ”AC55:EZ55”, ”AC56:EZ56”, ”AC57:EZ57”, ”AC58:EZ58”, ”AC59:EZ59”, ”AC60:EZ60”, ”AC61:EZ61”,”AC62:EZ62”, ”AC63:EZ63”, ”AC64:EZ64”, ”AC65:EZ65”, ”AC66:EZ66”,”AC67:EZ67”,”AC68:EZ68”,”AC69:EZ69”,”AC70:EZ70”,”AC71:EZ71”];
% Initializing matrices to store spectral data and defining wavelength range
data = zeros(19,128);
wavelengths = [365:5:1000];
% Plotting individual spectra for each sample
figure
for i = 1:19
    % Reading and transposing spectral data from Excel
    data(i,:) = xlsread(fileExcel, sheet, gammacells(i))';
    % Plotting the spectrum
    plot(wavelengths, dati(i, :));
    % Adding labels and title to the plot
    title('Spectra of the 19 samples (PrestoBlue and Alizarin Mixture)');
    xlabel('Wavelength (nm)');
    ylabel('Absorbance');
    % Holding the plot for overlaying multiple spectra
    hold on
end
% Initial concentrations of Alizarin, PrestoBlue, and Water in the mother solution
Presto_Blue = 150; %[mcrL]
Alizarin = 150;
Water = 700; %[mcrL]
% Initializing arrays to store calculated concentrations
Conc_k_A = zeros(19,1);
Conc_k_P = zeros(19,1);
```

```
% Calculating concentrations using the provided formulas
Conc_k_A(1) = Alizarin/1000;
Conc_k_P(1) = Presto_Blue/1000;
for i = 2:19
    Conc_k_A(i) = Conc_k_A(i-1)*750/1000;
    Conc_k_P(i) = Conc_k_P(i-1)*750/1000;
end
% Combining concentrations of Alizarin and PrestoBlue
Conc_k = [Conc_k_A Conc_k_P];
```

Listing 5.3: Code for Importing and Plotting Spectral Data for PrestoBlue and Alizarin Mixture

The plots of the 19 spectra we took into account for the creation and validation of the model are shown in the 5.11.

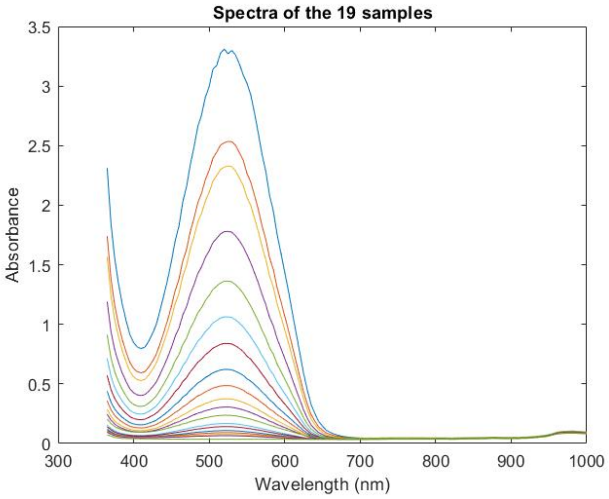
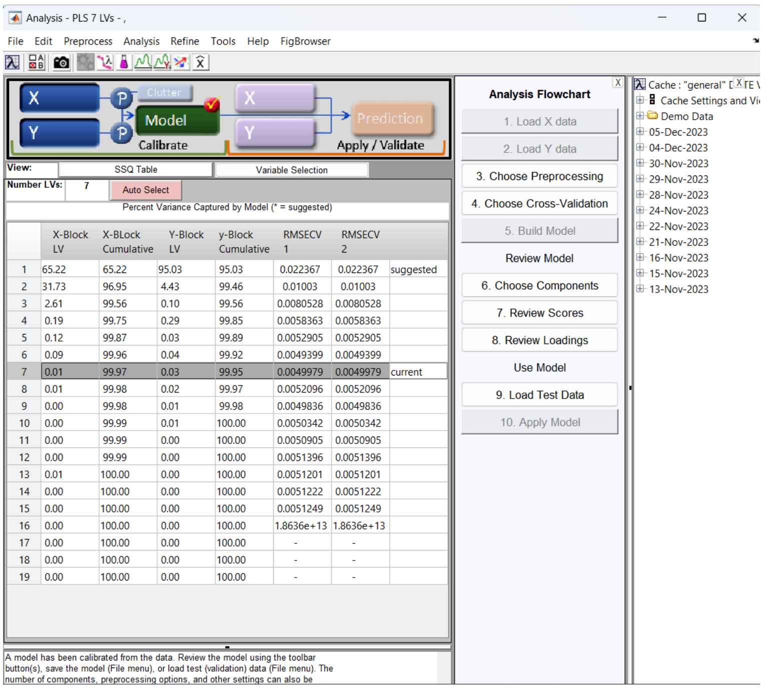


Figure 5.11: Spectrum of the 19 samples of solutions with different concentration of PrestoBlue and Alizarin mixture

PLS Analysis Results for Alizarin and PrestoBlue Solution

Initially, we imported the ”data” matrix containing all 19 spectra into the X calibration block, excluding two initially due to saturation issues. Subsequently, we imported the matrix containing the concentration values calculated using the specified formulas into the Y calibration block. With these datasets, we proceeded to create the model.

Upon reviewing the emerged values, it was observed that the optimal RMSECV occurred in the vicinity of 7 latent variables. Consequently, we recreated the model with these specifications to maximize predictive accuracy. Following this, we examined the model's scores.



View: SSQ Table | Variable Selection

Number LVs: 7 Auto Select

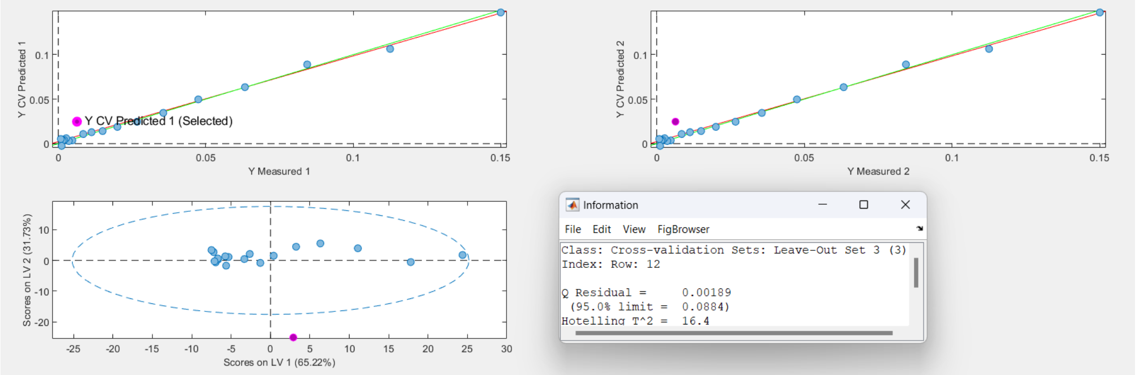
Percent Variance Captured by Model (* = suggested)

	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV 1	RMSECV 2	
1	65.22	65.22	95.03	95.03	0.022367	0.022367	suggested
2	31.73	96.95	4.43	99.46	0.01003	0.01003	
3	2.61	99.56	0.10	99.56	0.0080528	0.0080528	
4	0.19	99.75	0.29	99.85	0.0058363	0.0058363	
5	0.12	99.87	0.03	99.89	0.0052905	0.0052905	
6	0.09	99.96	0.04	99.92	0.0049399	0.0049399	
7	0.01	99.97	0.03	99.95	0.0049979	0.0049979	current
8	0.01	99.98	0.02	99.97	0.0052096	0.0052096	
9	0.00	99.98	0.01	99.98	0.0049836	0.0049836	
10	0.00	99.99	0.01	100.00	0.0050342	0.0050342	
11	0.00	99.99	0.00	100.00	0.0050905	0.0050905	
12	0.00	99.99	0.00	100.00	0.0051396	0.0051396	
13	0.01	100.00	0.00	100.00	0.0051201	0.0051201	
14	0.00	100.00	0.00	100.00	0.0051222	0.0051222	
15	0.00	100.00	0.00	100.00	0.0051249	0.0051249	
16	0.00	100.00	0.00	100.00	1.8636e+13	1.8636e+13	
17	0.00	100.00	0.00	100.00	-	-	
18	0.00	100.00	0.00	100.00	-	-	
19	0.00	100.00	0.00	100.00	-	-	

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be

Figure 5.12: Graph illustrating model parameters obtained with seven latent variables before dataset split into calibration and validation sets.

The scores plot promptly exposed a data point positioned outside the ellipse, suggesting the existence of an outlier. Acknowledging the possible repercussions of this outlier on model development, we identified it as the sample corresponding to row number 12 and emphasized its presence. Subsequently, we proceeded to remove this outlier. This strategic step not only bolstered the overall quality of the model but also heightened its predictive precision by mitigating potential adverse effects stemming from anomalous data points. The elimination of such anomalies augments the model's capability to discern meaningful patterns and relationships within the dataset, ultimately resulting in a more dependable and accurate predictive model. After cleaning and addressing outliers in the dataset, the next crucial step involved partitioning the data. Seventy percent was allocated for calibration purposes, while the remaining 30% was reserved for model validation. This deliberate split not only ensures the model's robustness but also facilitates its evaluation on previously unseen data. Following the dataset organization, the analysis of RMSEP trends played a pivotal role in influencing our strategic decision-making process. Notably, we observed a discernible trend during the ex-



```
Class: Cross-validation Sets: Leave-Out Set 3 (3)
Index: Row: 12

Q Residual =      0.00189
 (95.0% limit =  0.0884)
Hotelling T^2 =  16.4
```

Figure 5.13: Scores plot revealing outliers in the dataset. The graph illustrates the distribution of data points in the latent variable space, where a distinct point lies outside the typical pattern indicated by the ellipses. This outlier, marked for visibility, could potentially impact the model's robustness and predictive accuracy. Identifying and addressing outliers is crucial for refining the model and improving its overall performance.

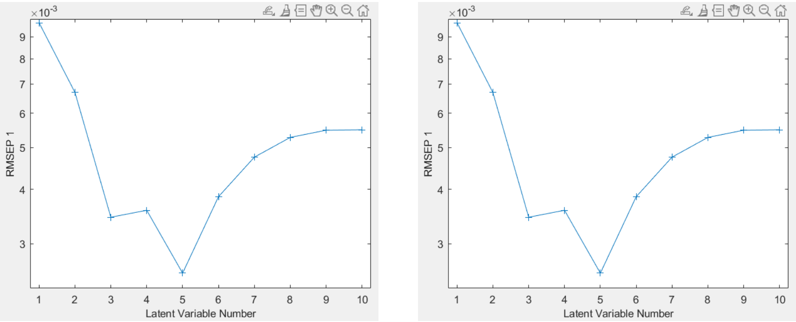
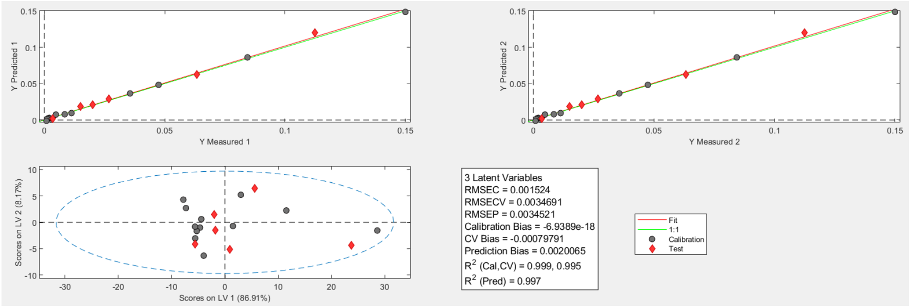


Figure 5.14: a) RMSEP for the substance 1; b) RMSEP for the substance 2

amination of RMSEP for varying numbers of latent variables. The reduction in RMSEP was particularly pronounced when transitioning from two to three latent variables. After careful consideration, we concluded that employing three latent variables yielded a more significant improvement in predictive accuracy compared to other configurations. This strategic choice aimed at achieving a balance between model complexity and performance. The subsequent execution of the model using the cleaned and divided dataset further enhances its ability to generalize and make accurate predictions on new and independent data.

In this specific analysis, unlike the two previous experiences, we found ourselves unable to use the ”Variable Selection” tool. This limitation arose because our Y dataset includes more than one column. The variable selection tool is designed to handle only one dependent variable at a time. Therefore, in addressing this situation, we adapted to the peculiarity of the Y matrix and proceeded without the use of this tool. In 5.15, we observe that the parame-



3 Latent Variables
RMSEC = 0.001524
RMSECV = 0.0034691
RMSEP = 0.0034521
Calibration Bias = -6.9389e-18
CV Bias = -0.00079791
Prediction Bias = 0.0020065
R^2 (Cal,CV) = 0.999, 0.995
R^2 (Pred) = 0.997

Figure 5.15: Graphical representation of model results with three plots: the first illustrates the relationship between known Y values and model-predicted Y values for the first substance, while the second plot depicts the same relationship for the second substance. The third plot displays the scores, which, in this specific case, are absent as all data points lie within the ellipse. The absence of outliers in the scores plot underscores the robustness of the model's predictions for both substances.

ters resulting from this analysis offer a detailed assessment of the model's performance. The mean values for the Y columns approach 0.0401802, indicating accurate overall concentration predictions. The low standard deviation (0.0408066) underscores the model's consistency in prediction, with data showing limited dispersion around the mean. Error indicators, such as RMSEC (0.00152397), RMSECV (0.00346906), and RMSEP (0.00345212), highlight notable precision in predicting concentrations during both calibration and on independent data. The absence of systematic trends is confirmed by values close to zero for Bias, CV Bias, and Pred Bias. Additionally, the high R^2 values (0.998797, 0.995451, 0.997045) indicate a strong correlation between known and predicted Y values during calibration, cross-validation, and prediction on independent data.

Overall, these results suggest that the model is accurate, well-fitted, and capable of reliably generalizing to new data. These parameters provide a solid foundation for our understanding of the relationship between the considered variables and constitute a valuable resource for future analyses within the scope of this research. In addition to the analysis of the model's statistical parameters, a scatter plot was generated to further evaluate the quality of the predicted data in comparison to the measured concentrations. This MATLAB code creates a visual representation of the relationship between the measured concentrations of substances 1 and 2 (Y_{m_1} and Y_{m_2}) and the corresponding predicted concentrations (Y_{p_1} and Y_{p_2}). The plot displays each concentration pair, using circles for measured data and asterisks for predicted data, with a clear legend for distinction. The x-axis represents the measured concentrations of substance 1, while the y-axis represents the measured concentrations of substance 2. This visualization aids in assessing how well the predicted concentrations align with the actual

```
% Load data from the provided MATLAB file
load('Data_Presto_Alizarin.mat')
% Separate the data into measured and predicted values for substances 1 and 2
Y_m_1 = Data_Presto_Alizarin(:, 1); % Measured data for substance 1
Y_m_2 = Data_Presto_Alizarin(:, 2); % Measured data for substance 2
Y_p_1 = Data_Presto_Alizarin(:, 3); % Predicted data for substance 1
Y_p_2 = Data_Presto_Alizarin(:, 4); % Predicted data for substance 2
% Create a figure to plot the measured and predicted data
figure
plot(Y_m_1, Y_m_2, 'o') % Scatter plot for measured data
hold on
plot(Y_p_1, Y_p_2, '*') % Scatter plot for predicted data
% Add legend and labels to the plot
legend(”Measured data”, ”Predicted data”)
xlabel(”Measured Concentration - Substance 1”)
ylabel(”Measured Concentration - Substance 2”)
title(”Measured vs Predicted Concentrations for Substances 1 and 2”)
```

Listing 5.4: Scatter plot illustrating the relationship between measured, known, and predicted concentrations for substances 1 and 2.

measured concentrations for both substances, providing valuable insights into the model's predictive accuracy.

5.3 Experiment on Glucose and Lactate Detection in a Culture Medium: Preliminary Phase of Experimentation

In the context of research aimed at the precise determination of concentrations of specific substances, such as glucose and lactate, the experiment was extended to the analysis of a culture medium present in the laboratory. The main objective was to assess the concentrations of these compounds within the sample, utilizing visible spectra acquired through the Tecan Infinite® 200 PRO spectrometer. The use of this advanced technology allowed for detailed spectra, providing crucial insights into the spectral characteristics of the target substances and significantly contributing to the understanding of the chemical dynamics of the examined culture medium.

