UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
Industrial Engineering

Cycle XXXVIII

TITLE OF THE Ph.D. THESIS

APPLICATION OF DATA ANALYTICS AND SPECTROSCOPY ON THE CHEESEMAKING PRODUCTION CHAIN

Scientific Disciplinary Sector(s)

ICHI-01/C

Ph.D. Student:	Leonardo Sibono
Supervisor:	Prof. Stefania Tronci
Co-Supervisor:	Prof. Massimiliano Grosso

Final exam. Academic Year 2024/2025
Thesis defence session: February 2026

Dedicated to my beloved friend Paola.

Contents

Abstract

In the last decade, the field of dairy has been significantly expanding its horizons through the integration of advanced analytical platforms and data analysis. The impact of such evolution goes beyond the mere dairy production, as it enables the design of new food functionalities (i.e., a food's ability to provide health benefits beyond basic nutrition by positively affecting one or more body functions), enables specific categories of consumers like religious groups, individuals with disrupted gut microbiota or athletes to access and consume products that were previously unavailable or unsuitable for them and allows for the safeguarding of the regional traditions associated with various dairy brands. The production of dairy products such as cheese often encounters several challenges in maintaining the intermediate product's properties within the desired trajectory along the manufacturing chain, due to the high variability of feedstock characteristics and process conditions. Dairy industry is therefore seeking new solutions to support consistent quality, comply with regulatory requirements under high production throughputs and improve process efficiency.

The present thesis aims to provide a rigorous methodology for investigating specific dairy topics that remain somewhat underexplored in the food research. Data analysis techniques were employed to examine the impact of process variables on the food matrix at selected stages of cheesemaking, following the product's path along its process line. As a first step, the issue of milk coagulation process monitoring through non-intrusive sensors is taken under consideration. Specifically, Raman spectroscopy combined with multivariate statistical methods is employed to assess its ability to keep track of milk gelation dynamics and to predict the curd cutting time, a key technological parameter in dairy processing. Accurate cutting time estimations are here reported for both model's calibration and validation. The following step is to exploit the whole potential of Raman spectroscopy in identifying the chemical bonds that behave as tracers for the coagulation process. This identification, integrated with appropriate signal preprocessing and the implementation of a kinetic model, enables the development of a monitoring system capable of providing real-time estimates of both concentration of compounds involved in milk coagulation and the reconstruction of curd's unmeasurable variables such as the elastic modulus, since its measurement requires the invasive extraction of samples. To this concern, the application Kalman filter demonstrates that Raman spectroscopy measurements can provide valuable inline information, even if model parameters are affected by uncertainty. Raman spectroscopy tool has

been then employed for the development of a fault detection system. Specifically, PCA-based multivariate statistical process control algorithms are proposed to analyze Raman spectra and detect the occurrence of the most common faults in the milk coagulation framework (i.e. temperature control and enzyme concentration faults). The overall results show that Raman spectroscopy can provide multitasked assistance in the cheese manufacturing process. The final part of the milk coagulation monitoring problem advances in system complexity, by implementing Raman spectroscopy and multivariate supervised tools to monitor the combined acid-rennet coagulation of milk. Such a system allows to provide a real time estimation of four metabolites involved in lactic acid fermentation (Glucose, galactose, lactose and lactic acid) along with pH. The multivariate models demonstrated good prediction (Q^2>0.5) for the response variable, further confirming the reliability of Raman spectroscopy for tracking dairy product quality within the coagulation stage.

The subsequent part of this dissertation shifts to the analysis of cheese metabolic profile, in order to study the effect of different process variables on the final product quality. The metabolomic study begins with the analysis of fatty acids and elemental composition of cheese sample produced from milk obtained in three different seasons (Summer, spring and winter). The analyses highlight that the levels of conjugated linoleic acid compounds were positively correlated with the spring season, while the length of the saturated fatty acids increased throughout the production seasons. At the same time, the problem of missing data is here addressed. The presence of "*Not a number*" cells in part of the fatty acid matrix is handled through Probabilistic modelling and discrimination algorithms, in order to both classify cheese samples on a production season basis while accounting for missing data and quantifying the missing information. To this regard, high classification accuracies (98 %) and good prediction performances on reconstructing data (R^2_{CV}=99.53%) were obtained. Afterwards, the effect of different sub-pasteurization heat treatments and different ripening times is studied through an untargeted metabolomics approach. Univariate tests on Gas Chromatography–Mass Spectrometry peaks complemented with false discovery rate correction are employed to identify potential biomarkers for different thermal treatment and ripening time for PDO Fiore Sardo cheese, resulting in the significance of biogenic amines, endocannabinoids, sugars and organic acids. Both multivariate and univariate models were employed to discern cheese samples produced under different thermization severity and ripening times, achieving acceptable accuracies (Accuracy in cross validation > 70%). Finally, the differences in the lipidic profile of Pecorino Romano cheese over different rennet sources are here investigated using liquid

chromatography-quadrupole time-of-flight tandem mass spectrometry. For the first time, a lipidomic study is conducted to differentiate vegetable-renneted (Cardoon) cheese samples from calf rennet-based product. The interaction between different rennet sources and cheese ripening times is also considered. Due to the high dimensionality and the presence of numerous non-informative variables within the lipidomic dataset, the performance of classification models can be significantly hindered. To address this issue, the statistical procedure proposed in this study foresees a variable selection strategy aimed at improving the classification accuracy of multivariate classifiers. In fact, when distinguishing between different types of rennet, the classifier's accuracy increases markedly (from 56% to 88%) when using selected features instead of the full set. The selected features are considered as biomarkers for distinguishing among the different renneting recipes. The biomarkers span several major lipid classes, including glycerophospholipids, triacylglycerols, monoacylglycerols, sphingolipids, various fatty acids and cholesterols.

Overall, the outcomes presented in this dissertation demonstrate that spectroscopic platforms coupled with data analytics tools can be a promising tool within the cheese manufacturing workflow, offering a holistic perspective which considers the production processes as an interconnected system rather than a set of isolated units.

Chapter 1: Introduction

This chapter delineates the motivations driving the thesis presented herein. A general discussion on the innovative contribution of the research is also provided, aimed at contextualizing the work within the broader scientific framework. More specific and technical aspects of innovation are addressed in detail in the subsequent chapters. Additionally, it offers an overview of the manuscript structure, by encompassing the topics discussed in each section of the manuscript. The conclusion of Chapter 1 comprises a list of conference participations and journal papers authored throughout the Ph.D. program.

1.1 Dairy industry: an overview

A dairy product is any food item that is derived from the milk of mammals such as cows, goats, and sheeps, and is processed through various techniques such as fermentation, coagulation, or churning. Common examples include milk, cheese, yogurt, butter, and cream to name a few (Roohi et al., 2018). Their production is performed in the framework of the "dairy processes", which generally involve the transformation of raw milk into a specific product variety. Dairy products are considered of high importance for their high nutritional and functional contents which confer them a crucial role in human diets across the world (Visioli & Strata, 2014). The dairy industry, like many other sectors of biotechnologies, has undergone a deep transformation in parallel with the major industrial revolutions. Starting from the late 18th century with the emergence of mechanization during industry 1.0, dairy processing began to shift from full-handcrafted operations to more structured methods which included the first genetic selection approaches, rudimentary veterinary care and primitive process mechanization empowered by the introduction of steam power. Information was mainly managed by oral tradition. With the advent of industry 2.0 in the late 19th and early 20th centuries, characterized by mass production and the use of electricity, dairy plants adopted more advanced machinery for pasteurization, refrigeration, bottling and milking, laying the foundation for modern hygiene and preservation standards. In the same period formulated feeds and controlled breeding practices were introduced in dairy processing, along with the first paper-based records, which paved the path towards the current data management procedures (Sangode, 2025). Industry 3.0, driven by automation and digital electronics in the second half of the 20th century, brought programmable logic controllers, process control systems and improved supply chain integration to the dairy sector. Basic data analysis is introduced in this period, promoting the development of standard milk storage techniques, disease management and pedigree selection (Hassoun et al., 2023; Sangode, 2025). Today, the global transition to Industry 4.0 is redefining dairy production through the integration of artificial intelligence, Internet of Things technologies, advanced data analytics, and Process Analytical Technology (Hassoun et al., 2023). These technological advancements are rapidly gaining momentum, giving rise to a newer version of dairy industry, which now refers to as Dairy 4.0 (Gehlot et al., 2022).