5.3.1 Culture Medium

In this experimental setup, the culture medium played a crucial role, with a focus on three distinct formulations: MO, M1, and M2. The baseline composition, MO, comprised α**-MEM**

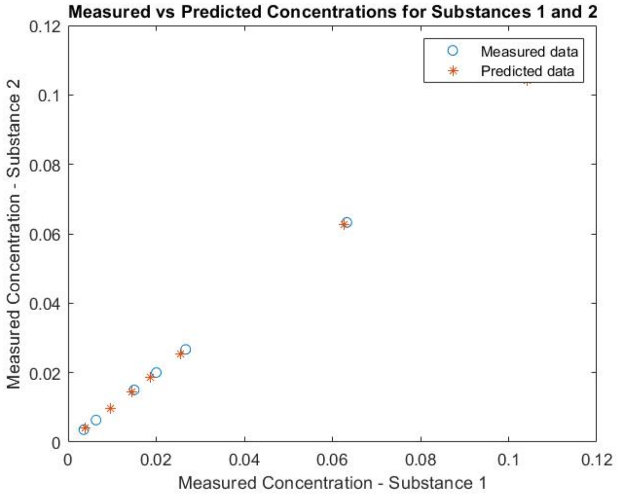


Figure 5.16: Figure displaying the measured concentrations of substances 1 and 2 compared to their predicted concentrations. Substance number one corresponds to Red Alizarin, while substance number two corresponds to PrestoBlue. Each point represents a concentration pair, with circles denoting measured data and asterisks denoting predicted data. The plot provides a visual assessment of the modelś predictive accuracy.

(Minimum Essential Medium), FBS (Fetal Bovine Serum), and **1% of Penicillin/Streptomycin**. The M1 formulation introduced 0.2 mM of **Ascorbic Acid 2-Phosphate** and 10 mM of **Glycerol 2-Phosphate**, enhancing antioxidant and metabolic aspects. M2, building upon M1, incorporated an additional component, 50 nM of **Melatonin**, known for its role in circadian rhythm regulation.

For experimental conditions, MO was exclusively used for *"Static" (S)* and *"Not conditioned" (NC)* cell cultures. In contrast, M1 and M2 formulations were applied in *"Conditioned" (C)* culture experiments. These variations aimed to create diverse microenvironments influencing concentrations of target substances like glucose and lactate within samples. Spectroscopic data from the Tecan Infinite® 200 PRO spectrometer were utilized to analyze and predict substance concentrations in the respective culture conditions.

The M1 composition, enriched with antioxidants and metabolic support, included Ascorbic Acid 2-Phosphate and Glycerol 2-Phosphate. The former, a derivative of ascorbic acid, contributes antioxidant properties to maintain a reducing environment. Glycerol 2-Phosphate serves as a source of glycerol and phosphate, supporting cellular energy metabolism and various biochemical processes.

M2, an extension of M1, introduced Melatonin (50 nM), a hormone involved in circadian rhythm regulation. In cell culture, melatonin may influence various cellular processes, potentially impacting metabolic activities. In summary, these variations in the culture medium aim to create nuanced environments affecting substance concentrations, providing a comprehensive understanding of experimental outcomes.

At this point, the cell cultures in the C and NC states underwent dynamic culture using a perfusion bioreactor, while cultures in the S state were treated with static culture. Static and dynamic cultures represent two distinct approaches in manipulating cellular growth conditions, each with unique characteristics significantly influencing cellular responses. In static culture, cells are maintained in a stationary environment, devoid of flow or agitation of the culture medium. This method is preferred when replicating uniform and constant growth conditions, similar to those cells may experience in stationary physiological contexts. Static cultures are often used to examine specific phenomena without the interference of dynamic variables.

On the other hand, dynamic cultures are conducted in bioreactors, systems where the culture medium is constantly stirred or flowing. This approach aims to reproduce a more dynamic environment, similar to physiological conditions where cells are exposed to nutrient flows and environmental changes. The continuous stirring or flow in bioreactors contributes to improving the uniform distribution of nutrients and oxygen, also promoting efficient removal of waste products. Dynamic cultures are particularly useful when simulating more realistic conditions, such as those occurring in biological tissues or tissue engineering contexts.

The key differences between the two methodologies lie in the static nature of the environment in static culture and the dynamism introduced by constant stirring or flow in dynamic culture. The choice between these two approaches depends on the specific goals of the experiment and the need to replicate the most relevant conditions for the ongoing study. Both methodologies play a crucial role in biological research and tissue engineering, providing fundamental tools to understand cellular dynamics in both static and dynamic contexts.

5.3.2 PLS Analysis for Metabolite Concentration Prediction

This initial experiment laid the groundwork for subsequent trials with an increased number of acquisitions, reaching a total of 63. The preliminary nature of this experiment allowed for the refinement of acquisition parameters and the establishment of a robust foundation for more extensive data collection.

For this experimental phase, given the uncertainty surrounding the outcomes and the nature of the spectra, acquisitions were conducted for samples labeled as C (Conditioned), NC (Non-Conditioned), and S (Static) on days 0, 10, and 21. In previous experiments, a greater number of wavelengths were excluded due to a better understanding of the spectral behavior of the solution. Multiple iterations were performed, and optimal results were consistently achieved by excluding wavelengths exhibiting saturation, particularly at 230 nm. Consequently, the spectrum importation process focused on the range from 235 nm to 1000 nm.

Importation of Absorption Spectra

The provided MATLAB code plays a crucial role in the analysis of absorbance spectra derived from a set of gamma cells within a scientific experiment. Its primary objective is to extract, process, and visualize absorbance data as it varies with wavelength. The code handles the extraction of absorbance spectra from nine distinct gamma cells, organizes the data into a matrix, and generates meaningful plots.

In the initial phase, specific gamma cells are identified and defined through the ‘gammacells‘ variable. This variable holds the cell ranges in the Excel sheet corresponding to each gamma cell. Subsequently, a matrix named ‘data‘ is initialized to store absorbance data, and a wavelength range is defined for the absorbance spectra. A figure is created to visually represent the spectra, and a loop is employed to extract and visualize absorbance spectra for each gamma cell. The resulting plots provide valuable insights into the characteristics of the absorbance data.

The code also includes the definition of concentrations for glucose ($Conc\_k\_glu$) and lactate ($Conc\_k\_lac$) for each gamma cell. hese concentrations are crucial for training the subsequent machine learning algorithm, allowing it to learn the relationship between absorbance spectra and concentration levels.

After training, the model can predict concentrations based on new, unseen absorbance spectra. This prediction capability is invaluable in the validation phase, where the model's accuracy and generalization to unseen data are assessed. The concentrations, organized in the matrix $Conc\_k$, play a pivotal role in both the training and validation steps, contributing to the robustness and reliability of the machine learning algorithm for absorbance data analysis.

In summary, this MATLAB code represents a pivotal component of the data analysis pipeline, facilitating a detailed examination of absorbance spectra and associated concentration variations across different gamma cells. The generated plots offer a visual representation of spectral characteristics, while the defined concentrations contribute to a comprehensive understanding of the experimental conditions.

```
% Definition of gamma cells from which to extract absorbance spectra
gammacells = [”C30:EZ30”, ”C31:EZ31”, ”C32:EZ32”, ”C33:EZ33”, ”C34:EZ34”,
    ”C35:EZ35”, ”C36:EZ36”, ”C37:EZ37”, ”C38:EZ38”]';
% Initialization of the matrix for absorbance data
```

```
data = zeros(9,154);
% Definition of wavelengths
wavelength = [235:5:1000];
% Creation of a figure to visualize the spectra
figure
% Loop to extract and visualize absorbance spectra for each gamma cell
for i = 1:9
    data(i,:) = xlsread(fileExcel, sheet, gammacells(i))';
    plot(wavelength, data(i, :));
    xlabel('Wavelength (nm)');
    ylabel('Absorbance')
    hold on
end
% Definition of concentrations of glucose and lactate for each gamma cell
Conc_k_glu = [4.88, 4.00, 4.22, 4.94, 4.83, 3.50 , 4.88,  4.38, 1.83]';
Conc_k_lac = [2.05, 2.15, 3.63, 2.08, 2.98, 4.24, 2.17, 3.14, 5.79]';
% Creation of a combined concentration matrix
Conc_k = [Conc_k_glu Conc_k_lac];
```

Listing 5.5: The code performs the extraction and visualization of absorbance spectra from various gamma cells in an experimental setup. Each gamma cell's spectrum is plotted, providing insights into the characteristics of the absorbance data. Additionally, concentrations of glucose and lactate for each gamma cell are defined and combined into a matrix for further analysis.

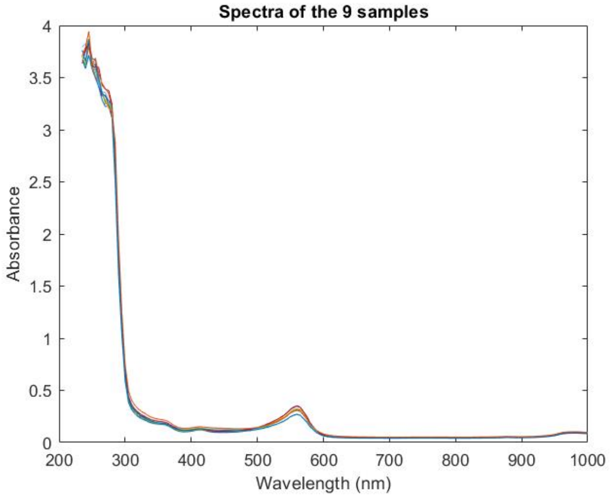
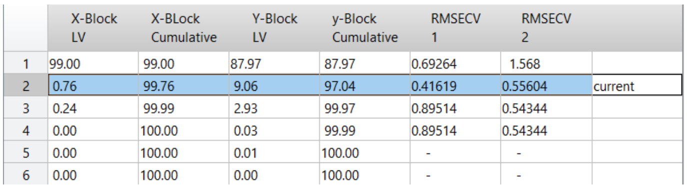


Figure 5.17: The figure displays absorption spectra from nine different samples categorized into three conditions: Non-Conditioned (NC), Conditioned (C), and Static (S). Each line represents the absorbance characteristics of a specific sample, offering a visual overview of spectral variations under distinct experimental conditions.

Application of the Machine Learning Algorithm and results

In the initial phase, we proceed with the data upload, placing all complete spectra in block X and the known concentrations in block Y. From the outset, the data is divided between those used for calibration and model creation and those designated for the validation phase. This division allocates 75% of the data for calibration and the remaining 25% for validation, achieved through the use of the Onion algorithm.

In this scenario, the system has six latent variables available but opts to use only two. The decision to limit the analysis to two latent variables is motivated by achieving improved RMSECV results with this specific configuration. Subsequently, it is necessary to select the type of preprocessing to apply. In spectral analysis, ”mean centering” and ”autoscaling” emerge as two data preprocessing approaches, each with specific advantages based on the nature of the analysis and spectral data characteristics.



	X-Block LV	X-BLock Cumulative	Y-Block LV	y-Block Cumulative	RMSECV 1	RMSECV 2	
1	99.00	99.00	87.97	87.97	0.69264	1.568	
2	0.76	99.76	9.06	97.04	0.41619	0.55604	current
3	0.24	99.99	2.93	99.97	0.89514	0.54344	
4	0.00	100.00	0.03	99.99	0.89514	0.54344	
5	0.00	100.00	0.01	100.00	-	-	
6	0.00	100.00	0.00	100.00	-	-	

Figure 5.18: The table presents various model performance metrics for different numbers of latent variables. The focus is on RMSECV, and the results indicate that selecting two latent variables achieves the lowest RMSECV, supporting the choice of this configuration for enhanced predictive accuracy.

”Mean centering” proves effective in removing offset effects in the data, bringing the mean of each wavelength to zero. This allows for a clear visualization of differences between spectra without altering the overall structure. This technique is particularly suitable when absolute variations in the data are of primary interest compared to relative variations between wavelengths.

On the other hand, ”autoscaling” involves normalizing each variable by dividing by its standard deviation. This process scales variables so that they have a mean of zero and a standard deviation of one. ”Autoscaling” is preferable when relative variations between variables are more relevant than absolute variations and when the measurement units of variables are different.

In the context of spectra, the choice between ”mean centering” and ”autoscaling” depends on the specific goals of the analysis. In the case of spectra from different experimental conditions, such as samples C, NC, and S, the choice of ”mean centering” is motivated by our focus on preserving the absolute shape of the spectra. This is crucial for identifying specific patterns and characteristic variations of each condition.

The decision to use ”mean centering” as a preprocessing technique is based on the goal of preserving the specificity of each sample’s spectra, allowing for an accurate analysis of variations in specific absorption regions while maintaining the consistency of spectral features. This choice provides a solid foundation for subsequent analysis, accurately highlighting spectral differences between the considered experimental conditions. Moving on to the model results, the figures display the comparison between measured and predicted concentrations for glucose and lactate. The model employing two latent variables demonstrates excellent performance, with low values of RMSEC (0.2322) and RMSECV (0.41619) indicating high accuracy during both calibration and cross-validation. The low RMSEP (0.33401) in prediction further underscores the model’s robustness in forecasting concentrations.

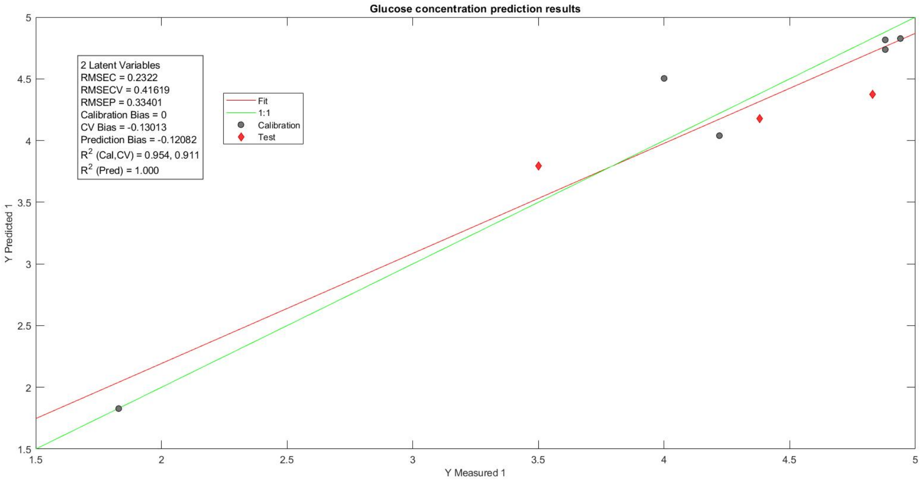


Figure 5.19: Comparison between measured and predicted glucose concentrations using the predictive model, highlighting the consistency between real values and those obtained through the machine learning algorithm.

The calibration bias of 0 signifies that the model’s predictions align well with the actual values during the training phase. The minor systematic errors indicated by the cross-validation bias (-0.13013) and prediction bias (-0.12082) have relatively small magnitudes, suggesting the model’s overall reliability.

The coefficient of determination (R^2) values provide insight into the model’s explanatory power. The high R^2 values for both calibration (0.951) and cross-validation (0.911) demonstrate the model’s ability to capture the variance in the data during training and testing. The perfect R^2 value of 1.000 in prediction indicates an exact fit, reinforcing the model’s capability to precisely predict concentrations.

In summary, the model with two latent variables exhibits exceptional accuracy and predictive power for glucose and lactate concentrations. The low errors, minimal biases, and high R^2 values collectively signify a robust and reliable model, making it suitable for practical applications and further experimentation. The two latent variable model for lactate concentration demonstrates exceptional predictive performance. The minimal values of RMSEC (0.19253) and RMSECV (0.55604) reflect high precision during calibration and cross-validation, respectively. The low RMSEP (0.2445) in prediction reinforces the model's reliability in forecasting lactate concentrations.

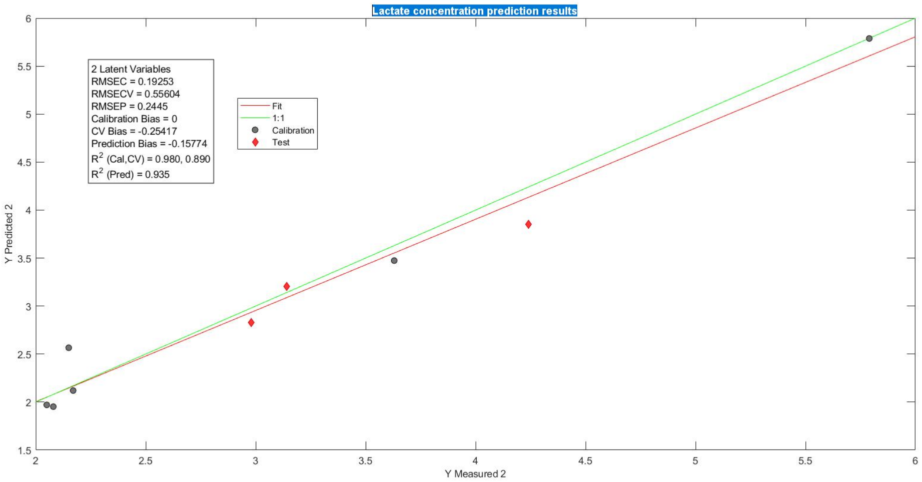


Figure 5.20: Comparison between measured and predicted lactate concentrations using the predictive model, highlighting the consistency between real values and those obtained through the machine learning algorithm.