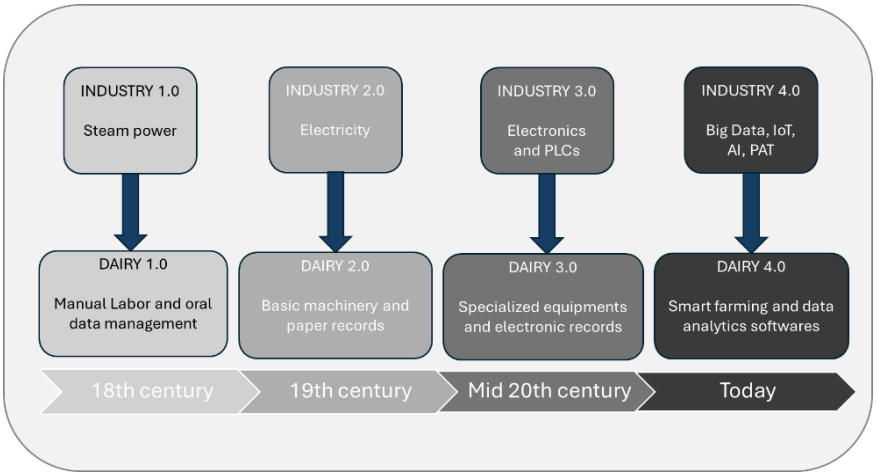


Figure 1.1 History of dairy practices and data management and their relationship with the respective industrial revolutions.

Such technological advancements were born to respond to the increasing market demand of dairy products. In 2023 the production of raw milk in EU's farms was estimated to be 160.8 million tonnes, which is the result of a steady yearly production increase of 0.8 million tonnes per year (EUROSTAT, 2025). Out of 160.8 million tonnes produced, 149.3 million tonnes were available to dairy sector in the same year. Approximately 70% of this amount has been used to make cheese and butter. Figure 1.2 presents the time evolution of total raw milk utilization in the dairy sector across different countries over the period 2014–2023. As it can be seen, with the exception of France, all the main milk-producer countries under investigation showed an yearly increase of raw milk utilization. The steady increase in raw milk volumes underscores the growing demand for dairy products and the corresponding pressure on production systems. This upward trend not only highlights the importance of optimizing processing capacity but also justifies the integration of advanced technologies aimed at improving efficiency while ensuring product quality.

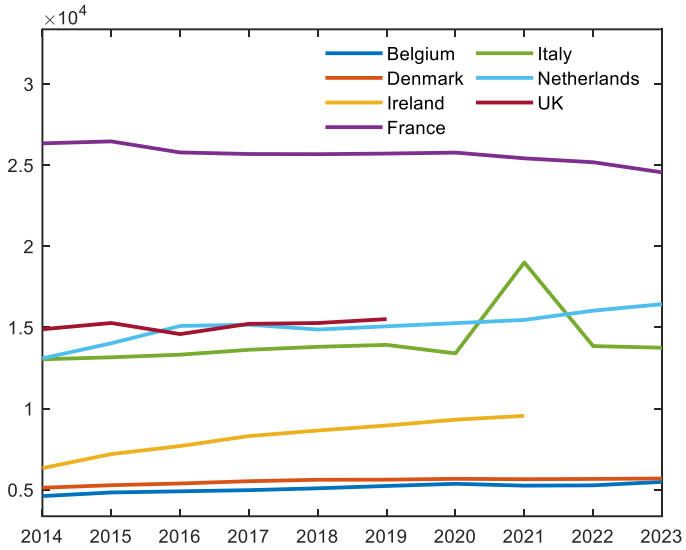


Figure 1.2 Total utilization of the raw milk available in dairies. Values are expressed in 1000 tonnes. Data were retreived from EUROSTAT (2023) datasets. Missing entries reflect incomplete or unavailable national submissions to Eurostat. Last updated: 22/05/2025

1.2 Motivation of the study

Despite its pivotal importance in the food sector, dairy industry has to face a range of complex challenges that threaten its sustainability, efficiency, and profitability. These challenges can be generally subdivided into four subcategories:

1) Economical challenges: along with the high productivity, high market instability is reaching the dairy market due to the strong price volatility of milk. Particularly, the sensitive nature of milk related to seasonal variations, unanticipated variation of supplies and shifts in product demand makes it difficult to plan long-term operations in dairy business (Rezitis & Kastner, 2021). Moreover, in the last few years, the market of dairy milk has to compete with that of plant-based alternatives, which are gaining appeal due to the sustainability and health concerns (e.g. milk intolerances), which poses the necessity of economical assessment of milk products manufacturing operations (Cardello et al., 2022).

2) Environmental challenges: despite the high nutritional value of milk-based products, the dairy industry is acknowledged as a large source of greenhouse gases emission, characterized by high water demand and waste production (Milani et al., 2011). Modern dairy practices are focusing on the environmental impact of the inherent processes and are developing new solutions to reduce the pollution pressure posed by wastewaters, packing and energy consumption for milk processing.
3) Animal welfare and ethical challenges: the pursuit of higher milk productivity must be balanced with the need to safeguard the health and well-being of dairy cattle, ensuring that ethical standards are upheld and intensive farming practices are avoided (Barkema et al., 2015). In this framework, genetics and genomic breeding have been implemented for improving cattle health and productivity (Barkema et al., 2015; Koeck et al., 2014).
4) Process-related challenges: the increase of milk productivity also introduces critical challenges related to hygiene management, particularly in terms of clean-in-place protocols, where the complexity of high volume production systems may pose contamination risks (Chowdhury et al., 2025). Additionally, the emphasis on product standardization, often driven by industrial-scale production goals, may conflict with the preservation of product diversity and traditional characteristics. This issue becomes particularly evident in the context of PDO (Protected Designation of Origin) labels, which require strict adherence to traditional production methods, regional raw materials, and defined sensory profiles. The pressure to scale up production while maintaining PDO compliance raises concerns about the authenticity and the potential dilution of cultural and geographical identity in favor of process uniformity (Laurent et al., 2017; Sibono et al., 2024).

Specific attention is here given to the fourth point, the process-related challenges. The focus on this topic is justified by a preliminary review of the literature, that suggests a relative gap in the technological advancement of dairy product quality monitoring and analysis (particularly in certain stages of cheese-making) when compared to the more extensively explored dimensions of economic, environmental, and animal welfare issues. A bibliometric research on Scopus database was performed to check the number of documents published between 1931 and 2026 regarding the specific fields associated to each of the four topics discussed above. Table 1.1 shows the search queries relevant to each field.

Table 1.1 Bibliometric research queries relevant to each topic of interest in the industrial challenges of dairy framework.

	Economic challenges	Environmental challenges	Animal welfare and ethical challenges	Process-related challenges
Queries	("Market" OR "production costs" OR "profitability" OR "economics") AND ("dairy" OR "milk production" OR "cheese")	("environmental impact" OR "carbon footprint" OR "sustainability" OR "wastewater") AND ("dairy" OR "milk production" OR "cheese")	("animal feed" OR "cow health" OR "livestock ethics" OR "genomics" OR "genetic selection") AND ("dairy" OR "milk production" OR "cheese")	("PAT" OR "metabolomics" OR "milk coagulation" OR "quality control" OR "process monitoring" OR "PDO") AND ("dairy" OR "milk production" OR "cheese")

It is worth specifying that a broader use of OR logic was applied in the formulation of the query associated to process-related challenges, aiming to emphasize any eventual discrepancy in the number of publications. From the bibliometric analysis, it resulted that the largest number of documents was from the Economical field (14364 documents), followed by the animal welfare and ethics (11111 documents) and the environmental analysis (10301 documents). In contrast, the number of publications addressing process-related aspects was significantly lower (5625), suggesting that this area remains relatively underexplored within the current scientific literature. Results are reported in figure 1.3.

The bibliometric research suggests that there is a relative gap in the technological advancement of process monitoring and quality analysis (particularly in certain stages of dairy production, such as cheese-making) when compared to the more extensively explored dimensions of economic, environmental, and animal welfare issues, albeit such topics are strongly interconnected between each other.

Therefore, the research effort presented in this thesis arises from the notable discrepancy between the attention given to diverse dairy subtopics and that of dairy food quality monitoring

(in the broadest sense of the term), which is the cornerstone of technological advancement in this field.

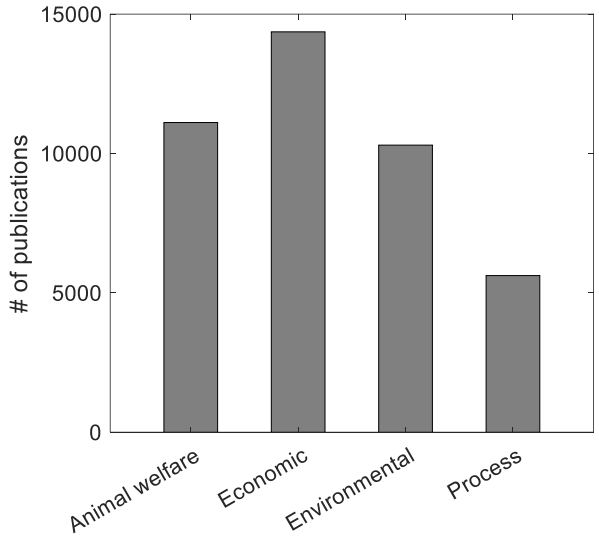


Figure 1.3 Number of publications retrieved for each investigated challenge. These data are updated to June 2025.

Upon this gap, it becomes apparent that current methodologies for industrial process control in dairy food production need to be extended. In this framework there is a progressive shift toward data-rich and interdisciplinary approaches (Agregán et al., 2021). These strategies aim to integrate advanced sensors, data analytics and process modeling to describe the inherent complexity of dairy systems and enable real-time decision making along the production chain. In the last two decades, spectroscopic techniques gained traction due to their versatility and potential to serve different analytical purposes along the dairy production chain. Different technologies are being developed with the aim of maximizing information acquisition from the process, tailored to the specific objectives of the analysis.

A common denominator across many of these analytical approaches is the application of Chemometrics, which has proven to be a reliable and powerful tool for food quality assessment (Smaoui et al., 2023). Chemometrics can be defined as "*the science of relating chemical measurements made on a chemical system to the property of interest through the application of mathematical or statistical methods*" (Wulandari et al., 2022). Such discipline is studied and employed to design experiments, analyze chemical data, and extract meaningful information from

complex, multivariate datasets typically generated by analytical instruments. In the field of food processes, large data sets collected from spectroscopical techniques are analyzed through chemometrics to obtain the relevant knowledge of the food matrix, aiming at gaining understanding of how various stimuli influence specific properties of interest, which are target of quality monitoring and control (Panikuttira et al., 2018). Despite chemometrics and pattern recognition emerged in the 1970s (Wold, 1976), the field has continued to evolve, driven by advances in computational power, machine learning algorithms, and the increasing complexity of analytical data (Li et al., 2025). Particularly, the constant development of advanced analytical devices has made the analysis of complex systems increasingly dependent on technologies that produce high-dimensional data (Li et al., 2025). These analytical instruments include Raman spectroscopy, InfraRed (IR), fluorescence, Nuclear Magnetic Resonance (NMR) and Mass Spectroscopy (MS)-based chromatography, which recently proved to be applied efficiently within the dairy process line (Caboni et al., 2019; Panikuttira et al., 2019; Puerta et al., 2017; Scano et al., 2019). Nevertheless, the application of these platforms in the dairy sector still demands to be further expanded. At present, these techniques are often studied individually, each from a highly specific point of view, neglecting their potential integration into a more comprehensive, process-wise perspective (Hayes et al., 2023).

The integration of chemometrics with spectroscopic methods such as Raman, IR, and NIR enables the implementation of Process Analytical Technology (PAT) strategies for real-time monitoring and control of dairy food quality (Panikuttira et al., 2018). On the other hand, the combination of chemometrics with data-intensive platforms like MS chromatography allows to explore in detail the molecular profile of dairy products. These two branches of knowledge, which cover the methodological rationale of this dissertation, are further discussed in the following subsections.

1.2.1 Application of PAT in dairy industry

The term Process Analytical Technology gained significant recognition following the 2004 publication of the Guidance for Industry by the U.S. Food and Drug Administration (FDA). In this document, PAT is defined as "*a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality*" (FDA, 2004). PAT principles were introduced to improve process understanding and

ensure consistent product quality through real-time monitoring, which is an approach that relies on time-resolved measurements to determine whether subsequent actions should be performed. Although originally developed for pharmaceutical manufacturing, its principles have been extended to other sectors, including food processing, where they offer valuable tools for enhancing quality control and production efficiency (Pomerantsev & Rodionova, 2012). The growing interest in extending PAT rationale within diverse sectors, including dairy, lies in its ability to provide non-invasive measurements during processing, thanks to the non-contact principles of the sensors featured in the PAT scheme (Minnich et al., 2016). However, the successful implementation of these principles requires a deep understanding of how each stage of the production chain contributes to the overall quality of the final product. To this concern, Pomerantsev and Rodionova (2012) reported a five step systematic procedure for PAT implementation, which are briefly summarized in the following list:

1) Feasibility analysis: preliminary laboratory-phase evaluation of a proposed analytical solution. This phase aims to verify whether the method performs reliably under controlled conditions and to demonstrate its potential to the manufacturer. Feasibility studies typically include testing the models' quantification ability and their robustness under varying process conditions, before moving to validation procedures.
2) Quality by Design (QbD): QbD approach enables systematic design, analysis, and control of manufacturing processes by identifying, early on, the performance attributes that define the characteristics of the end-product. Central to this strategy is the definition of Critical Process Parameters (CPPs), which are the key operational variables that must be monitored and controlled to ensure quality consistency (Maillot et al., 2021). These parameters directly influence the Critical Quality Attributes (CQAs), which are the specific characteristics of the final product that are directly related to its quality, safety, or efficacy (Schofield et al., 2015). By establishing the relationship between CPPs and CQAs, QbD facilitates the production of consistently high-quality products, while also contributing to reduced waste and manufacturing costs. Importantly, QbD principles should be applied both to process development and to final product design.
3) Actions before the process: a crucial aspect of PAT is the routine monitoring of raw material properties prior to processing. In the pharmaceutical industry, this ensures batch-to-batch consistency and compliance with quality standards. In the food sector, raw material variability is often biological in origin and can significantly influence product

quality. Such an aspect is explored in this work through the study of the effect of milk seasonality.

4) Actions in the process: transferring PAT solutions from laboratory or pilot scale to full-scale industrial settings often encounters significant challenges such as nonideal production conditions, limited flexibility in modifying process parameters, and the inability to test systems under extreme fault scenarios. The implementation of inline and online monitoring tools in real world scenarios is a critical step of path implementation.
5) Actions after the process: although the implementation of PAT solutions can reduce the need for extensive end-product testing, it cannot fully replace final quality control, which remains essential to confirm that product quality meets the required standards. Additionally, the reliability of the models employed in the production line may degrade over time due to various factors (e.g. aging of sensors and detectors or environmental variations). Therefore, continuous validation and maintenance of predictive models by means of end-product verifications is a critical component of any PAT strategy.

Before delving into the essence of the PAT application in dairy it is important to define the classification of PAT measurement systems according to their connectivity with the process. In particular, measurement systems are classified into four main categories (Minnich et al., 2016):

Offline: the sample is taken from the process and analyzed in a separate laboratory. Such category does not belong to real-time measurement class.

Atline: the sample is collected and real-time analyzed near the production area using a device that is disconnected from the process.

Online: measurement is performed automatically and continuously, with sample diverted from the process stream and introduced into the analyzer by means of a by-pass.

Inline: In situ measurement. The sensor is directly inserted into the process stream for real-time measurement without any sample removal.

PAT solutions have to provide a real time picture of the process state through inline or online analyzers.

In the last few years the dairy food sector has been starting to integrate PAT paradigm following the path of implementation of other high-interest industrial areas (Grassi et al., 2022). Cheese production involves several processing steps that affect the quality of the final product which

require the implementation of PAT tools. Specifically, five cheesemaking stages are suited to the adoption of PAT (Panikuttira et al., 2018):

1) Milk quality composition
2) Standardization of milk
3) Milk coagulation and syneresis
4) Moisture and salt level control of curd
5) Ripening of cheese

Several tools have been applied to develop online monitoring systems for such processing steps. Generally, milk quality composition and standardization are monitored through spectroscopic techniques such as MIR and FT-IR (Cassoli et al., 2011; Panikuttira et al., 2019; Soyeurt et al., 2009); NIR, fluorescence and light backscattering are used to monitor the milk coagulation process (Castillo, 2006; Felfoul et al., 2022; Lyndgaard et al., 2012; Nicolau et al., 2015), dielectric spectroscopy is used to monitor cut curd moisture and salt content (Fagan et al., 2005) and imaging techniques for cheese ripening monitoring (Alinaghi et al., 2023; Loddo et al., 2022).

The production of consistent high-quality product while following the increase of production rates is still an open challenge (Panikuttira et al., 2018). Such an issue arises from the strong sensitivity of cheese properties to the quality of raw and in-process materials. There is the need to develop advanced tools capable of detecting variability in the properties of dairy products along their production chain, supporting high-throughput manufacturing while complying with sanitary standards and process environment constraints. The employment of complex chemometric strategies to develop and validate prediction model is the future trend of PAT implementation in cheese manufacturing, which will lead to improved traceability across the production line, supported by the automation of the key process stages.