A calibration bias of 0 indicates a strong alignment between the model's predictions and the actual values during the training phase. Although minor systematic errors are suggested by the cross-validation bias (-0.25417) and prediction bias (-0.15774), their relatively small magnitudes imply overall robustness.

The coefficient of determination (R^2) values offer insights into the model's explanatory capability. High R^2 values for calibration (0.980) and cross-validation (0.890) underscore the model's capacity to capture lactate data variance during training and testing. A prediction R^2 value of 0.935 signals a robust fit, affirming the model's accuracy in predicting lactate concentrations.

In conclusion, the two latent variable model exhibits outstanding accuracy and predictive power for lactate concentrations. The minimal errors, negligible biases, and high R^2 values collectively affirm a reliable and robust model suitable for practical applications and further investigations.

To further enhance the predictive capabilities, an advanced data transformation using a polynomial transformation with squared and cubed cross terms is applied. This choice aims to create a more accurate and adaptable predictive model, especially considering complex nonlinear relationships between wavelengths and absorbance in the spectral data.

The introduction of polynomial and cross terms improves the model's predictive capacity by increasing flexibility, allowing for more accurate predictions compared to rigid models. The transformation simplifies the data structure, focusing on significant components and reducing the impact of irrelevant noise or variations.

The effectiveness of this transformation depends on the specific nature of spectral data. A careful analysis of different transformation techniques is essential to select the most suitable strategy, ensuring maximum predictive accuracy and optimal interpretability of the model.

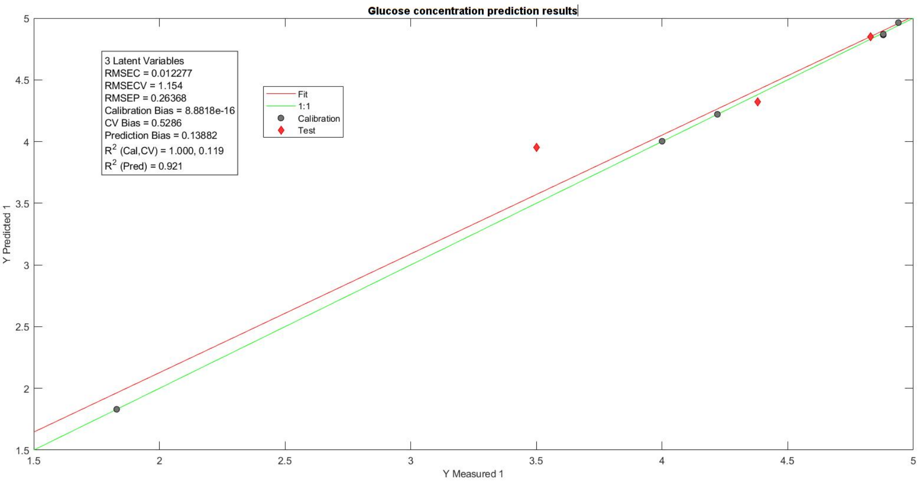


Figure 5.21: Comparison between measured and predicted glucose concentrations using the predictive model, highlighting the consistency between real values and those obtained through the machine learning algorithm. It's noteworthy that data transformation methods were employed to enhance the accuracy of the predictions.

The introduction of polynomial transformations, including squared and cubed cross terms, has significantly improved the predictive performance of the model, particularly in estimating glucose concentrations. This improvement is evident through key metrics. Firstly, the Root Mean Square Error of Cross-Validation (RMSECV) for the transformed model is significantly lower, indicating heightened accuracy in predicting glucose concentrations during cross-validation. This reduction signifies a closer alignment with actual glucose levels in the validation set.

Moreover, the calibration bias in the transformed model is negligible, indicating a more bal-

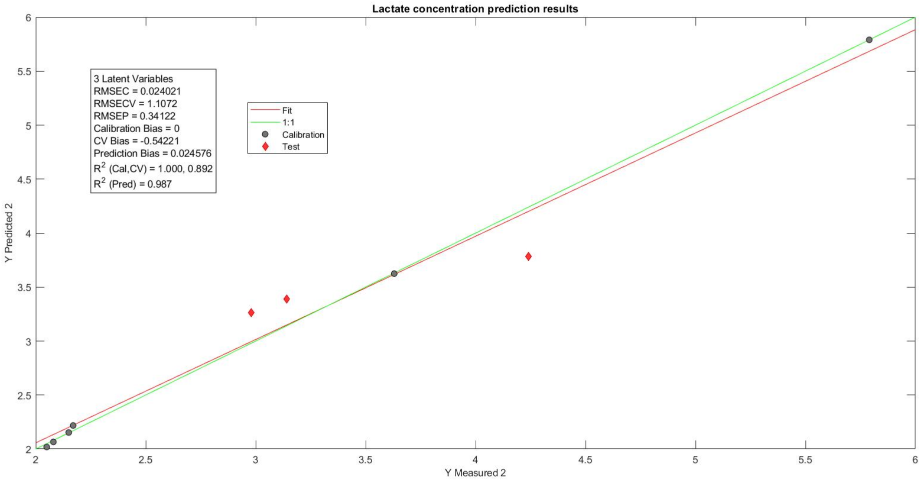


Figure 5.22: Comparison between measured and predicted lactate concentrations using the predictive model, highlighting the consistency between real values and those obtained through the machine learning algorithm. It's noteworthy that data transformation methods were employed to enhance the accuracy of the predictions.

anced fit during the calibration phase. This suggests that the transformed model tends to avoid systematic overestimation or underestimation of glucose concentrations during the calibration process. Additionally, the R2 values, representing the goodness of fit, are higher for both calibration and cross-validation in the transformed model. This signifies not only a superior fit to the calibration data but also a more effective generalization to unseen data, capturing underlying patterns and relationships.

The benefits of the transformed model also extend to the prediction of lactate concentrations, albeit to a slightly lesser extent. The lower RMSECV and a slightly higher R2 (CV) for lactate indicate improved accuracy and a better fit to the data during cross-validation.

In conclusion, the incorporation of polynomial transformations with squared and cubed cross terms proves to be a valuable enhancement for the model's predictive capabilities. The transformed model demonstrates an improved ability to capture non-linear relationships in the data, resulting in superior accuracy, minimized bias, and enhanced generalization. Therefore, the transformed model stands out as the preferred choice, offering a more nuanced understanding of the complex relationships within the spectral data, especially in the precise prediction of glucose concentrations.

Model Configuration Evaluation: Transformation Impact on Predictive Performance

Within the analysis of predictive models for glucose and lactate concentrations, a thorough evaluation of different configurations was conducted, encompassing models without any transformation and models adopting polynomial transformation with the introduction of cubic and square terms. These approaches yielded distinct results, providing a detailed overview of model performance and adaptability to the spectral data under consideration.

Without Transformation, the models exhibit valid results with accurate predictions, especially for glucose and lactate. The absence of transformations simplifies the model, making it more interpretable, with high R^2 (Pred) values indicating good adaptability to new observations.

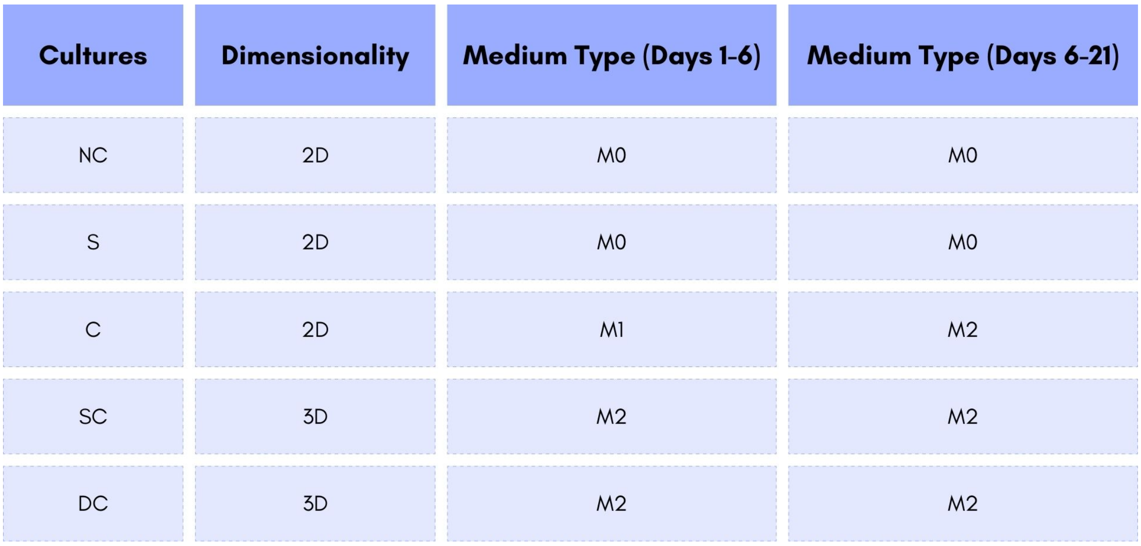
With Transformation, transformed models significantly improve prediction accuracy, especially for glucose. The near absence of Calibration Bias suggests a more balanced calibration, while elevated R^2 (Pred) values indicate good adaptability to new observations.

The choice between models with or without transformation is guided by a balanced consideration of predictive accuracy and model complexity. The transformation approach offers clear advantages in terms of prediction accuracy but introduces greater complexity. The final decision should reflect the specific objectives of the analysis, considering data availability and the need for a balance between accuracy and interpretability.

In conclusion, since it is preferable for our application to achieve greater precision in prediction, we are willing to accept the additional complexity of the model. However, it is crucial to carefully evaluate the trade-off between complexity and precision based on the specific objectives of our analysis.

5.4 Experiment on Glucose and Lactate Detection in a Culture Medium: Impact of Experimental Conditions on Predictive Modeling

This experimental phase involves a diversified set of five cell cultures, each representing a specific experimental condition. The NC (Non-Conditioned) and C (Conditioned) cultures are two-dimensional and differ based on the use of conditioned (C) or non-conditioned (NC) medium. The S culture (Static) follows a static mode, while the DC (Dynamic Conditioned) and SC (Static Conditioned) cultures involve the use of scaffold supports, differing in the dynamic nature of the culture condition. In the table shown in the figure 5.23, we have summarized the main characteristics of the cultures used for the analysis. The introduction of



Cultures	Dimensionality	Medium Type (Days 1-6)	Medium Type (Days 6-21)
NC	2D	M0	M0
S	2D	M0	M0
C	2D	M1	M2
SC	3D	M2	M2
DC	3D	M2	M2

Figure 5.23: Characteristics of Different Cultures

different culture media poses a significant challenge in the prediction phase, especially when applying Partial Least Squares (PLS) software to spectral data. To mitigate this complexity, analyses were conducted using exclusively spectra from identical cultures, ensuring greater consistency in the results. It is crucial to note that the spectra were acquired in the visible range, ranging from 235 nm to 1000 nm. Since lactate is one of the substances of interest for concentration prediction from spectra, it would be ideal to use the optimal acquisition range for this substance, which, in the case of lactate, is in the MIR. However, due to the availability of laboratory instrumentation allowing measurements in the visible range, this solution was chosen in the initial testing phases, considering also the high cost of technology required for MIR range acquisition.
The MATLAB code used for data import has been omitted in this introduction, as it has been previously explained in the preceding paragraphs.

5.4.1 Bidimentional Static Culture Experiment: Spectral Analysis and Predictive Modeling

The first experiment conducted during this phase involved acquiring spectra for the S culture on days 0, 6, 8, 10, 12, 14, 16, 18, and 21. As evident, samples for days 2 and 4 are not present in the corresponding vials due to compromise. To ensure data integrity, such data were excluded from the collection. After acquiring spectra with the TECAN 200 Infinite spectrometer, a data analysis was performed. Specifically, an initial modeling with a set of 10

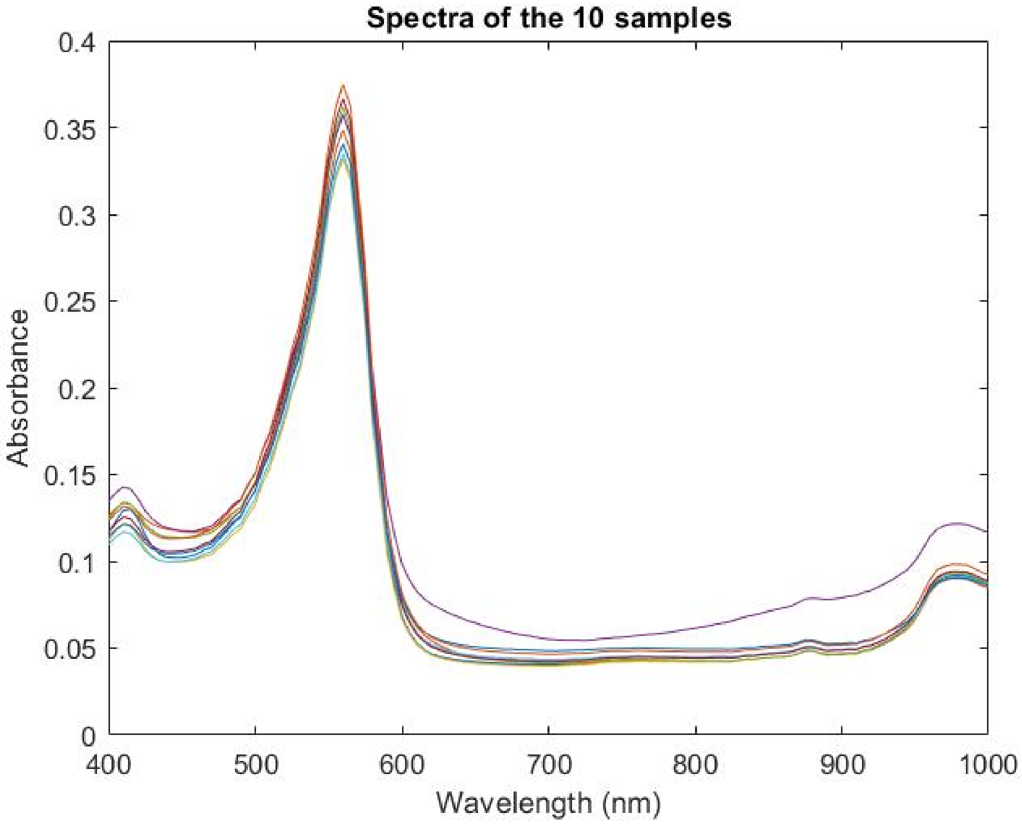


Figure 5.24: Acquired spectra for sample S from 400 to 1000 nm.

spectra was conducted. However, results revealed a sample in the validation set, particularly in lactate concentration prediction, showing significantly deviated results from the fit curve. The removal of this specific sample, identified as the data from row number 9, was carried out to explore potential improvements in the final model and predictions.

Modeling Results

After acquiring spectra with the TECAN 200 Infinite spectrometer, a data analysis was performed. Specifically, an initial modeling with a set of 10 spectra was conducted. However, results revealed a sample in the validation set, particularly in lactate concentration prediction, showing significantly deviated results from the fit curve. The removal of this specific sample, identified as the data from row number 9, was carried out to explore potential improvements in the final model and predictions.

Before delving into the removal of a specific sample, the modeling results exhibited remarkable precision, with low values of RMSEC, RMSECV, and RMSEP for both glucose and lactate. However, it is important to highlight that lactate showed a relatively low R^2 CV of 0.0445795, indicating limited generalization capacity during cross-validation.

Subsequently, in the model creation process, it was observed that a sample in the validation set for lactate displayed significantly deviated results from the fit curve. The removal of this specific sample, identified as the data from row number 9, was carried out to explore potential improvements in the final model and predictions.

Before the removal of the problematic sample, the model for glucose exhibited high precision during the training phase, as indicated by the RMSEC (Root Mean Square Error of Calibration) of 0.0183117. This parameter represents the average quadratic error between observed and predicted data during training. The high value of R^2 Cal (Coefficient of Determination Calibration) of 0.997737 indicated an excellent fit of the model to the data during this phase. However, during cross-validation, the model showed lower generalization capacity to new data, as highlighted by the RMSECV (Root Mean Square Error of Cross-Validation) of 0.607814 and the R^2 CV (Coefficient of Determination Cross-Validation) of 0.659822. The RMSECV represents the average quadratic error during cross-validation, while the R^2 CV measures the model's ability to generalize to new data. The removal of the sample led to an improvement in these parameters, suggesting greater stability in glucose predictions for data not used in model construction.

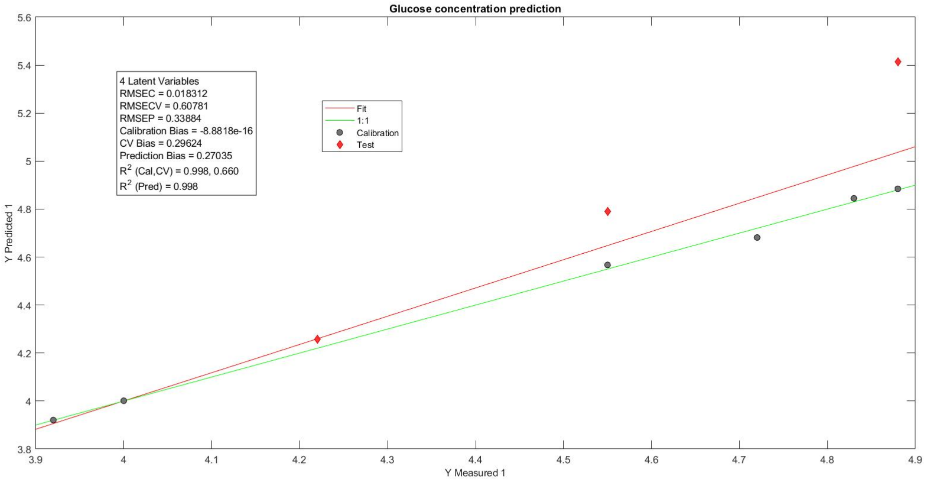


Figure 5.25: Scatter Plot of Measured vs. Predicted Y-values (Original Model without Sample Removal). The model for glucose exhibits high precision during training (RMSEC = 0.0183117, R^2 Cal = 0.997737). However, cross-validation indicates lower generalization capacity (RMSECV = 0.607814, R^2 CV = 0.659822).

For lactate, the original model exhibited limited generalization capacity, indicated by a low R^2 CV of 0.0445795 during cross-validation. After the removal of the problematic sample, a significant improvement occurred, with a notable increase in R^2 CV, suggesting greater adaptability of the model to new observations. In the analysis of prediction parameters after removal, the RMSEP (Root Mean Square Error of Prediction) for glucose slightly increased to 0.414195, indicating greater variability in predictions for new data not used in model construction. However, the R^2 Pred (Coefficient of Determination Prediction) remained high at 1, confirming that the model has good adaptability to new data. For lactate, the variations were more evident, with a significant decrease in RMSEP to 0.101089, indicating greater precision in predictions for new data. The R^2 Pred increased to 1, confirming a perfect adaptability of the model to new observations.

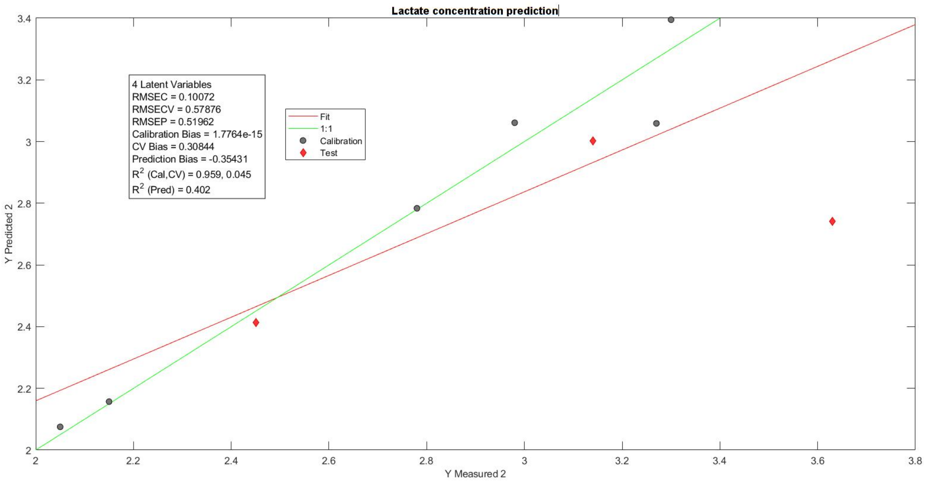


Figure 5.26: Scatter Plot of Measured vs. Predicted Lactate Concentrations (Original Model without Sample Removal). The original lactate model demonstrates limited generalization capacity during cross-validation (RMSECV = 0.0445795, R^2 CV = 0.0445795).

Conclusion

After a thorough analysis of the results obtained before and after the removal of the problematic sample in the PLS modeling for predicting glucose and lactate concentrations, important considerations emerge.

The evaluation of prediction parameters indicates that, despite a slight increase in RMSEP for glucose after removal, the model retains remarkable precision, as evidenced by an R^2 Pred of 1. The tendency to predict slightly higher values, indicated by the positive pred bias, may be acceptable depending on the analysis goals.

For lactate, the removal of the problematic sample led to significant improvements. RMSEP

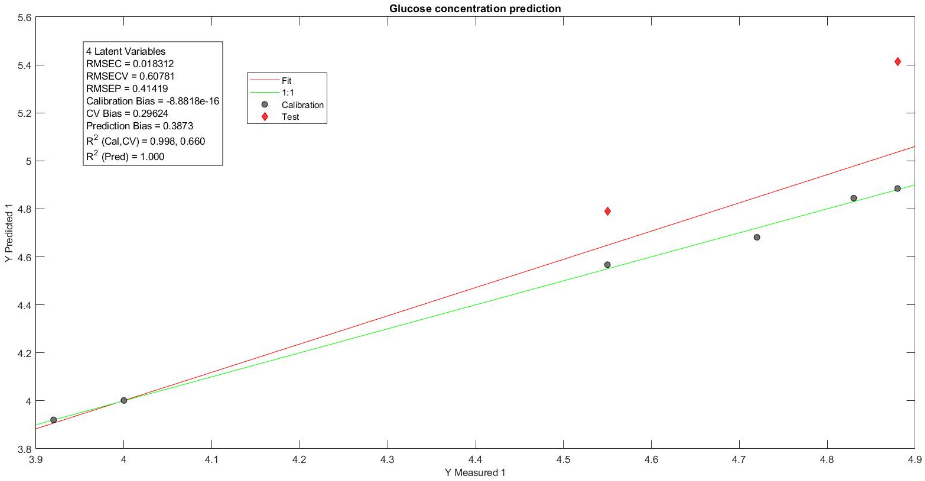


Figure 5.27: Scatter Plot of Measured vs. Predicted Y-values (Model after Sample Removal). The removal of the problematic sample improves cross-validation parameters, suggesting greater stability in glucose predictions for new data (RMSECV and R^2 CV improvement).

has notably decreased, and R^2 Pred has risen to 1, suggesting a perfect adaptability of the model to new observations. This indicates that the removal procedure contributed to a more accurate and reliable prediction for lactate.

The choice between using the model with or without the removal of the specific sample should be guided by the specific goals of the analysis. In the context of this experiment, where accurate prediction of lactate is crucial, sample removal seems justified. Significant improvements in prediction parameters, especially for lactate, suggest that this procedure enhances the overall reliability of the model.

However, it is crucial to consider the trade-off between precision and model stability. If the primary goal is highly precise lactate prediction, sample removal is the preferable choice. On the other hand, if stability and model generalization are priorities, the decision may require a more in-depth evaluation.

In conclusion, the specific sample removal procedure seems to provide a significant improvement in concentration predictions, especially for lactate. The decision to adopt this procedure should be based on a careful evaluation of the specific analysis goals and the desired balance between precision and model stability.

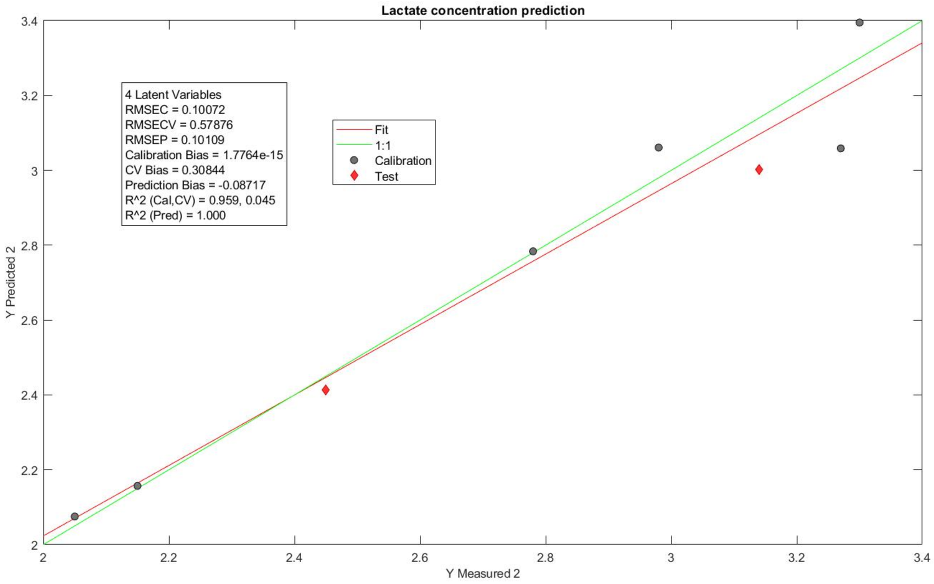


Figure 5.28: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The removal of the problematic sample results in significant improvements in prediction parameters. RMSEP decreases to 0.101089, indicating greater precision in predictions for new data. The R^2 Pred increases to 1, confirming a perfect adaptability of the model to new observations.

5.4.2 Bi-dimensional Non-Conditioned Culture Experiment: Spectral Analysis and Predictive Modeling

To broaden the scope of our investigation, we extended our study to non-conditioned cultivation, characterized by the constant presence of M0 medium throughout the experiment's duration. As customary, soil replacement occurs on the sixth day, introducing a significant transition in environmental composition.

Our approach focuses on spectral analysis of this non-conditioned cultivation, utilizing the optical range. This methodology allows us to explore spectral variations over time, providing a detailed perspective on the cultivation dynamics and interactions among M0 soil components. In our analysis, we chose to commence from 400 nm, maintaining a targeted approach that preserves relevant information for lactate and glucose concentrations. This decision was motivated by the common practice of excluding early wavelengths in spectral analyses to reduce electronic noise, enhance baseline stability, and optimize predictive models.

Special attention was dedicated to the analysis and evaluation of predictive models for glucose and lactate concentrations. Using spectral data in the optical range aims to develop reliable and accurate models to predict variations in these crucial substances in non-conditioned cultivation. This integrated approach, combining spectral analysis and predictive modeling, forms a robust foundation for in-depth understanding of ongoing biological and chemical processes. Its implementation is essential for optimizing non-conditioned cultivation conditions, contributing to advancements in scientific research.

Modeling Results

Particular attention will be paid to the evaluation of predictive models for the detection of glucose and lactate concentrations. Using spectral data in the optical range, we aim to develop reliable models that can accurately predict changes in these key substances in the context of unconditioned culture. This integrated approach, combining spectral analysis and predictive modeling, will provide a solid basis for understanding ongoing biological and chemical processes, as well as for the optimization of future unconditioned cultures. Again, the culture was sampled on days 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 21. In this case, the sample referring to day 4 was absent. In addition, we were not placed in the optimum conditions for the realiza-

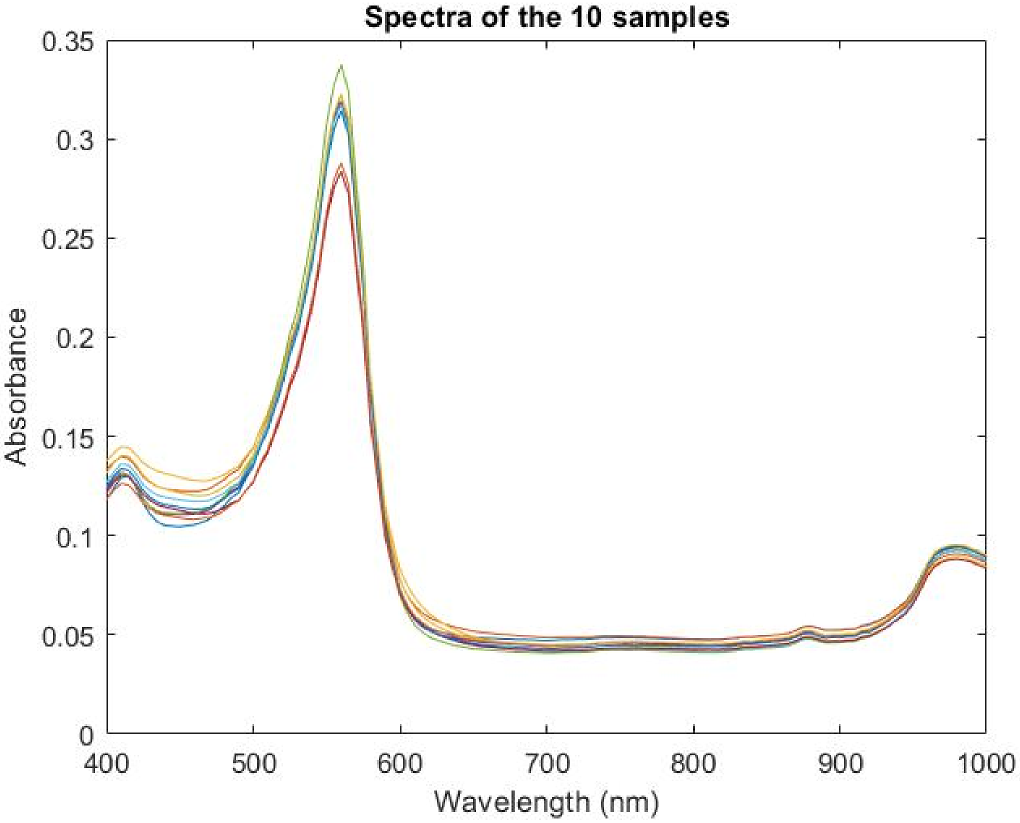


Figure 5.29: Acquired spectra for sample S from 400 to 1000 nm.

tion of the model, as we have the substitution of the culture medium, which obviously entails variations. Furthermore, the data of the known concentrations provided to us are the result of the mathematical average between two concentration measurements on two different samples. Also in this analysis, we started with glucose, immediately finding some difficulties with the predictive model of glucose. The very low RMSEC (Root Mean Square Error of Calibration) value of 0.0243313 testifies to the model's ability to fit the training data excellently. However, during the cross-validation phase, the slightly higher RMSECV (Root Mean Square Error of Cross-Validation), with a value of 1.01104, indicates a possible excessive complexity of the model or the presence of noise in the data.

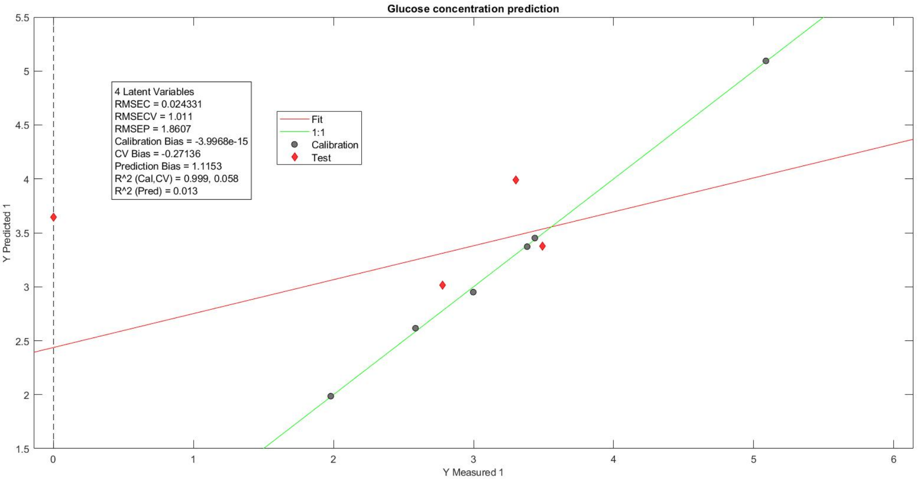


Figure 5.30: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

The RMSEP (Root Mean Square Error of Prediction) value during prediction on unseen data is moderately low, registering 1.86068. This value, although indicative of good predictive ability, is higher than the calibration and cross-validation parameters, suggesting some uncertainty in prediction on new data.

The R^2 (Coefficient of Determination) takes on extremely high values during calibration (0.99936), indicating an almost complete explanation of the variation in the training data. However, during cross-validation (0.0577914) and the prediction phase (0.0129411), the R^2 is significantly lower, suggesting a potential challenge in generalizing the model to new data.