1.2.2 Foodomics and Metabolomics

In the context of living systems analysis, omics emerges as a set of various disciplines which aim at quantify and identify different biomolecules such as genes, metabolites and mRNA, in order to relate which characterizes a specific state and time of a biological substrate (Subedi et al., 2022). An important goal of this group of disciplines within the food processes field is to perform molecular profiling of the samples of interest subjected to different stimuli, in order to identify which of the investigated molecules serve as indicators of those specific conditions. These

molecules are commonly referred to as biomarkers (García-Cañas et al., 2010). When multiple omics disciplines are applied (either individually or in combination) to the field of nutrition, the resulting area of study is referred to as foodomics (Ibáñez et al., 2013). Three different approaches emerge in foodomics (Aslam et al., 2017; Ibáñez et al., 2013):

Genomics: the study of the complete genetic material of organisms. In the food sector, genomics is applied to identify desirable traits, enhance breeding programs, and optimize microbial strains for fermentation processes.

Proteomics: the large-scale study of the full set of proteins expressed in a biological system. In the context of food, proteomics is employed to monitor protein composition changes during processing or identify bioactive peptides.

Metabolomics: the comprehensive analysis of low-molecular-weight metabolites present in a biological sample. Metabolomics is crucial for profiling the biochemical fingerprint of food products when subjected to different process stimuli, geographical origins or other technological treatments.

Such main categories branch out into different subdisciplines such as epigenomics, peptidomics, lipidomics and metallomics, to name a few (Capozzi & Bordoni, 2013). Current metabolomics strategies mainly rely on MS and NMR approaches (Alseekh & Fernie, 2018), embracing several fields of research such as food, medicine and drugs development (Buescher & Driggers, 2016; Caboni et al., 2020; Del Carratore et al., 2016).

Metabolomics can be broadly categorized into two approaches: targeted and untargeted. Targeted metabolomics is hypothesis-driven and focuses on the quantification of a predefined set of metabolites, typically associated with specific pathways or physiological functions. This method facilitates the data analysis due to the limited number of molecules under study along with the possibility to perform their absolute quantification (Gorrochategui et al., 2016). In contrast, untargeted metabolomics adopts an exploratory, hypothesis-generating approach, aiming to capture a comprehensive snapshot of the metabolome without prior selection of specific compounds. This approach is particularly valuable in food science for uncovering novel markers associated with technological treatments, geographical origin, or spoilage processes (Fabiano et al., 2011; Gorrochategui et al., 2016; Sibono et al., 2025). Although more complex in terms of data processing and interpretation, untargeted metabolomics provides a broad view of the biochemical

picture of food matrices, enabling a deep understanding of how processing conditions influence food quality at the molecular level. On the other hand, when using single spectroscopic platforms, a complete metabolome coverage and precise annotation are often prevented from the vast chemical diversity of metabolites and the intrinsic constraints of current analytical methods, presence of isomers and other factors that contribute in the occurrence of false annotations (Nash & Dunn, 2019).

Despite the dairy research benefits the successful applications of foodomics, metabolomics and their relevant subcategories there are still two main open challenges that need to be addressed:

1) Several process-related factors require in-depth analytical investigation to understand their impact on the final product quality. For instance, the influence of different thermal treatments on the cheese metabolome and its interplay with the ripening process (Caboni et al., 2019), the effect of various rennet types on the lipidomic profile of cheese, or the preservation of traditional features of a specific dairy product (Sibono et al., 2024).
2) The integration of multiple analytical approaches often generates high-dimensional and complex datasets, which can present significant challenges in terms of data interpretation and meaningful extraction of information (Valdés et al., 2022).

Overall, while foodomics offers valuable tools for the analysis of food systems, its effective application still requires the refinement of the analytical and data analysis methods, which will make these approaches more accessible and practically useful across the food industry.

1.3 Goals of the research

This dissertation focuses on the current challenges of dairy industry using a *process-oriented* approach: the specific process stages of the cheesemaking production line are investigated through data analytics. However, such process steps were not considered separately between each other but were instead considered as a portion of the whole process chain, which is characterized by a complex interaction between its parts. Due to technical limitations, it was not possible to provide a full quantitative assessment of the interconnection among all process steps, but the intentions of this study are to propose a set of rigorous methods to evaluate dairy food quality, by following the flow line of the cheesemaking chain, according to the scheme provided in figure 1.4. Despite the numerous variations of cheesemaking process which depend on product

type and manufacturing practices, the diagram presented outlines the generic structure that was considered in this study.

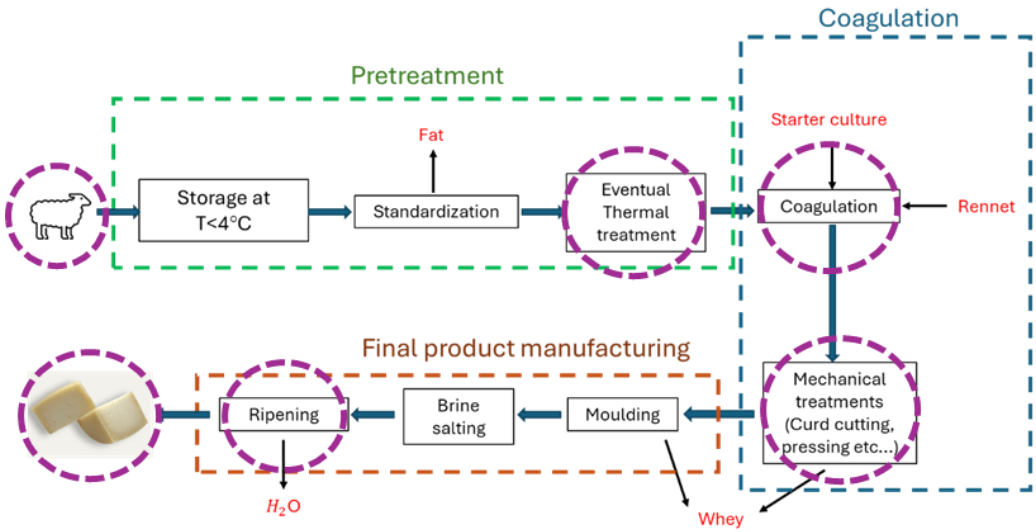


Figure 1.4 General process flow of cheesemaking, illustrating the main stages from raw milk to the final cheese product. Violet, dashed circles represent the process steps subjected to investigation in this thesis.

The present study integrates two spectroscopically-based frameworks: metabolomics, used for the characterization of dairy product metabolites, and PAT, applied for real-time monitoring and control of processing conditions. The application of such disciplines supported the development of the following innovative contributions:

- Application of Raman spectroscopy for milk rennet coagulaltion monitoring: NIR Spectroscopy is the analytical platform that is typically employed for milk gelation. Despite such technique allowed to overcome the problems related to other invasive or atline measurements (knife test, rheology, lactodynamography) its high sensitivity to water content prevents to track part of the signal of interest provided by organic molecules. In this thesis Raman spectroscopy is employed to fulfill various tasks in the milk coagulation monitoring framework, including cutting time prediction, elastic modulus prediction, and fault detection.
- Monitoring the combined acid-enzymatic coagulation of milk: although milk renneting is one of the most extensively studied topics in coagulation research, the combined use of starter cultures and rennet remains the most preferred approach in the dairy industry. This is due to the high functionality of lactic acid bacteria (LAB), which contributes to the

development of the distinctive sensory and nutritional characteristics of cheese. Despite the growing demand for extending the monitoring techniques in the dairy manufacturing chain (Hayes et al., 2023), the integration of different spectroscopic sensors for monitoring milk coagulation is yet to be fully implemented. In the preliminary study reported in this dissertation, Raman and pH are employed to monitor the acid-rennet coagulation of milk, offering a valuable strategy to estimate the concentration of acidification metabolites without relying on unsuitable measurements based on chromatographic techniques.

- Missing data inference from multiblock data structures: matrices affected by missing values are often encountered when using different analytical platforms, including data from dairy samples. In this study, various techniques are employed to reconstruct the cheese fatty acid profiles by using the partial information provided by metal content. Such reconstruction is then used to enhance the discrimination performances of a linear classifier tasked to discern cheese samples according to their production season.
- Combined effect of thermal treatment and ripening time on cheese metabolic profile: The effect of different process stimuli, despite being widely studied, yet remains to be extended. Particularly, the type of thermal treatment is strongly related to technological protocol to be followed in order to comply with Protected Designation of Origin (PDO) requirements for a given product. The detection of specific biomarkers can assist in identifying cheese products that do not comply with the production guidelines established by the certification schemes.
- Lipidomic profile of plant and animal-based rennet: The use of vegetal rennet in cheese production serves as a trademark for dairy products intended to meet vegetarian and certain religious dietary requirements. While the impact of different rennet sources on the cheese peptidomic profile has been extensively studied due to the enzyme's central role in shaping the proteic composition of the curd, the potential effects on the lipidic profile have received no attention. The contribution provided in the present dissertation can promote the expansion of the set of biomarkers useful for distinguishing cheeses produced with different types of rennet.
- Application of tailored variable selection methods for foodomics problems: When dealing with complex MS-based platforms that generate high-dimensional datasets, the presence of numerous uninformative variables can compromise the performance of classification

models. To address this, variable selection problem was undertaken prior to models calibration, aiming to enhance classification accuracy by eliminating irrelevant or noisy features.

It is important to specify that the experimental choices were tailored to align with the specific objectives of each contribution presented above. Specifically, the metabolomic investigations were conducted on sheep milk, as it represents a key raw material for the production of several PDO-labeled cheeses in the Sardinian region, the main location of experimental analysis. On the other hand, the development of the PAT-based monitoring system required the use of reconstituted cow milk powder. This choice was made to facilitate standardization, repeatability, and material handling during the system development. In light of this, the selection of the raw materials made in this research effort is expected to simplify the methodology application.

In conclusion, this work embraces a holistic perspective by combining molecular-level information obtained through metabolomics with process-level insights provided by PAT tools, employed to achieve a broader knowledge of how specific processing decisions influence final product quality.

1.4 Thesis outline

The structure of this dissertation reflects the evolving state of the food matrix under examination along the production process, in accordance with the main stages of the cheesemaking process outlined in the reference flowchart (Figure 1.4). Accordingly, the thesis begins with the development of the monitoring system for the milk coagulation process, as this stage directly involves the milk, and then moves towards metabolomic analyses conducted on the final cheese product. Therefore, the content of this thesis is structured as follows.

Chapter 2 describes the mathematical methodologies employed in this work. The models are presented in their general form, while the specific applications and adaptations are detailed in the subsequent chapters.

Chapter 3 addresses the problem of milk coagulation monitoring through a novel non-invasive spectroscopic technique. After a short depiction of the state of the art, the methodology behind the application of Raman spectroscopy as a PAT tool for process monitoring is presented. Afterwards, the results obtained by such methodology are reported. The multivariate spectroscopy signal provided by Raman sensors is used to predict the cutting time, to shed light on the chemical bonds related to rennet coagulation of milk and finally develop a kinetic model of the process.

In chapter 4, the problem of coagulation monitoring is taken closer to the industrial demand by focusing on the development of a fault detection system of milk rennet coagulation. First, different fault detection algorithms are explored to identify the optimal configuration, which yielded the maximum fault detection rate while minimizing the false alarm rate when temperature and rennet concentration anomalies were introduced in the system. Once the optimal detection model is identified, a fault diagnosis procedure is developed to understand the source of the process failure and eventually assess whether correcting actions can be taken.

Chapter 5 accounts for the application of Raman spectroscopy to monitor the dynamic evolution of acid-rennet coagulation of milk. In this study, the rationale presented in chapter 3 is applied to the complex dynamics of bacterial fermentation of milk by merging an online spectroscopic tool with the offline measurements of sugars and organic acids concentration, for the sake of process monitoring. In this study the pH decreasing during coagulation is also considered as a reference tool. This study bridges the application of PAT with the offline analytical perspective provided by

metabolomics, where high-resolution MS techniques are employed to characterize the cheese matrix, as suggested by the *Actions after the process* practices.

After examining the PAT rationale, the cheese metabolomic point of view is taken into account following the process scheme (Figure 1.4).

Chapter 6 examines the influence of milk seasonality on the fatty acid and elemental composition of Pecorino Romano PDO cheese. Special attention is given to Conjugated Linoleic Acid (CLA) molecules due to their significant contribution to the cheese's nutritional profile. The aim of this section is two-fold: to investigate the variation in saturated and unsaturated fatty acid content across three distinct seasons (winter, spring, and summer), and to develop a classification model capable of discriminating among seasonal groups, even under conditions of incomplete chemical composition data (i.e., with missing values in the dataset). By addressing the problem of missing values, the presented algorithm managed to efficiently infer the missing information and to predict the class membership despite part of information being hidden.

Chapter 7 focuses on the joint effect of milk thermal treatment and ripening time on the metabolic profile of Fiore Sardo PDO cheese. The problem of biomarker identification in presence of multiple testing problem (high type I error probability) is addressed. Once the molecules of interest are identified, the application of classification algorithms was employed as a screening tool for the compliance with PDO-label guidelines in terms of thermal treatment and ripening time.

In Chapter 8 the coagulation problem is investigated under a different point of view. Specifically, the effect of plant-based and animal-based rennet enzymes on the cheese lipidome is taken into consideration through Liquid chromatography-quadrupole time-of-flight mass spectrometry (LC-QTOFMS) measurements. The interaction between rennet type and ripening time is also considered for both biomarker identification and sample classification. In this investigation, data matrix was affected by the so called "curse of dimensionality" (presence of numerous uninformative variables), which undermines the performances of any classification algorithm. Therefore, a variable selection strategy was developed to improve the accuracies of a multivariate statistical model for discerning both rennet type and cheese ripening time.

Chapter 9 comprises the conclusion of the dissertation.

As it can be seen, each topic is addressed with a dual perspective: on one hand, the chemical characterization of the system under investigation, and on the other, the development or

application of statistical strategies to manage issues posed by data complexity. This approach reflects the difficulties arising from real-world scenarios where chemical characterization and statistical challenges emerge simultaneously and are intertwined. For this reason, the topics addressed in this thesis are approached in a unified manner that intends to connect the analytical understanding with data interpretation.

1.5 Contributions in international journals

Here follow the journal publications included in the research conducted in this dissertation:

Sibono L, Grosso M, Tronci S, Errico M, Addis M, Vacca M, Manis C, Caboni P (2023) Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction. Metabolites 13:. https://doi.org/10.3390/metabo13070877

Sibono L, Manis C, Zucca F, Atzori L, Errico M, Tronci S, Casula M, Dedola A, Pes M, Caboni P, Grosso M (2024) Metabolomic profiling of Fiore Sardo cheese: Investigation of the influence of thermal treatment and ripening time using univariate and multivariate classification techniques. Food Chem 456:. https://doi.org/10.1016/j.foodchem.2024.139930

Sibono L, Tronci S, Hedegaard MAB, Errico M, Grosso M (2025) Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy. Appl Food Res 5:. https://doi.org/10.1016/j.afres.2025.100701

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. (2025). PAT-driven dairy processing: a rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics. Analytical and Bioanalytical Chemistry. https://doi.org/10.1007/s00216-025-05965-2

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. (2025). Raman spectroscopy as screening tool for detecting anomalies during rennet coagulation of milk. Processes. https://doi.org/10.3390/pr13113519

Sibono, L., Grosso, M., Errico, M., Tronci, S., Lai, G., Pes, M., Addis, M., Caboni, P., & Manis, C. (2025). Untargeted Lipidomics by LC-QTOF-MS for Biomarker Selection and Discrimination of plant- and animal-derived Rennet in Cooked-Curd ovine cheese at Different Ripening Stages. Food Research International. https://doi.org/10.1016/j.foodres.2025.117987

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. Raman spectroscopy as a non-invasive tool for monitoring combined acid-rennet coagulation of milk. The manuscript is under preparation for submission.

The following publications refer to research topics explored during the PhD activity that fall outside the scope of the present thesis. Their development was supported by the competencies and methodologies acquired during the PhD work.