During the analysis of the graph, two points, 8 and 9, were found to show significant discrepancies with the model. These points were excluded to improve the predictive accuracy of the

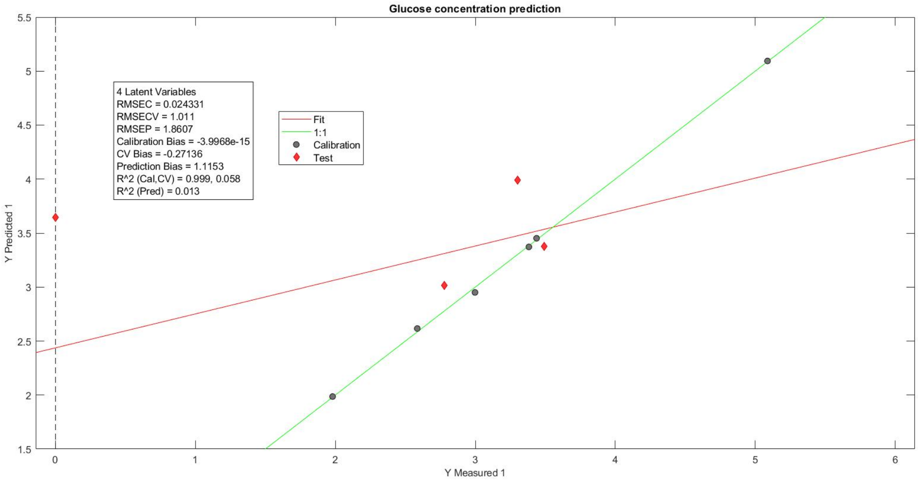


Figure 5.31: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

model. The initial RMSEC value was remarkably low (0.0243313), indicating the exceptional fit of the model to the training data. However, the higher RMSECV (1.01104) during cross-validation suggested possible excessive model complexity or the presence of noise in the data. After removing the outlier points, the RMSECV remained almost unchanged, but RMSEP experienced a significant improvement, dropping to 0.516896. This reduction indicates greater accuracy in predictions on new data. The initial R^2 Cal was extremely high (0.99936), indicating an excellent explanation for the variation in the training data. However, during both cross-validation and prediction, there was a significant increase in the R^2 values, reaching 0.0577914 and 1, respectively. This suggests a significant improvement in the ability to generalize the model to new data, overcoming the initial challenges of low R^2 in cross-validation and prediction. In the course of our analysis, our focus was on creating a predictive model for measuring lactate concentration. Initially, the model had the following parameters:

- RMSEC: 0.578098
- RMSECV: 1.29598
- RMSEP: 0.5211
- R^2 Cal: 0.616028
- R^2 CV: 0.300714

- R^2 Pred: 0.403273

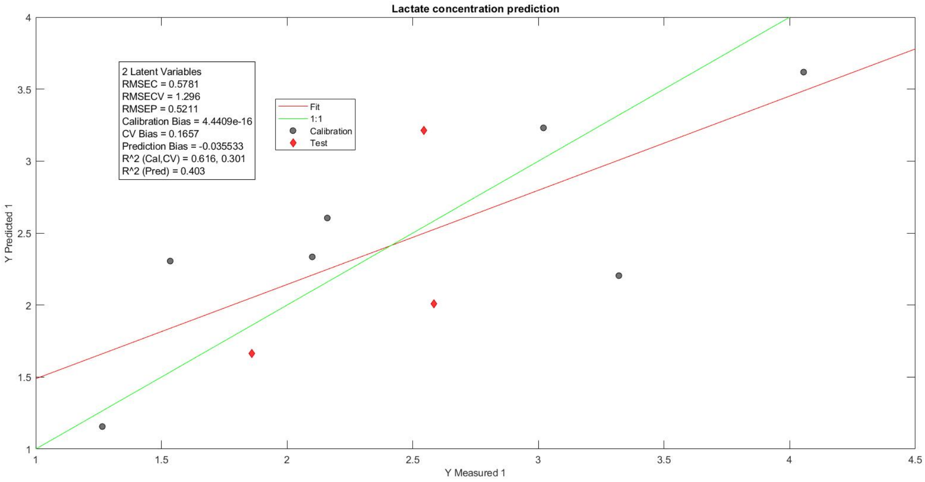


Figure 5.32: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

Notably, after obtaining the model and observing the graph and data points, we identified a point in the calibration set that was slightly distant from the fit curve, with a lactate concentration measurement of 2.545 and a prediction of 3.21107. The removal of this point resulted in the following changes to the model:

- RMSEC: 0.578098
- RMSECV: 1.29598
- RMSEP: 0.430686
- R^2 Cal: 0.616028
- R^2 CV: 0.300714
- R^2 Pred: 1

A detailed analysis of the two models reveals some important considerations. Before removing the outlier point, the model exhibited good precision during calibration (R^2 Cal: 0.616028) but showed some complexity during non-calibration phases, including cross-validation (R^2 CV: 0.300714) and the prediction phase (R^2 Pred: 0.403273).

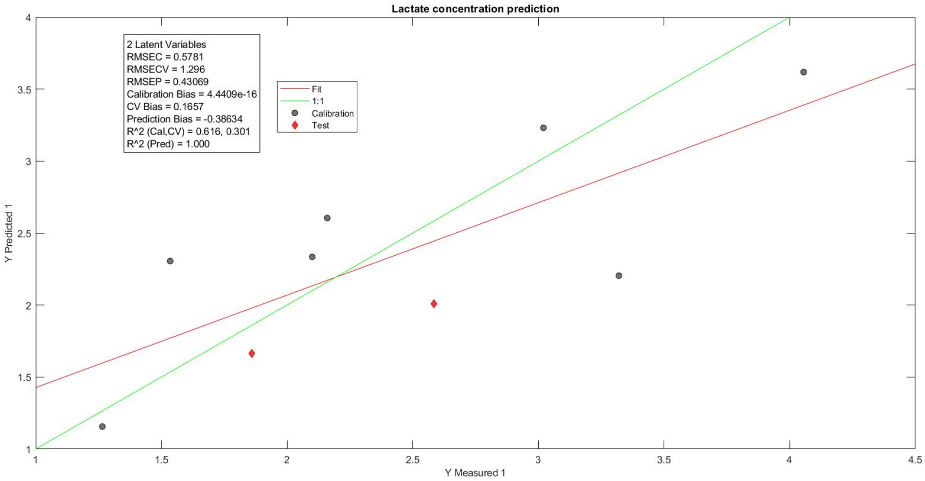


Figure 5.33: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

After excluding the problematic point, a significant improvement in the model's performance was observed. The reduction in RMSEP to 0.430686 indicates greater precision in predictions on new data. Moreover, achieving the maximum R^2 Pred value of 1 suggests a complete explanation of the variation in prediction data.

The decision to remove the outlier point is supported by the fact that its presence compromised the precision of the model, especially in non-calibration phases such as cross-validation and prediction. Correcting this discrepancy contributed to better adaptability of the model to the data and increased reliability in predictions. This optimization is crucial to ensuring the validity and effectiveness of the predictive model, providing more accurate and generalizable results in the measurement of lactate concentrations.

The initial R^2 Cal was extremely high (0.99936), indicating an excellent explanation for the variation in the training data. However, during both cross-validation and prediction, there was a significant increase in the R^2 values, reaching 0.0577914 and 1 respectively. This suggests a significant improvement in the ability to generalise the model to new data, overcoming the initial challenges of low R^2 in cross-validation and prediction.

Conclusion

In conclusion, the in-depth spectral analysis and evaluation of predictive models for glucose and lactate concentrations in non-conditioned cultivation provided a detailed perspective on the dynamics of our specific experiment. The use of the optical range and attention to spectral details helped develop accurate models to predict variations in these key substances within the context of our non-conditioned culture. The removal of outlier points enhanced the precision and reliability of predictions, underscoring the importance of an integrated approach involving spectral analysis and predictive modeling. These results establish a solid foundation for further exploration and optimization in future cultivation experiments, contributing to our understanding of ongoing biological and chemical processes.

5.4.3 Bi-dimentional Conditioned Culture Experiment: Spectral Analysis and Predictive Modeling

In this section, we will create a predictive model using PLS to predict the concentration of lactate and glucose within a conditioned two-dimensional culture. In this case, the culture was sampled on days 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 21. Notably, the sample referring to the fourth day was absent. In Figure 5.34, the ten spectra acquired through the spectrometer are

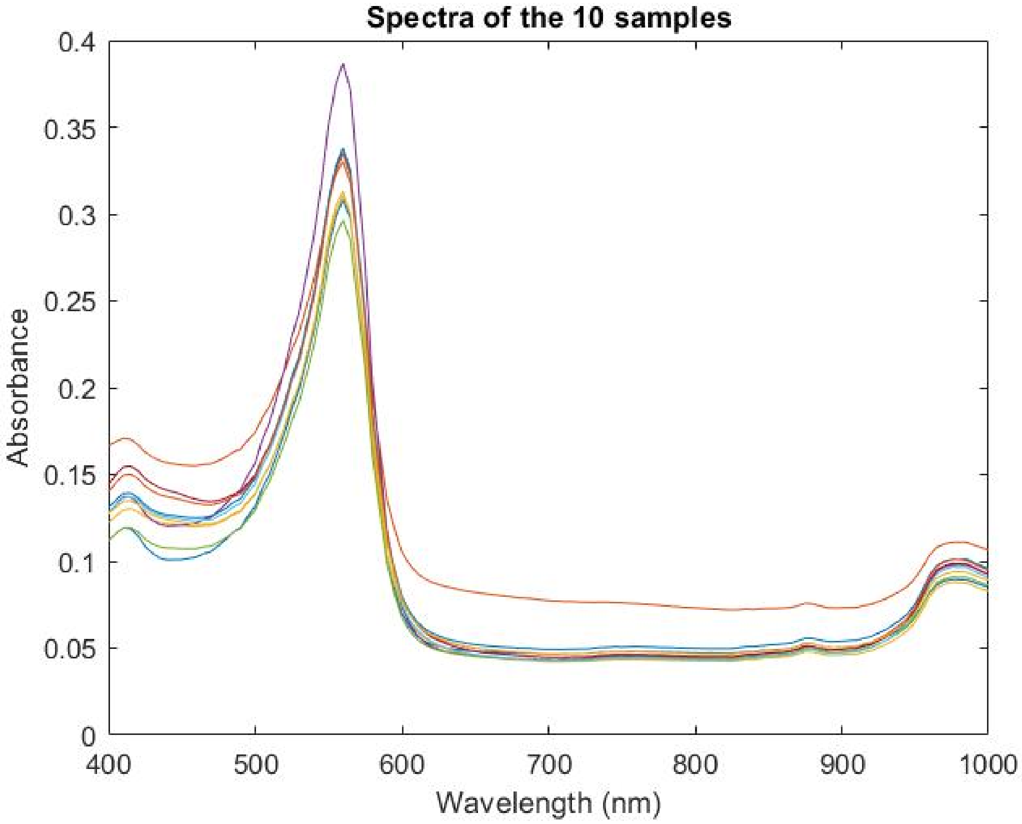


Figure 5.34: Acquired spectra for sample C from 400 to 1000 nm.

presented, all within the visible range. For this specific case, we preferred to consider information starting from the wavelength of 400 nm. The choice to initiate the analysis from 400 nm is motivated by the common practice of excluding information related to the initial wavelengths in spectral analyses. This decision was made to reduce sensitivity to electronic noise, enhance baseline stability, and optimize the predictive model, while ensuring that relevant information for lactate and glucose concentration is preserved in our analysis.

Modeling Results

As part of our investigation into conditioned cultivation, achieving precise predictions of glucose and lactate concentrations is crucial to demonstrate the capability of Partial Least Squares (PLS) in developing predictive models for substance concentrations. It is important to note that, in our study, glucose and lactate serve as key substances for analysis, and our primary focus is on showcasing the effectiveness of PLS in generating accurate predictive models for these concentrations.

Nevertheless, our PLS-based approach has encountered notable challenges that could compromise the precision of predictions. Changes in cultivation conditions, particularly the transition from M1 to M2 soil between the sixth and twenty-first days, along with the absence of samplings on the fourth day, may introduce complex variations in glucose production, directly impacting the validity of the predictive model. In this section, we will closely examine the performance of our PLS model, considering the temporal dynamics and peculiarities of our conditioned cultivation, to assess the robustness and reliability of the obtained predictions.

The accurate prediction of glucose concentration is pivotal in our investigation into conditioned cultivation. The model, initially constructed using the entire dataset, exhibited the following performance metrics for glucose concentration: Before Glucose Outlier Removal:

- RMSEC: 0.510162
- RMSECV: 2.14233
- RMSEP: 4.29748
- R^2 Cal: 0.695992
- R^2 CV: 0.179894
- R^2 Pred: 0.0185977

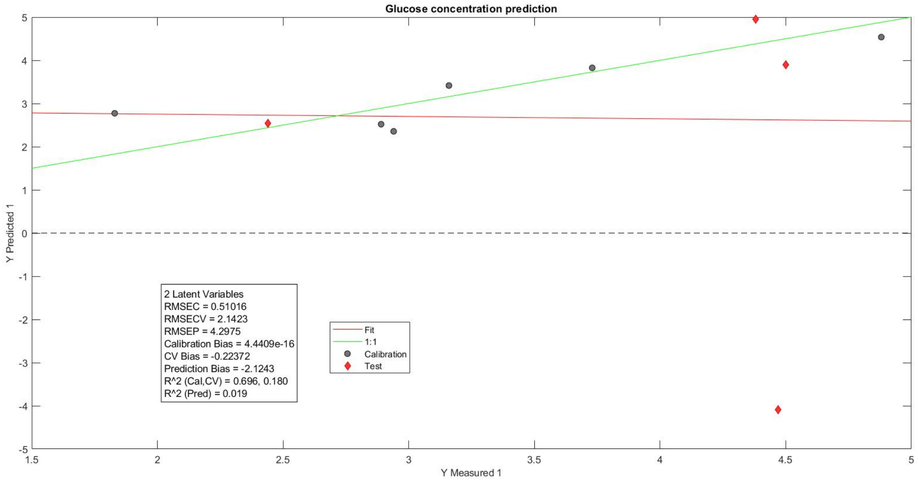


Figure 5.35: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

Subsequently, during the application of the variable selector, data point 7, identified as an outlier specifically for glucose concentration, was removed to enhance the model's accuracy. After Glucose Outlier Removal (Point 7):

- RMSEC: 0.510162
- RMSECV: 2.14233
- RMSEP: 0.482205
- R^2 Cal: 0.695992
- R^2 CV: 0.179894
- R^2 Pred: 0.76752

The targeted removal of glucose outlier (Point 7) has led to significant improvements in the model's accuracy for predicting glucose concentration. This strategic choice resulted in a substantial reduction in RMSEP, indicating increased precision in predicting new glucose data. R^2 Pred has significantly increased, highlighting a better ability of the model to generalize to glucose samples outside the training set.

It is crucial to emphasize that outlier removal is a practice that requires caution and a thorough

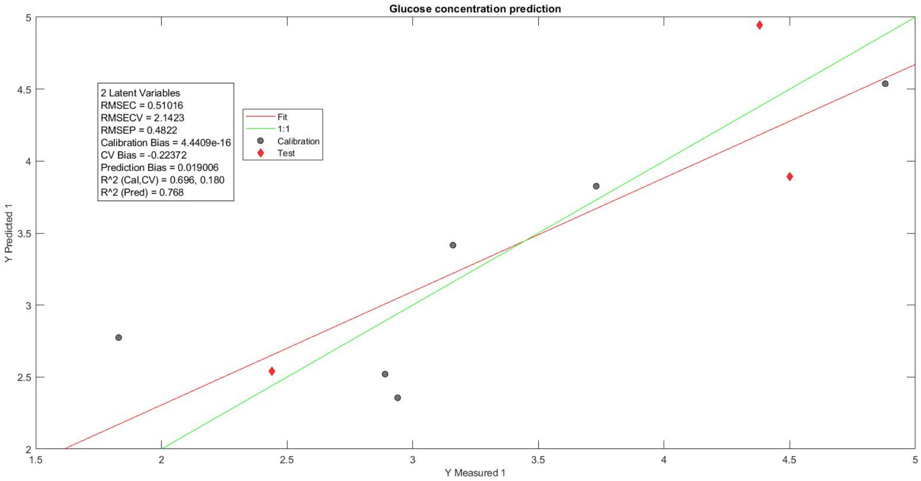


Figure 5.36: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

understanding of the experimental context. However, in our case, this operation was fundamental to ensuring a more reliable and accurate predictive model for glucose concentration in our investigation into conditioned cultivation.

A comprehensive evaluation of the lactate concentration predictive models involved comparing their performance before and after strategically removing specific samples. Notably, two samples—one from the calibration set and another from the validation set—were excluded to understand their influence on model robustness.