Sibono L, Tronci S, Hajrizaj R, Christensen K V., Errico M, Grosso M (2023) Optimization and kinetic analysis of untreated brewers' spent grain saccharification process via enzymatic hydrolysis. Biochem Eng J 198:. https://doi.org/10.1016/j.bej.2023.109044

Sibono L, Tronci S, Grosso M, Hajrizaj R, Errico M (2023) Brewer's Spent Grain to Bioethanol through a Hybrid Saccharification and Fermentation Process. Chem Eng Trans 99:7–12. https://doi.org/10.3303/CET2399002

Lisci S, Tronci S, Grosso M, Hajrizaj R, Sibono L, Karring H, Gerganov A, Maschietti M, Errico M (2024) Valorizing brewer's spent grain: A sequential pathway of supercritical extraction, hydrolysis, and fermentation. Chem Eng Sci 285:. https://doi.org/10.1016/j.ces.2023.119620

Sibono L, Grosso M, Tejedor-Calvo E, Casula M, Marco-Montori P, Garcia-Barreda S, Manis C, Caboni P (2025) A critical analysis of Adaptive Box-Cox transformation for skewed distributed data management: Metabolomics of Spanish and Argentinian truffles as a case study. Anal Chim Acta 1345:. https://doi.org/10.1016/j.aca.2025.343704

Coradduzza D, Sibono L, Tedde A, Marra S, De Miglio MR, Zinellu A, Medici S, Mangoni AA, Grosso M, Madonia M, Carru C (2025) Diagnostic Stratification of Prostate Cancer Through Blood-Based Biochemical and Inflammatory Markers. Diagnostics 15:. https://doi.org/10.3390/diagnostics15111385

Coradduzza, D., Cruciani, S., Sibono, L., Tedde, A., Zinellu, A., Maioli, M., Cogoni, A. A., De Miglio, M. R., Medici, S., Madonia, M., Angius, A., Grosso, M., & Carru, C. (2025). Diagnostic relevance of Humanin, GAS5 and miR-21/miR-103 in prostate disease risk stratification. *Clinical and Experimental Medicine*, *25*(1). https://doi.org/10.1007/s10238-025-01810-z

Finally, the participations at webinars, national and international conference are listed below:

Brewer's Spent Grain to Bioethanol through a Hybrid Saccharification and Fermentation Process. ICHEAP 2023, International conference, Naples (Italy), May 21-24, 2023.

Multivariate statistics approach to infer metabolite profile variations in Pecorino Romano Cheese . XI CCM, International conference, Padua (Italy), June 27-30, 2023.

Process Analytical Technology in Food: Rheo-Raman Monitoring of Milk Coagulation as a Case Study . GRICU 2025, National conference, Ischia (Italy), September 14-17, 2025.

Webinar "Dairy Products: From Starter to Structure" held by "Process journal", Editor: MDPI. September 10, 2025.

1.6 References

Agregán, R., Echegaray, N., Nawaz, A., Hano, C., Gohari, G., Pateiro, M., & Lorenzo, J. M. (2021). Foodomic-based approach for the control and quality improvement of dairy products. Metabolites, 11(12). https://doi.org/10.3390/metabo11120818

Alinaghi, M., Nilsson, D., Singh, N., Höjer, A., Saedén, K. H., & Trygg, J. (2023). Near-infrared hyperspectral image analysis for monitoring the cheese-ripening process. Journal of Dairy Science, 106(11), 7407–7418. https://doi.org/10.3168/jds.2023-23377

Alseekh, S., & Fernie, A. R. (2018). Metabolomics 20 years on: what have we learned and what hurdles remain? Plant Journal, 94(6), 933–942. https://doi.org/10.1111/tpj.13950

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., & Rasool, M. H. (2017). Proteomics: Technologies and their applications. Journal of Chromatographic Science, 55(2), 182–196. https://doi.org/10.1093/chromsci/bmw167

Barkema, H. W., von Keyserlingk, M. A. G., Kastelic, J. P., Lam, T. J. G. M., Luby, C., Roy, J. P., LeBlanc, S. J., Keefe, G. P., & Kelton, D. F. (2015). Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. Journal of Dairy Science, 98(11), 7426–7445. https://doi.org/10.3168/jds.2015-9377

Buescher, J. M., & Driggers, E. M. (2016). Integration of omics: more than the sum of its parts. Cancer & Metabolism, 4(1). https://doi.org/10.1186/s40170-016-0143-y

Caboni, P., Maxia, D., Scano, P., Addis, M., Dedola, A., Pes, M., Murgia, A., Casula, M., Profumo, A., & Pirisi, A. (2019). A gas chromatography-mass spectrometry untargeted metabolomics approach to discriminate Fiore Sardo cheese produced from raw or thermized ovine milk. Journal of Dairy Science, 102(6), 5005–5018. https://doi.org/10.3168/jds.2018-15885

Caboni, P., Scano, P., Sanchez, S., Garcia-Barreda, S., Corrias, F., & Marco, P. (2020). Multi-platform metabolomic approach to discriminate ripening markers of black truffles (Tuber melanosporum). Food Chemistry, 319. https://doi.org/10.1016/j.foodchem.2020.126573

Capozzi, F., & Bordoni, A. (2013). Foodomics: A new comprehensive approach to food and nutrition. Genes and Nutrition, 8(1), 1–4. https://doi.org/10.1007/s12263-012-0310-x

Cardello, A. V., Llobell, F., Giacalone, D., Roigard, C. M., & Jaeger, S. R. (2022). Plant-based alternatives vs dairy milk: Consumer segments and their sensory, emotional, cognitive and situational use responses to tasted products. Food Quality and Preference, 100. https://doi.org/10.1016/j.foodqual.2022.104599

Cassoli, L. D., Sartori, B., Zampar, A., & Machado, P. F. (2011). An assessment of Fourier Transform Infrared spectroscopy to identify adulterated raw milk in Brazil. International Journal of Dairy Technology, 64(4), 480–485. https://doi.org/10.1111/j.1471-0307.2011.00711.x

Castillo, M. (2006). Cutting time prediction methods in cheese making. Encyclopedia of Agricultural, Food, and Biological Engineering, 1, 1–7.

Chowdhury, M. A. H., Reem, C. S. A., Rahman, S. M., & Sarkar, F. (2025). Comprehensive approaches for ensuring microbial safety in the dairy industry: Monitoring systems, inhibitory strategies and future prospects. Food Control, 168. https://doi.org/10.1016/j.foodcont.2024.110894

Del Carratore, F., Lussu, M., Kowalik, M. A., Perra, A., Griffin, J. L., Atzori, L., & Grosso, M. (2016). Statistical Health Monitoring Applied to a Metabolomic Study of Experimental Hepatocarcinogenesis: An Alternative Approach to Supervised Methods for the Identification of False Positives. Analytical Chemistry, 88(16), 7921–7929. https://doi.org/10.1021/acs.analchem.5b03078

EUROSTAT. (2025). Milk collection (all milks) and dairy products obtained - annual data. EUROSTAT. https://doi.org/DOI : 10.2908/apro_mk_pobta

Fabiano, A., Gazzolo, D., Zimmermann, L. J. I., Gavilanes, A. W. D., Paolillo, P., Fanos, V., Caboni, P., Barberini, L., Noto, A., & Atzori, L. (2011). Metabolomic analysis of bronchoalveolar lavage fluid in preterm infants complicated by respiratory distress syndrome: Preliminary results. Journal of Maternal-Fetal and Neonatal Medicine, 24(SUPPL. 2), 56–59. https://doi.org/10.3109/14767058.2011.606977

Fagan, C. C., Everard, C., O'Donnell, C. P., Downey, G., & O'Callaghan, D. J. (2005). Prediction of inorganic salt and moisture content of process cheese using dielectric spectroscopy. International Journal of Food Properties, 8(3), 543–557. https://doi.org/10.1080/10942910500269816

FDA. (2004). Guidance for industry, PAT-A framework for innovative pharmaceutical development, manufacturing and quality assurance.

Felfoul, I., Bouazizi, A., Attia, H., & Karoui, R. (2022). Monitoring of acid-induced coagulation of dromedary and cow's milk using different complementary analytical techniques. Food Control, 136. https://doi.org/10.1016/j.foodcont.2022.108867

García-Cañas, V., Simó, C., León, C., & Cifuentes, A. (2010). Advances in Nutrigenomics research: Novel and future analytical approaches to investigate the biological activity of natural compounds and food functions. Journal of Pharmaceutical and Biomedical Analysis, 51(2), 290–304. https://doi.org/10.1016/j.jpba.2009.04.019

Gehlot, A., Malik, P. K., Singh, R., Akram, S. V., & Alsuwian, T. (2022). Dairy 4.0: Intelligent Communication Ecosystem for the Cattle Animal Welfare with Blockchain and IoT Enabled Technologies. Applied Sciences (Switzerland), 12(14). https://doi.org/10.3390/app12147316

Gorrochategui, E., Jaumot, J., Lacorte, S., & Tauler, R. (2016). Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. TrAC - Trends in Analytical Chemistry, 82, 425–442. https://doi.org/10.1016/j.trac.2016.07.004

Grassi, S., Strani, L., Alamprese, C., Pricca, N., Casiraghi, E., & Cabassi, G. (2022). A FT-NIR process analytical technology approach for milk renneting control. Foods, 11(1). https://doi.org/10.3390/foods11010033

Hassoun, A., Garcia-Garcia, G., Trollman, H., Jagtap, S., Parra-López, C., Cropotova, J., Bhat, Z., Centobelli, P., & Aït-Kaddour, A. (2023). Birth of dairy 4.0: Opportunities and challenges in adoption of fourth industrial revolution technologies in the production of milk and its derivatives. Current Research in Food Science, 7. https://doi.org/10.1016/j.crfs.2023.100535

Hayes, E., Greene, D., O'Donnell, C., O'Shea, N., & Fenelon, M. A. (2023). Spectroscopic technologies and data fusion: Applications for the dairy industry. Frontiers in Nutrition, 9. https://doi.org/10.3389/fnut.2022.1074688

Ibáñez, C., Simó, C., García-Cañas, V., Cifuentes, A., & Castro-Puyana, M. (2013). Metabolomics, peptidomics and proteomics applications of capillary electrophoresis-mass spectrometry in Foodomics: A review. Analytica Chimica Acta, 802, 1–13. https://doi.org/10.1016/j.aca.2013.07.042

Koeck, A., Loker, S., Miglior, F., Kelton, D. F., Jamrozik, J., & Schenkel, F. S. (2014). Genetic relationships of clinical mastitis, cystic ovaries, and lameness with milk yield and somatic cell score in first-lactation Canadian Holsteins. Journal of Dairy Science, 97(9), 5806–5813. https://doi.org/10.3168/jds.2013-7785

Laurent, C., Hulin, S., Agabriel, C., Chassaing, C., Botreau, R., & Monteils, V. (2017). Co-construction of an assessment method of the environmental sustainability for cattle farms involved in a Protected Designation of Origin (PDO) cheese value chain, Cantal PDO. Ecological Indicators, 76, 357–365. https://doi.org/10.1016/j.ecolind.2017.01.035

Li, M., Cai, Q., Zhang, T., Tang, H., & Li, H. (2025). Progress of Complex System Process Analysis Based on Modern Spectroscopy Combined With Chemometrics. Journal of Chemometrics, 39(2). https://doi.org/10.1002/cem.70006

Loddo, A., Di Ruberto, C., Armano, G., & Manconi, A. (2022). Automatic Monitoring Cheese Ripeness Using Computer Vision and Artificial Intelligence. IEEE Access, 10, 122612–122626. https://doi.org/10.1109/ACCESS.2022.3223710

Lyndgaard, C. B., Engelsen, S. B., & Van Den Berg, F. W. J. (2012). Real-time modeling of milk coagulation using in-line near infrared spectroscopy. Journal of Food Engineering, 108(2), 345–352. https://doi.org/10.1016/j.jfoodeng.2011.07.029

Maillot, C., Sion, C., De Isla, N., Toye, D., & Olmos, E. (2021). Quality by design to define critical process parameters for mesenchymal stem cell expansion. Biotechnology Advances, 50. https://doi.org/10.1016/j.biotechadv.2021.107765

Milani, F. X., Nutter, D., & Thoma, G. (2011). Invited review: Environmental impacts of dairy processing and products: A review. Journal of Dairy Science, 94(9), 4243–4254. https://doi.org/10.3168/jds.2010-3955

Minnich, C., Hardy, S., & Krämer, S. (2016). Stopping the Babylonian Confusion: An Updated Nomenclature for Process Analyzers in PAT Applications. Chemie-Ingenieur-Technik, 88(6), 694–697. https://doi.org/10.1002/cite.201500188

Nash, W. J., & Dunn, W. B. (2019). From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography–mass spectrometry data. TrAC Trends in Analytical Chemistry, 120, 115324. https://doi.org/10.1016/j.trac.2018.11.022

Nicolau, N., Buffa, M., O'Callaghan, D. J., Guamis, B., & Castillo, M. (2015). Estimation of clotting and cutting times in sheep cheese manufacture using NIR light backscatter. Dairy Science and Technology, 95(4), 495–507. https://doi.org/10.1007/s13594-015-0232-7

Panikuttira, B., O'Shea, N., Tobin, J. T., Tiwari, B. K., & O'Donnell, C. P. (2018). Process analytical technology for cheese manufacture. International Journal of Food Science and Technology, 53(8), 1803–1815. https://doi.org/10.1111/ijfs.13806

Panikuttira, B., Payne, F. A., O'Shea, N., Tobin, J. T., & O'Donnell, C. P. (2019). Evaluation of a fluorescence and infrared backscatter sensor to monitor acid induced coagulation of skim milk. Innovative Food Science and Emerging Technologies, 54, 219–224. https://doi.org/10.1016/j.ifset.2019.04.011

Pomerantsev, A. L., & Rodionova, O. Y. (2012). Process analytical technology: A critical view of the chemometricians. Journal of Chemometrics, 26(6), 299–310. https://doi.org/10.1002/cem.2445

Puerta, F., Wessel, W., García, V., López, E., Rovira, S., Alvarez, D., Roa, I., & López, M. B. (2017). A preliminary study of the use of a Raman laser sensor to monitor coagulation and syneresis for the online control of cheesemaking. International Journal of Dairy Technology, 70(4), 607–609. https://doi.org/10.1111/1471-0307.12439

Rezitis, A. N., & Kastner, G. (2021). On the joint volatility dynamics in international dairy commodity markets*. Australian Journal of Agricultural and Resource Economics, 65(3), 704–728. https://doi.org/10.1111/1467-8489.12433

Roohi, Zaheer, M. R., & Gupta, A. (2018). Current development and future perspectives of microbial enzymes in the dairy industry. Enzymes in Food Biotechnology: Production, Applications, and Future Prospects, 287–302. https://doi.org/10.1016/B978-0-12-813280-7.00017-7

Sangode, P. B. (2025). Dairy 4.0: Embracing Digital Technology for Sustainable Farming:AReview. Asian Journal of Dairy and Food Research, 44(2), 169–178. https://doi.org/10.18805/ajdfr.DR-2242

Scano, P., Cusano, E., Caboni, P., & Consonni, R. (2019). NMR metabolite profiles of dairy: A review. International Dairy Journal, 90, 56–67. https://doi.org/10.1016/j.idairyj.2018.11.004

Schofield, T., Robbins, D., & Miró-Quesada, G. (2015). Critical quality attributes, specifications, and control strategy. AAPS Advances in the Pharmaceutical Sciences Series, 18, 511–535. https://doi.org/10.1007/978-1-4939-2316-8_21

Sibono, L., Grosso, M., Tejedor-Calvo, E., Casula, M., Marco-Montori, P., Garcia-Barreda, S., Manis, C., & Caboni, P. (2025). A critical analysis of Adaptive Box-Cox transformation for skewed distributed data management: Metabolomics of Spanish and Argentinian truffles as a case study. Analytica Chimica Acta, 1345. https://doi.org/10.1016/j.aca.2025.343704

Sibono, L., Manis, C., Zucca, F., Atzori, L., Errico, M., Tronci, S., Casula, M., Dedola, A., Pes, M., Caboni, P., & Grosso, M. (2024). Metabolomic profiling of Fiore Sardo cheese: Investigation of the influence of thermal treatment and ripening time using univariate and multivariate classification techniques. Food Chemistry, 456. https://doi.org/10.1016/j.foodchem.2024.139930

Smaoui, S., Tarapoulouzi, M., Agriopoulou, S., D'Amore, T., & Varzakas, T. (2023). Current State of Milk, Dairy Products, Meat and Meat Products, Eggs, Fish and Fishery Products Authentication and Chemometrics. Foods, 12(23). https://doi.org/10.3390/foods12234254

Soyeurt, H., Bruwier, D., Romnee, J. M., Gengler, N., Bertozzi, C., Veselko, D., & Dardenne, P. (2009). Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. Journal of Dairy Science, 92(6), 2444–2454. https://doi.org/10.3168/jds.2008-1734

Subedi, P., Moertl, S., & Azimzadeh, O. (2022). Omics in Radiation Biology: Surprised but Not Disappointed. Radiation, 2(1), 124–129. https://doi.org/10.3390/radiation2010009

Valdés, A., Álvarez-Rivera, G., Socas-Rodríguez, B., Herrero, M., Ibáñez, E., & Cifuentes, A. (2022). Foodomics: Analytical Opportunities and Challenges. Analytical Chemistry, 94(1), 366–381. https://doi.org/10.1021/acs.analchem.1c04678

Visioli, F., & Strata, A. (2014). Milk, dairy products, and their functional effects in humans: A narrative review of recent evidence. Advances in Nutrition, 5(2), 131–143. https://doi.org/10.3945/an.113.005025

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. Pattern Recognition, 8(3), 127–139. https://doi.org/10.1016/0031-3203(76)90014-5

Wulandari, L., Idroes, R., Noviandy, T. R., & Indrayanto, G. (2022). Application of chemometrics using direct spectroscopic methods as a QC tool in pharmaceutical industry and their validation. Profiles of Drug Substances, Excipients and Related Methodology, 47, 327–379. https://doi.org/10.1016/bs.podrm.2021.10.006

Chapter 2: Mathematical methods

This chapter presents the mathematical methodologies employed in this thesis and the algorithms developed to address the problems under consideration. This discussion focuses on the theoretical foundations of these approaches. In particular, multivariate analysis models will be briefly presented along with their possible variants. Univariate statistical techniques and state estimation models are also discussed. The softwares used to implement the algorithms described herein are reported in the following chapters.