Initially, the lactate concentration predictive model displayed the following parameters:

- RMSEC: 0.59073
- RMSECV: 4.64854
- RMSEP: 1.72715
- Bias: -8.88178e-16
- CV Bias: 1.65875
- Pred Bias: -0.251501
- R^2 Cal: 0.81412

- R^2 CV: 0.170865

- R^2 Pred: 0.644816

Following the strategic exclusion of the second sample from the calibration set and the eighth sample from the validation set, the model exhibited adjustments:

- RMSEC: 0.67645

- RMSECV: 3.2348

- RMSEP: 0.285631

- Bias: 0

- CV Bias: 0.0578471

- Pred Bias: 0.0897236

- R^2 Cal: 0.722489

- R^2 CV: 0.0784432

- R^2 Pred: 1

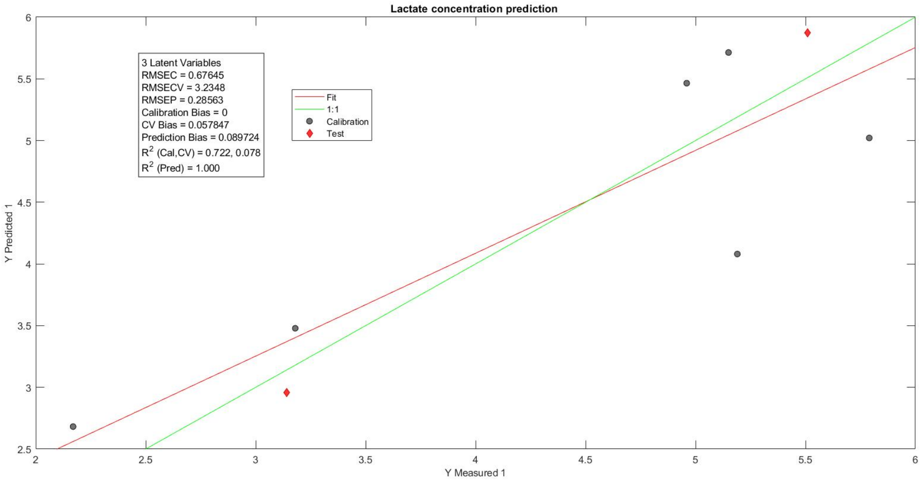


Figure 5.37: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

The rationale behind the sample exclusions was grounded in a meticulous examination of their impact on the model's predictive capabilities. Sample 8 from the validation set displayed a substantial deviation, justifying its removal due to its pronounced outlier status. Similarly, Sample 2 from the calibration set warranted exclusion to enhance overall model accuracy, despite exhibiting a minor discrepancy.

The comparative analysis indicated a nuanced trade-off. While the removal of Sample 8 significantly improved RMSEP and Pred Bias, addressing a pronounced outlier, the exclusion of Sample 2 introduced a slight reduction in R^2 Cal. Nevertheless, the net effect was an overall enhancement in predictive precision. These strategic sample exclusions were imperative for cultivating a more robust and reliable predictive model for lactate concentrations within the conditioned two-dimensional culture (C).

Conclusion

The thorough analysis of predictive models for lactate and glucose concentrations within conditioned cultures has yielded significant insights, revealing the impact of crucial decisions in the initial phase of cultivation.

Initially, the model for lactate concentration exhibited promising precision and adaptability parameters, as highlighted by RMSEC, RMSECV, and R^2 Cal. However, careful examination revealed the presence of influential samples that affected the model's robustness. The strategic exclusion of specific samples led to a substantial improvement in predictive accuracy and model adaptability, confirming the importance of a detailed and targeted data analysis.

The variation in culture medium, transitioning from M1 to M2, emerged as a critical element. The transition on the sixth day could have introduced substantial variations in cell metabolic dynamics, directly impacting lactate and glucose concentrations. This discontinuity poses a challenge in creating stable predictive models and may be responsible for some of the observed variations in results.

In conclusion, despite initial challenges, refining the model through targeted exclusions has proven to be crucial in obtaining more accurate results. The variability in the culture medium underscores the importance of carefully addressing environmental conditions in such analyses, paving the way for further explorations and refinements in experimental design.

5.4.4 Static Culture Experiment: Spectral Analysis and Predictive Modeling

To conduct further experiments, we also considered a static culture. This culture is three-dimensional on a scaffold and has the characteristic of being realized by providing M2 culture medium for the entire duration of the experiment. Similar to the other cultures, medium replacement occurs on the sixth day, creating a significant transition in environmental composition. Our approach will focus on the spectral analysis of this non-conditioned culture, utilizing the optical range. This will allow us to explore and understand spectral variations over time, providing a detailed perspective on the culture dynamics and interactions among the components of the M0 soil. Once again, the decision to commence the analysis from 400 nm is motivated by the common practice of excluding information related to early wavelengths in spectral analyses. This decision aims to reduce sensitivity to electronic noise, enhance baseline stability, and optimize the predictive model, while ensuring that information relevant to

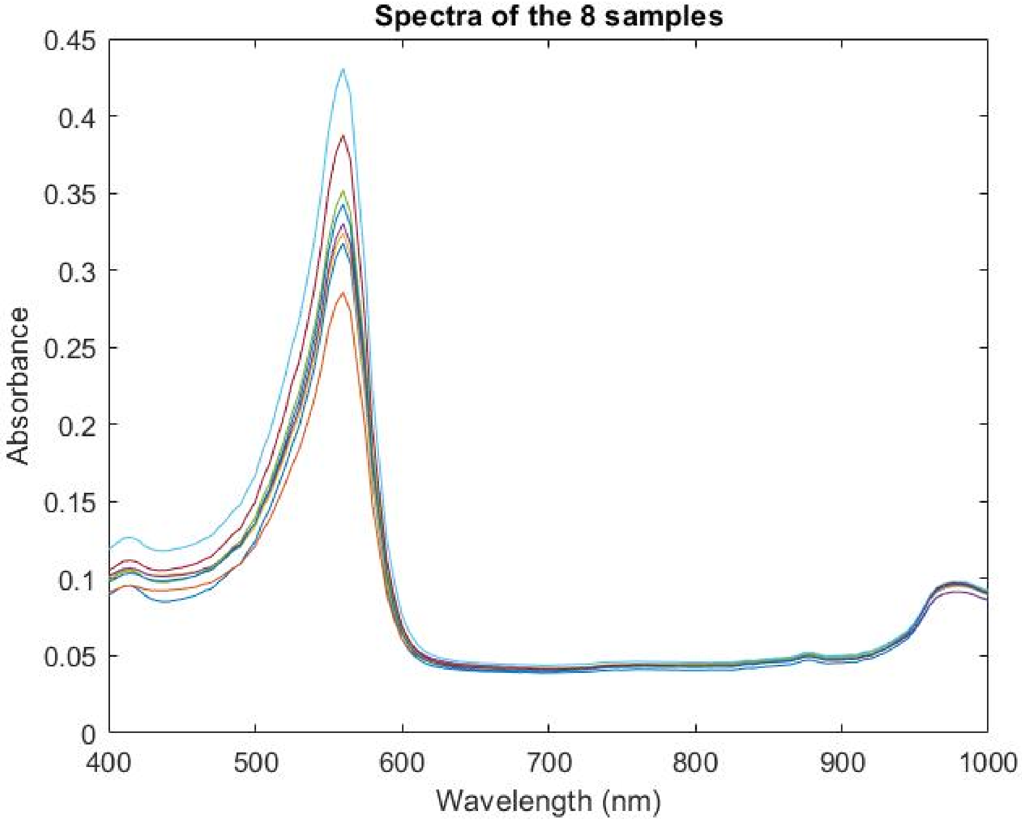


Figure 5.38: Acquired spectra for sample SC from 400 to 1000 nm.

lactate and glucose concentrations is preserved in our analysis.
Special attention will be devoted to the evaluation of predictive models for detecting glucose and lactate concentrations. By utilizing spectral data in the optical range, our goal is to develop reliable models that can accurately predict changes in these key substances within the context of non-conditioned cultivation. This integrated approach, combining spectral analysis and predictive modeling, will establish a solid foundation for understanding ongoing biological and chemical processes and optimizing future non-conditioned cultures.
In this case as well, the culture was sampled on days 0, 2, 4, 6, 8, 10, 12, 14, 16, 18. However, a complication arose around the 16^{th} day, where mold appeared. Consequently, the culture was compromised, and it was deemed appropriate to discard the measurements acquired on the 18_{th} day.

Modeling Results

Thanks to the proactive removal of the compromised culture, in the case of glucose, no additional removals were necessary. The obtained model immediately presents excellent parameters, indicative of high precision in predicting glucose concentrations.

- The **RMSEC** (Root Mean Square Error of Calibration) is 0.312709, highlighting the model's excellent adaptation to the training data.
- The **RMSECV** (Root Mean Square Error of Cross-Validation) is 0.732089, indicating good generalization ability to unseen data during cross-validation.
- The **RMSEP** (Root Mean Square Error of Prediction) is 0.0659966, reflecting the model's accuracy in predictions on new data.
- The **R² Cal** (Coefficient of Determination for Calibration) is 0.364766, suggesting satisfactory explanation of variation in training data.
- The **R² CV** (Coefficient of Determination for Cross-Validation) is 0.0649299, and despite being relatively low, the maximum value of **R² Pred** indicates a highly predictive model on new data, implying a complete explanation of variation.

In conclusion, the obtained model for predicting glucose concentrations is robust and highly reliable, demonstrating remarkable accuracy and generalization capabilities to previously unseen data. The analysis of the lactate model, both before and after the removal of the outlier point, reveals crucial insights into its performance characteristics. Before removal, the model

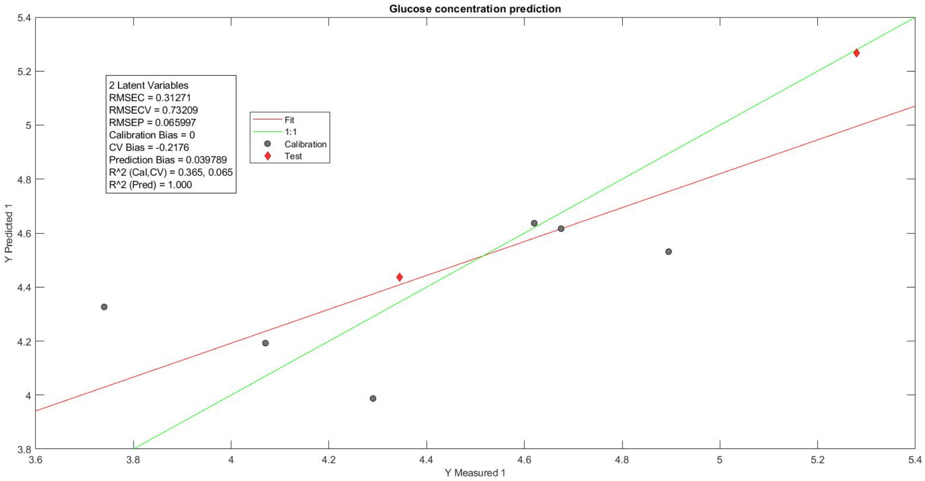


Figure 5.39: Scatter Plot of Measured vs. Predicted Y-values for Glucose. The scatter plot illustrates the model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

demonstrated exceptional accuracy in fitting the training data, as indicated by the low RMSEC (Root Mean Square Error of Calibration) of 0.0336682. However, challenges arose during cross-validation (RMSECV: 0.34924) and prediction on unseen data (RMSEP: 0.637734), suggesting difficulties in generalizing to new observations. The coefficient of determination values (R^2) further emphasized this, with a high R^2 Cal (0.939239) for the calibration set but considerably lower R^2 values for cross-validation (0.542459) and prediction (0.0227324). After the removal of the problematic point, the model's parameters showed notable changes. While RMSEC remained unchanged, maintaining the excellent fit to the training data, RMSEP significantly improved to 0.16619, indicating enhanced predictive accuracy on new observations. The maximum R^2 Pred of 1 after removal signifies a complete explanation of the variation in lactate concentrations on new, unseen data, a substantial improvement from the initial low R^2 Pred. Importantly, the model's generalization ability, as indicated by R^2 CV (0.542459), remained consistent even after the removal. In summary, the removal of the outlier point played a pivotal role in refining the lactate model, addressing challenges in generalization and significantly improving its predictive accuracy. These enhancements contribute to the model's reliability and effectiveness, ensuring more accurate and trustworthy predictions of lactate concentrations in practical applications.

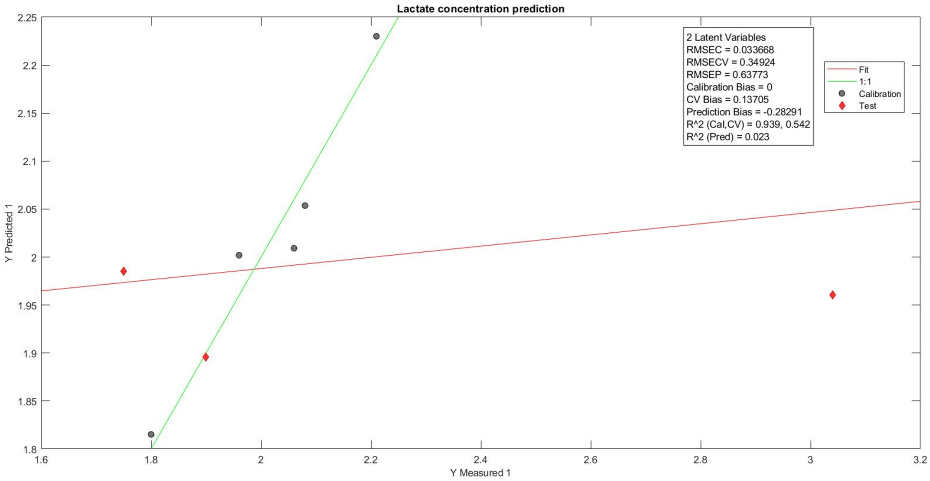


Figure 5.40: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

Conclusion

In conclusion, the exploration of a static culture experiment provided valuable insights into the spectral analysis and predictive modeling of glucose and lactate concentrations. The static culture, realized in a three-dimensional scaffold with continuous exposure to M2 culture medium, offered a unique perspective on the cultivation dynamics.

The spectral analysis, conducted within the optical range from 400 to 1000 nm, allowed for a detailed examination of variations over time, shedding light on the intricate interactions among the components of the M0 soil.

For glucose predictions, the model exhibited outstanding parameters, with low RMSEC, RMSECV, and RMSEP values, indicating high precision, generalization ability, and accuracy on new data. The scatter plot visually confirmed the model's reliability and performance.

In the case of lactate, the initial model faced challenges in generalization, as indicated by relatively higher RMSECV and RMSEP values. The removal of a problematic point led to a significant enhancement, resulting in improved predictive accuracy, as reflected in the reduced RMSEP and elevated R^2 Pred to 1.

These findings emphasize the significance of proactive removal of compromised cultures and outlier points in refining predictive models. The enhanced accuracy and generalization capabilities contribute to the robustness of the models, ensuring their reliability in predicting

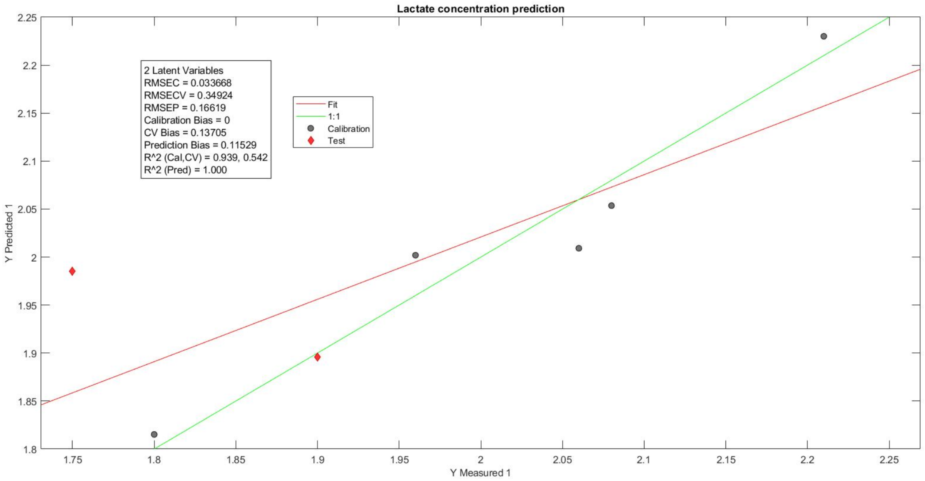


Figure 5.41: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

glucose and lactate concentrations. This experiment, despite the challenge posed by mold around the 16^{th} day, highlights the importance of meticulous data curation and analysis for meaningful outcomes in future non-conditioned cultures.

5.4.5 Dynamic Culture Experiment: Spectral Analysis and Predictive Modeling

In this section, we will create a predictive model using PLS to predict the concentration of lactate and glucose within a DC culture. In this case as well, the culture was sampled on days 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 21, and showed no defects; all samples were available and not deteriorated by molds.

Figure 5.42 depicts the eleven spectra acquired using the spectrometer, all within the visible range. For this specific case, we preferred to consider information starting from the wavelength of 400 nm, as the earlier part exclusively exhibited an initial peak.

The choice to initiate the analysis from 400 nm is motivated by the common practice of excluding information related to the initial wavelengths in spectral analyses. This decision was made to reduce sensitivity to electronic noise, enhance baseline stability, and optimize the predictive model, while ensuring that relevant information for lactate and glucose concentration is preserved in our analysis.

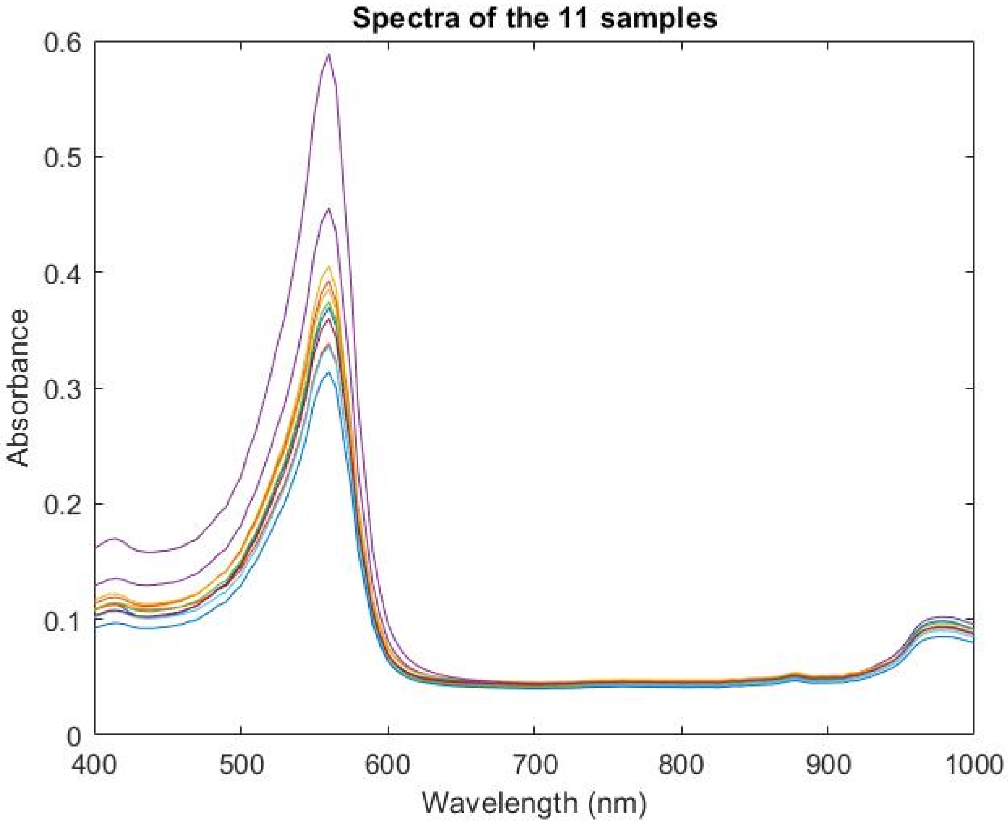


Figure 5.42: Acquired spectra for sample DC from 400 to 1000 nm.

Modeling Results

The analysis of glucose concentration initially presented promising results, showcasing a predictive model with commendable precision. The comprehensive set of statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, underlined the accuracy of the initial model in capturing glucose dynamics within the cell cultures.

However, the model's robustness was put to the test with the inclusion of all samples, leading to an RMSEP of 1.00048 (Figure 5.43). While indicative of decent predictive capability, there was room for enhancement. It was at this juncture that a pivotal decision was made to scrutinize and address the impact of individual samples on predictive accuracy. The removal of the sample at row 8, identified as a potential outlier, brought about a transformative change in the model's performance. Post-removal, the RMSEP saw a substantial drop to 0.75326, signaling a remarkable improvement in prediction precision for glucose concentration. This adjustment, combined with the steady Bias, indicated the model's maintained fairness in pre-

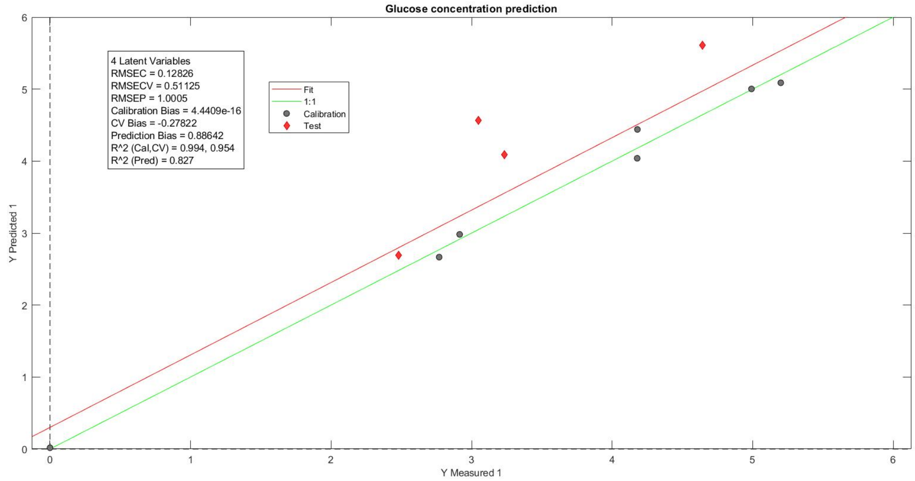


Figure 5.43: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

dictions. The Pred Bias exhibited a noteworthy decline from 0.886418 to 0.676187, reflecting a refinement in the model's predictive tendencies after the removal of the problematic sample. Most notably, the R^2 Pred soared from 0.827095 to an impressive 0.978253, portraying the model's newfound adaptability to new observations with near perfection.

The modeling results for lactate concentration offer a detailed exploration of the predictive model, shedding light on its performance with and without the inclusion of a specific data point.

Upon meticulous examination, Row 9 emerged as an anomaly, evident in its substantial deviation from the model's predictions. This data point, with a measured value (Y Measured) of 2.86, significantly differed from the model's prediction (Y Predicted: 1.33238), resulting in a considerable Y Residual of -1.52762. To ensure the model's reliability, this outlier was excluded from further analyses.

The initial model, considering all samples, demonstrated commendable accuracy with RMSEC (0.235146), RMSECV (0.738069), and RMSEP (0.933185). While the model exhibited a strong fit to calibration data (R^2 Cal: 0.889052), limitations were evident in cross-validation (R^2 CV: 0.228413) and prediction (R^2 Pred: 0.971399). Pred Bias suggested a slight underestimation of values (-0.339336). Post-exclusion of the outlier at Row 9, the lactate prediction model underwent a remarkable transformation. RMSEP saw a substantial reduction from 0.933185 to 0.37342, indicating enhanced precision in predicting new data. Pred Bias shifted

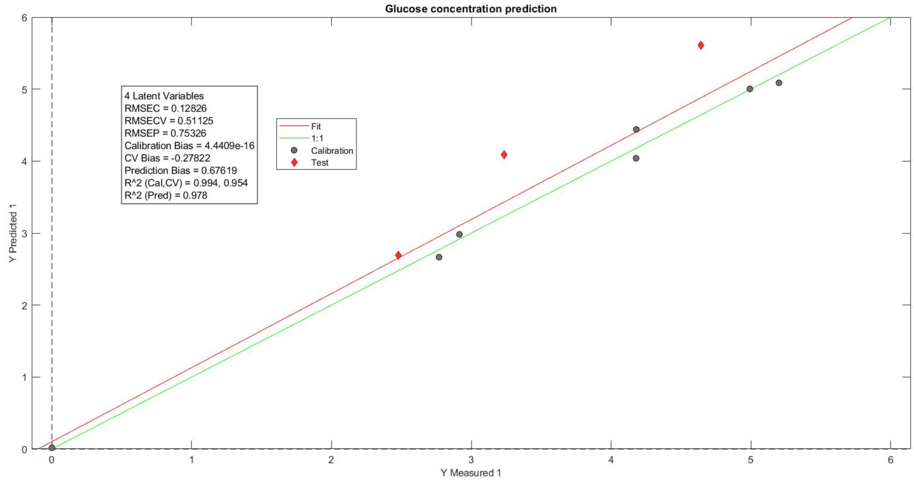


Figure 5.44: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

positively from -0.339336 to 0.254806, reflecting a correction in the model's tendency to underestimate lactate concentrations. Most notably, the R^2 Pred reached a perfect score of 1, showcasing flawless adaptability to new observations. The rationale behind excluding Row 9 was rooted in its substantial deviation from the expected pattern, as evidenced by the significant difference between measured and predicted values. This strategic exclusion not only improved overall model accuracy but also underscored the critical role of meticulous data curation in refining predictive models for precise outcomes in complex analyses.

Conclusion

In this comprehensive exploration of predictive modeling for glucose and lactate concentrations within a DC culture, a meticulous approach was undertaken to refine and enhance the accuracy of the models. The initial phase involved the acquisition of eleven spectra, spanning the visible range, from a culture sampled at various time points. Notably, our analysis focused on wavelengths from 400 nm, excluding the initial portion to reduce electronic noise and optimize baseline stability.

The modeling results for glucose concentration showcased commendable precision initially, as illustrated in Figure 5.43. However, the inclusion of all samples revealed areas for improvement, prompting a closer examination. The pivotal decision to exclude a specific sample (Figure 5.44), identified as a potential outlier, led to a transformative change. Subsequent

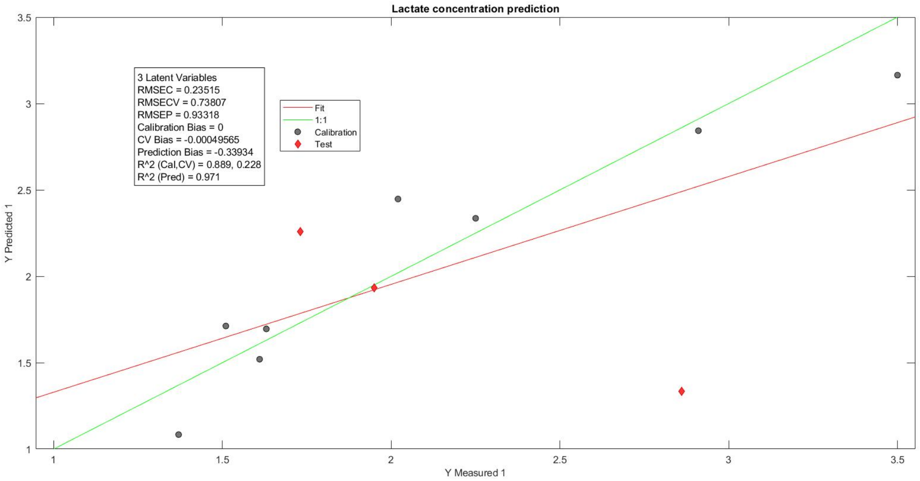


Figure 5.45: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model before Sample Removal). The scatter plot illustrates the initial model's performance in predicting lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

enhancements in prediction parameters, including a substantial drop in RMSEP and an impressive increase in R^2 Pred, emphasized the significance of meticulous sample curation in refining the glucose predictive model.

Similarly, the lactate concentration model underwent a detailed assessment, revealing an anomaly in Row 9 (Figure 5.45). Strategic exclusion of this outlier (Figure 5.46) resulted in a precision boost, with improved RMSEP and R^2 Pred. This process reaffirmed the critical role of data curation in optimizing predictive models.

In conclusion, the iterative refinement of the predictive models for glucose and lactate concentrations demonstrates a commitment to precision and reliability. These enhancements underscore the importance of scrutinizing individual samples, addressing anomalies, and refining models for robust predictions in the dynamic context of cell cultures. This iterative process not only advances the accuracy of predictive models but also highlights the necessity of rigorous data curation in ensuring the reliability of analytical outcomes.

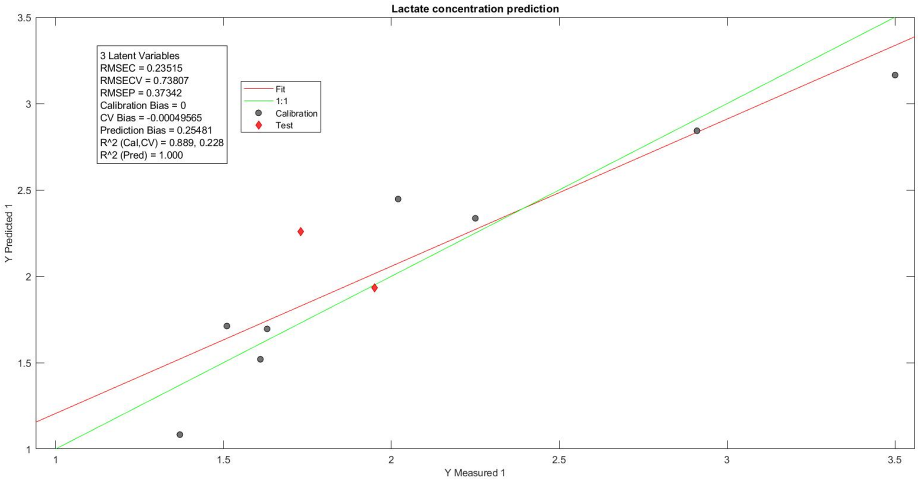


Figure 5.46: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting Lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy.

5.4.6 Non-Conditioned and Static Culture Experiment: Spectral Analysis and Predictive Modeling for Glucose and Lactate Concentrations

In our approach to modeling glucose and lactate concentrations, we initially focused on creating effective predictive models, addressing the complexity arising from differences in culture media. The decision to model individual cultures was motivated by the use of various culture media.

From the outset, we were aware of challenges related to the cultures, including the absence of data for culture S on day 2 and both cultures NC and S on day 4. It is worth noting that despite challenges related to the change in media on day 6 and compromised samples, a robust model for glucose was successfully obtained. Interestingly, NC culture samples are represented by a relative value, obtained by averaging two measurements from distinct samples, without the possibility of tracing back to the original sample. This clarification underscores our targeted approach to addressing the specific challenges of individual cultures, ensuring accurate and informative modeling of glucose and lactate concentrations.

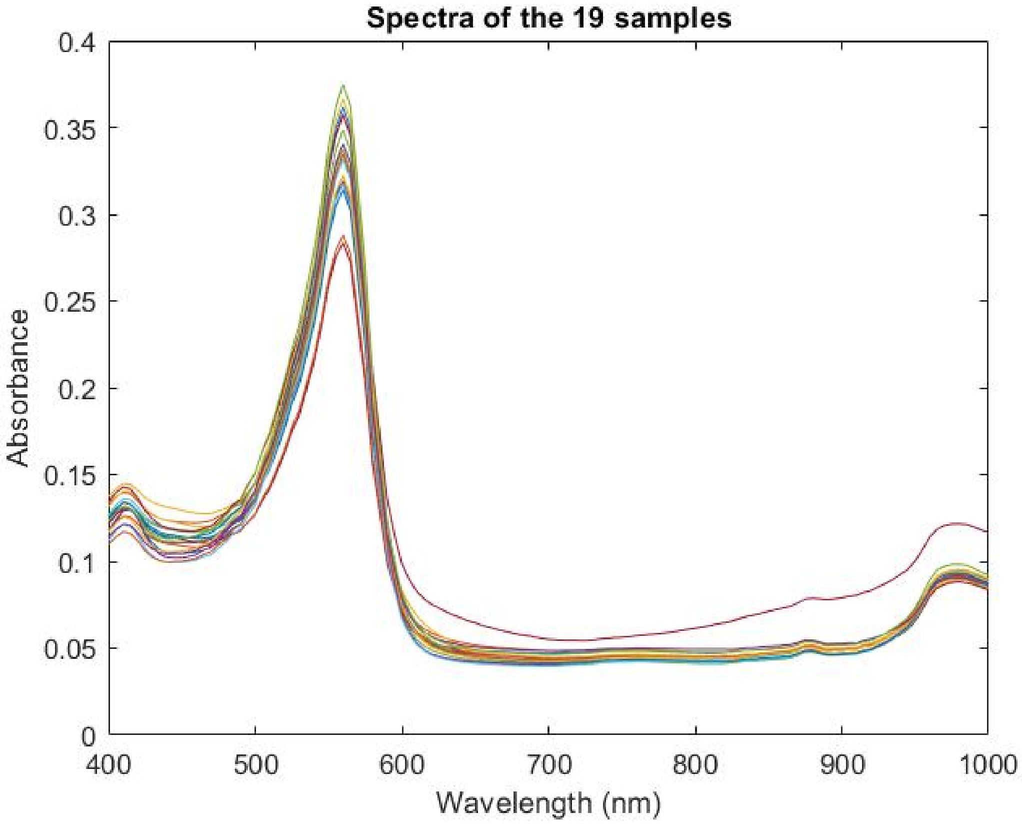


Figure 5.47: Acquired spectra for NC and S samples from 400 to 1000 nm.

Modeling Results

During the analysis of the initial model, an anomalous point was identified in the test set, with a measured value of 0 significantly differing from the predicted value of 2.94108. The removal of this point led to substantial improvements in the model parameters. Subsequent results show an RMSEC of 0.645152, an RMSECV of 1.06682, and an RMSEP of 0.462268. These values indicate good adaptability to training data, reasonable generalization to new data, and acceptable precision in predicting observations not included in the calibration phase. The removal of the disturbance point significantly improved the consistency and accuracy of the model, highlighting the importance of careful data cleaning. Despite intrinsic challenges in the experimental process, the final results underscore the validity of the obtained model. Specifically, the R² Cal, indicating the explanation of variation in training data, reached a value of 0.449466, suggesting a good ability of the model to adapt to such data. The R² CV, measuring the ability to generalize to new data, was 0.0654148, indicating that the model maintains a decent ability to adapt to data not used in the calibration phase. Finally, the R² Pred, repre-

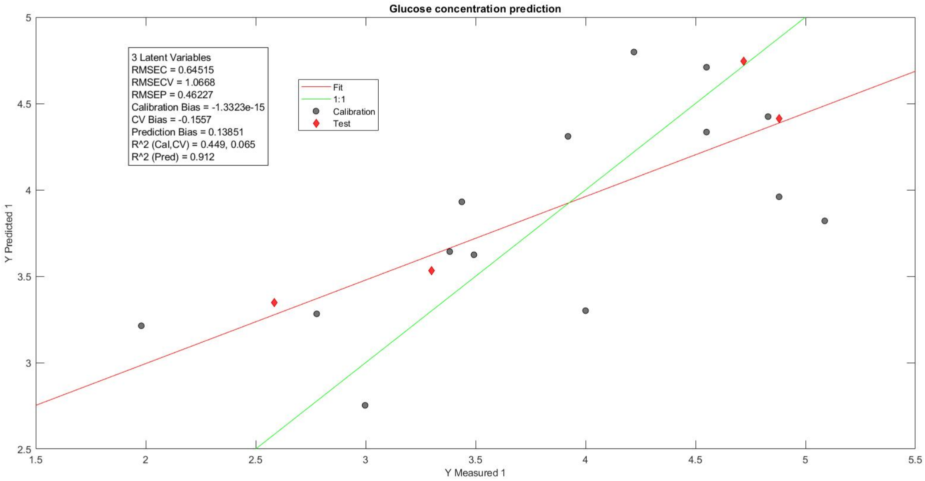


Figure 5.48: Scatter Plot of Measured vs. Predicted Y-values for Glucose (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting Glucose concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy

senting the ability to predict new data, reached the notable value of 0.911628, emphasizing the effectiveness of the model in predicting glucose concentrations outside the calibration phase. Now, we have moved on to lactate. We have again separated the data into test and training sets and applied variable selection. The analysis of lactate model parameters before and after the removal of the two problematic samples reveals significant differences, underscoring the effectiveness of the data cleaning procedure. Before removal, the model had an RMSEC of 0.301662, indicating good adaptability to training data, but a higher RMSECV of 0.682809, suggesting moderate generalization to data not included in the calibration. The RMSEP of 0.64361 indicated acceptable precision in predicting new data. However, determination coefficients (R^2) for calibration (0.790191), cross-validation (0.172615), and prediction (0.647276) indicated challenges in explaining variation. After the removal of the problematic samples, there was a significant improvement in all parameters. The RMSEC decreased to 0.250547, highlighting excellent adaptability to training data. Despite the increase in RMSECV to 0.812982, the model maintains good generalization to new data. The significantly reduced RMSEP to 0.330117 indicates higher precision in predicting data not included in the calibration phase. The determination coefficients R^2 for calibration (0.843587), cross-validation (0.0351263), and prediction (0.922145) reflect a notable improvement in the model's ability to explain variation in the data.

In summary, the removal of problematic samples significantly contributed to refining the lac-

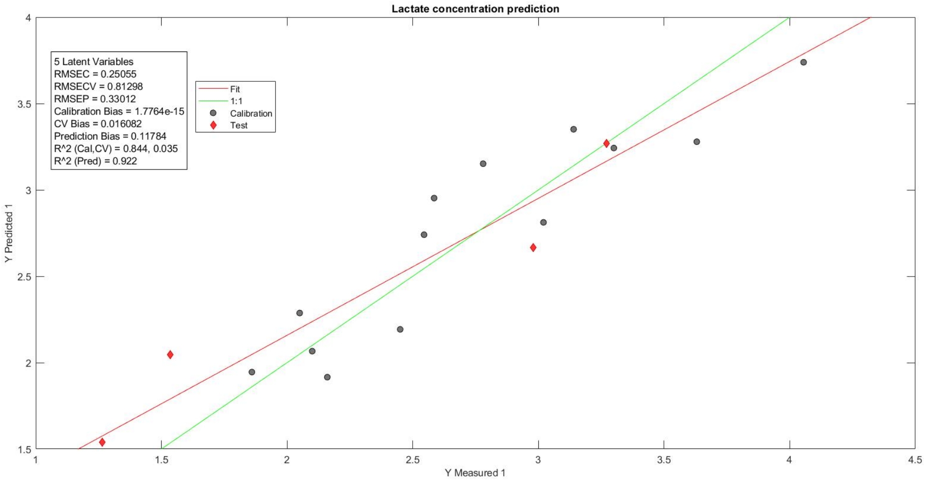


Figure 5.49: Scatter Plot of Measured vs. Predicted Y-values for Lactate (Model after Sample Removal). The scatter plot illustrates the initial model's performance in predicting Lactate concentrations, showcasing high precision and a good fit. Statistical metrics, including RMSEC, RMSECV, RMSEP, Bias, CV Bias, Pred Bias, R^2 Cal, R^2 CV, and R^2 Pred, highlight the model's accuracy

tate model, enhancing its ability to adapt to training data and generalize to new data. This optimization is crucial to ensuring more accurate and reliable predictions of lactate concentrations in practical settings.

Conclusion

In conclusion, the modeling of glucose and lactate concentrations has successfully tackled challenges associated with differences in culture media and compromised samples. The targeted approach to data cleaning has once again proven crucial for obtaining robust and reliable predictive models. The removal of anomalous points has significantly enhanced the consistency and accuracy of the models, underscoring the importance of careful analysis and targeted interventions to address any disturbances in the data. The overall results indicate the validity of the obtained models, contributing to a better understanding and prediction of glucose and lactate concentrations in both experimental and practical settings.

Bibliography

[1] Mostafa H Ahmed, Mohini S Ghatge, and Martin K Safo. “Hemoglobin: structure, function and allostery”. In: *Vertebrate and invertebrate respiratory proteins, lipoproteins and other body fluid proteins* (2020), pp. 345–382.

[2] Fahmida Alam et al. “Lactate biosensing: The emerging point-of-care and personal health monitoring”. In: *Biosensors and Bioelectronics* 117 (2018), pp. 818–829.

[3] Krzysztof B Beć, Justyna Grabska, and Christian W Huck. “NIR spectroscopy of natural medicines supported by novel instrumentation and methods for data analysis and interpretation”. In: *Journal of Pharmaceutical and Biomedical Analysis* 193 (2021), p. 113686.

[4] Krzysztof B Beć and Christian W Huck. “Breakthrough potential in near-infrared spectroscopy: Spectra simulation. A review of recent developments”. In: *Frontiers in chemistry* 7 (2019), p. 48.

[5] Michele Bellancini et al. “Development of a CO2 Sensor for Extracorporeal Life Support Applications”. In: *Sensors* 20.13 (2020), p. 3613.

[6] M Blanco and INIR Villarroya. “NIR spectroscopy: a rapid-response analytical tool”. In: *TrAC Trends in Analytical Chemistry* 21.4 (2002), pp. 240–250.

[7] Nienke Bosschaart et al. “A literature review and novel theoretical approach on the optical properties of whole blood”. In: *Lasers in medical science* 29 (2014), pp. 453–479.

[8] George A Brooks. “Lactate as a fulcrum of metabolism”. In: *Redox biology* 35 (2020), p. 101454.

[9] Scott H Brown. “Multiple linear regression analysis: a matrix approach with MATLAB”. In: *Alabama Journal of Mathematics* 34 (2009), pp. 1–3.

[10] Karthik Budidha et al. “Identification and quantitative determination of lactate using optical spectroscopy—towards a noninvasive tool for early recognition of sepsis”. In: *Sensors* 20.18 (2020), p. 5402.

[11] Subhasri Chatterjee et al. “In-silico investigation towards the non-invasive optical detection of blood lactate”. In: *Scientific Reports* 11.1 (2021), p. 14274.

[12] Chien-Ming Chen et al. “Development of an enzymatic assay system of d-lactate using d-lactate dehydrogenase and a UV-LED fluorescent spectrometer”. In: *Journal of Pharmaceutical and Biomedical Analysis* 116 (2015), pp. 150–155.

[13] Hoeil Chung et al. “Simultaneous measurements of glucose, glutamine, ammonia, lactate, and glutamate in aqueous solutions by near-infrared spectroscopy”. In: *Applied spectroscopy* 50.2 (1996), pp. 270–276.

[14] Lígia Regina Tomás Coelho. “Vascularization: plant decellularization and electrospinning techniques for the development of small and medium caliber blood vessels”. PhD thesis. 2018.

[15] Luke J Currano et al. “Wearable sensor system for detection of lactate in sweat”. In: *Scientific reports* 8.1 (2018), p. 15890.

[16] Mirosław Antoni Czarnecki et al. “Advances in molecular structure and interaction studies using near-infrared spectroscopy”. In: *Chemical reviews* 115.18 (2015), pp. 9707–9744.

[17] P Doshi and Anagha Panditrao. “Optical sensor system for hemoglobin measurement”. In: *International Journal of Computational Engineering Research* 3.7 (2013), pp. 41–5.

[18] Kevin Dunn. *Process Improvement Using Data*. Learnche.org, 2023.

[19] Ahmed Badr Eldin and I Akyar. “Near infra red spectroscopy”. In: *Wide spectra of quality control. InTech, Rijeka, Croatia* (2011), pp. 237–248.

[20] C Lovatt Evans et al. “The utilization of blood sugar and lactate by the heart-lung preparation”. In: *The Journal of Physiology* 80.1 (1933), p. 21.

[21] Christopher D Foucher and Robert E Tubben. “Lactic acidosis”. In: *StatPearls [Internet]*. StatPearls Publishing, 2022.

[22] Francesca Grassi, Daniela Negrini, and Carlo Adolfo Porro. “Fisiologia umana”. In: *Poletto Ed* (2015).

[23] Julian Haas and Boris Mizaikoff. “Advances in mid-infrared spectroscopy for chemical analysis”. In: *Annual Review of Analytical Chemistry* 9 (2016), pp. 45–68.

[24] Rabeay YA Hassan. “Advances in electrochemical nano-biosensors for biomedical and environmental applications: From current work to future perspectives”. In: *Sensors* 22.19 (2022), p. 7539.

[25] H Michael Heise and Robert Schulenburg. “Near-infrared spectroscopy for medical, food and forage applications”. In: *Molecular and Laser Spectroscopy*. Elsevier, 2022, pp. 189–247.

[26] Manish Jadhav et al. “Haemoglobin Measurement Using Non-Invasive Technique: State of the art”. In: *INFORMATION TECHNOLOGY IN INDUSTRY* 9.3 (2021), pp. 669–675.

[27] IS Kucherenko, Ya V Topolnikova, and OO Soldatkin. “Advances in the biosensors for lactate and pyruvate detection for medical applications: A review”. In: *TrAC Trends in Analytical Chemistry* 110 (2019), pp. 160–172.

[28] Xiaolu Li et al. “Lactate metabolism in human health and disease”. In: *Signal transduction and targeted therapy* 7.1 (2022), p. 305.

[29] John Loftus. *On The Development of Control Systems Technology for Fermentation Processes*. The University of Manchester (United Kingdom), 2017.

[30] Michael J McShane, Gerard L Cote, and Clifford H Spiegelman. “Assessment of partial least-squares calibration and wavelength selection for complex near-infrared spectra”. In: *Applied Spectroscopy* 52.6 (1998), pp. 878–884.

[31] Technology Network. *UV-Vis Spectroscopy: Principle, Strengths and Limitations and Applications*. 05 12 2023. 2023. URL: `https://www.technologynetworks.com/analysis/articles/uv-vis-spectroscopy-principle-strengths-and-limitations-and-applications-349865`.

[32] Ion Olaetxea et al. “Machine learning-assisted Raman spectroscopy for pH and lactate sensing in body fluids”. In: *Analytical Chemistry* 92.20 (2020), pp. 13888–13895.

[33] GIUSEPPE Palleschi. “Biosensori in medicina”. In: *Caleidoscopio* 42 (1989).

[34] Margaret E Payne et al. “Printed, flexible lactate sensors: Design considerations before performing on-body measurements”. In: *Scientific reports* 9.1 (2019), p. 13720.

[35] Michael H Penner. “Basic principles of spectroscopy”. In: *Food analysis* (2017), pp. 79–88.

[36] *Spettroscopia infrarossa: spettroscopia del medio infrarosso, del vicino infrarosso e del lontano infrarosso/terahertz.*

[37] Raffaele Vitale et al. “Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial”. In: *Analytica Chimica Acta* (2023), p. 341304.

[38] Gordon CC Yang and Sheng-Wei Chan. “Photocatalytic reduction of chromium (VI) in aqueous solution using dye-sensitized nanoscale ZnO under visible light irradiation”. In: *Journal of Nanoparticle Research* 11 (2009), pp. 221–230.

[39] Xiaomeng Zhang et al. “Functional surface engineering of quantum dot hydrogels for selective fluorescence imaging of extracellular lactate release”. In: *Biosensors and Bioelectronics* 80 (2016), pp. 315–322.

[40] Ivan Zorin et al. “Advances in mid-infrared spectroscopy enabled by supercontinuum laser sources”. In: *Optics Express* 30.4 (2022), pp. 5222–5254.

Acknowledgments

With deep gratitude, I dedicate these words of thanks. This journey has been more than just an academic experience; it has been a personal journey of growth and learning spanning two years. It has been a period marked by significant personal and intellectual growth, characterized by challenging but also rewarding experiences. These years have been a collection of experiences where moments of difficulty intertwined with moments of achievement. Throughout it all, I have had the privilege of meeting numerous individuals whose support and encouragement have propelled me forward, urging me to always strive for the best.

With sincere gratitude, I now wish to thank those who have accompanied me along this journey. Their guidance, encouragement, and belief in my abilities have been pillars of strength during times of uncertainty and difficulty. This thesis represents not only the result of my academic commitment but also the fruit of the collective support and effort of those who have helped me along the path of growth during these years.

I wish to express profound gratitude to my family for their constant support and patience during these years of study. Their encouragement and understanding have been crucial in facing challenges and persevering until the end. In particular, I would like to thank Chiara for her unconditional support and for being by my side at every moment. Our outings to celebrate our successes have been precious moments that have made this journey even more special.

I wish to express my deepest gratitude to Prof. Marco Tartagni for his immense availability and the trust he has placed in me throughout the course of my research. His valuable advice and constant support have significantly contributed to the success of this work.

A special thanks goes to Leonardo and Ouma, who have assisted me at every moment and provided valuable advice when needed. Their support has been essential in facing challenges and overcoming obstacles along the way.

I also want to thank Prof. Emanuele Domenico Giordano and Prof. Joseph Lovecchio for their assistance in the laboratory and for guiding me in the preparation of the necessary samples for my research. A heartfelt thanks also goes to Gianluca Squadrani, a friend and classmate, for his patience and support in conducting laboratory experiments. His availability and help have been essential for the success of this work.

A heartfelt thanks also goes to Eurosets s.r.l. of Medolla, for their valuable contribution to the design and commercialization of biomedical devices. Their suggestions and support regarding

sensor requirements for cardiac surgery, autotransfusion, and vital support in extracorporeal circulation have been extremely helpful and appreciated.
Heartfelt thanks to all of you!