2.1 Univariate models

2.1.1 ANOVA test

The ANalysis Of VAriance (ANOVA) is a statistical method used to examine the influence of one or more independent variables, also known as factors, on a dependent variable. This technique allows to determine whether there are significant differences between the means of a dependent variable obtained from different treatments of such factors, as well as to explore possible interaction effects. In the present thesis, analyses were conducted considering two factors. Consequently, for practicality, the following discussion focuses on the two-way ANOVA model. Considering two factors denoted by letter A and B with a number of levels (or treatments) equal to a and b, the effect model can be expressed as follows (Montgomery, 2013):

$$x_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk} \quad (2.1)$$

Where x_{ijk} represents the *k*-th experimental replication related to *i* Level of factor A and *j* level of factor B, while ε_{ijk} is the residual for the *ijk* observation. In a two-way ANOVA test, the null hypothesis (H0) being tested concerns the equality of treatment effects with respect to the following factors:

factor A:

$$\begin{cases} H0: \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ H1: \text{at least one } \tau_i \neq 0 \end{cases}$$

factor B:

$$\begin{cases} H0: \gamma_1 = \gamma_2 = \cdots = \gamma_b = 0 \\ H1: \text{at least one } \gamma_j \neq 0 \end{cases}$$

Interaction between ripening time and milk thermal treatment:

$$\begin{cases} H0: (\tau\gamma)_{ij} = 0 \ \forall \ i \neq j \\ H1: \text{at least one } (\tau\gamma)_{ij} \neq 0 \end{cases}$$

The ANOVA table can be constructed to test the significance of single effects and interaction by decomposing the total sum of squares (SS_T) into different components. The total sum of squares (SS_T) is defined as:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E \quad (2.2)$$

Where SS_E is the sum of squares errors, while SS_A, SS_B and SS_{AB} are the sums of squares of treatments for factor A, factor B and the interaction term respectively.

The sum of squares for each effect can be calculated as follows:

$$SS_A = bn \sum_{i=1}^{a} (\bar{x}_{i..} - \bar{x}_{...})^2 \quad (2.3)$$

$$SS_B = an \sum_{j=1}^{b} (\bar{x}_{.j.} - \bar{x}_{...})^2 \quad (2.4)$$

$$SS_{AB} = n \sum_{i=1}^{a} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} - \bar{x}_{...})^2 \quad (2.5)$$

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij.})^2 \quad (2.6)$$

Where:

$$\bar{x}_{i..} = \frac{1}{b\,n} \sum_{j=1}^{b} \sum_{k=1}^{n} x_{ijk} \quad (2.7)$$

$$\bar{x}_{.j.} = \frac{1}{a\,n} \sum_{i=1}^{a} \sum_{k=1}^{n} x_{ijk} \quad (2.8)$$

$$\bar{x}_{ij.} = \frac{1}{n} \sum_{k=1}^{n} x_{ijk} \quad (2.9)$$

$$\bar{x}_{...} = \frac{1}{ab\,n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} x_{ijk} \quad (2.10)$$

Finally, the mean sum of squares (MS) for each effect can be calculated as follows:

$$MSA = \frac{SS_A}{a-1} \quad (2.11)$$

$$MSB = \frac{SS_B}{b-1} \quad (2.12)$$

$$MSAB = \frac{SS_{AB}}{ab-a-b+1} \quad (2.13)$$

$$MSE = \frac{SS_E}{a\,b\,(n-1)} \quad (2.14)$$

The F-statistic for each effect can be calculated as the ratio of the mean sum of squares to the mean sum of squares of the residual error:

$$f_A = \frac{\mathrm{MSA}}{\mathrm{MSE}} \quad (2.15)$$

$$f_B = \frac{\mathrm{MSB}}{\mathrm{MSE}} \quad (2.16)$$

$$f_{AB} = \frac{\text{MSAB}}{\text{MSE}} \quad (2.17)$$

The null hypothesis for each effect is that the corresponding effect is zero. If the F-statistic is larger than the critical value of the F-distribution at a given significance level α, the null hypothesis is rejected, and the effect is considered to be significant. Vice versa, if the observed value of Fisher value is smaller than the critical value given by the test's significance level the test is concluded with the non-rejection of null hypothesis.

2.1.2 False discovery rate correction

The level of significance α represents the probability of incorrectly rejecting the null hypothesis. When the significance of factors is tested on multiple variables, the overall probability $\tilde{\alpha}$ of incorrectly rejecting at least one hypothesis can be calculated through the complement of the probability of independent events:

$$\tilde{\alpha} = 1 - (1 - \alpha)^{\mathrm{r}} \quad (2.18)$$

In order to identify the variables that are mostly responsible of variations in the factors of interest, while avoiding the uncontrolled cumulation of $\tilde{\alpha}$, Benjamini-Hochberg method (Benjamini & Hochberg, 1995) can be employed. This method is based on the idea of controlling the proportion of false discoveries among all rejected hypotheses (i.e. variables significance). Given $p_1, p_2 \ldots p_m$ *m* p-values obtained from multiple hypothesis tests and sorted in ascending order, where *m* is the total number of tests, the False Discovery Rate (FDR) can be defined as the expected value of the proportion of false discoveries out of the total number of rejected hypotheses:

$$\text{FDR}=\text{E}(\text{FP}/\text{R}) \quad (2.19)$$

Where FP is the number of false positive and R is the number of rejections (True positives + False positives). The goal of Benjamini-Hochberg procedure is to find the *l* quantity such that:

$$l = \max\{g : p_{(g)} \leq \frac{g}{m} q^*\} \quad (2.20)$$

Where g is the g-th p-value subjected to FDR correction and q^* is the desired FDR (Benjamini, Y. & Hochberg, 1995). q^* was set equal to 0.05. Once the l value is computed, the null hypotheses related to the first $p_1, p_2 \ldots p_l$ p-values are rejected. Conversely, the remaining null hypotheses are not rejected, even if the relative p-value is less than 0.05 (Benjamini & Hochberg, 1995).

Hence, the FDR correction helps to balance the trade-off between type I error (the rejection of a valid null hypothesis) and type II error (the acceptance of an invalid null hypothesis) in multiple hypothesis testing. This method enables one to undertake a secondary screening process (Downwards ANOVA test) aimed at identifying the most significant variables.

2.1.3 t-test based binary classification model

In the present thesis, hypothesis testing has also been employed as a classification tool. A t-test based classifier merges the hypotheses test on significant difference between two-sample means with the Monte Carlo cross-validation rationale. In such cross-validation method, a portion of the samples from group A are randomly selected to form the first training set. Afterwards, a group of samples equal in size to the first training set, is taken from the group B to create the second training set. The remaining samples are left to make the test set for group A and B, respectively. The purpose of the t-test procedure is to determine whether each individual observation in the two test sets is significantly different from the average of the training sets. Let x_j^{test1} and x_j^{test2} be the experimental observation belonging to the test sets selected from group A and B respectively, and $\bar{x}^{trainA}$ and $\bar{x}^{trainB}$ the first and second training set means, respectively. In this case, *j* is the *j*-th observation for each test set $j = 1,2 \dots n_{test}$. It is important to specify that for a balanced testing $n_{testA} = n_{testB} = n_{test}$ and $n_{trainA} = n_{trainB} = n_{train}$. The hypotheses tests are performed under the assumptions of independent normally distributed data samples, characterized by equal variances. The equal-variance assumption was supported by employing a pooled variance estimate, which provides a more stable and reliable estimation of data variance due to the accounting of all samples in the training sets.

S_P is the pooled variance and is defined as follows (Liao & Akritas, 2007):

$$S_P = \frac{(n_{trainA}-1)S_A+(n_{trainB}-1)S_B}{n_{trainA}+n_{trainB}-2} = \frac{S_A+S_B}{2} \tag{2.21}$$

The idea of this model is that if a sample is taken from the test set of A group, which indicates that it truly belongs to group A, and is then incorporated into the mean of group B through a t-test, the difference between groups A and B is supposed to decrease and vice versa. The following hypotheses testing procedure is therefore applied:

First t-test: x_j^{testA} is placed with A group training set, a new mean is calculated and tested versus $\bar{x}^{trainB}$:

$$t_{1,j}^{testA} = \frac{\frac{(n_{train}\bar{x}^{trainA}+x_j^{testA})}{n_{train}+1} - \bar{x}^{trainB}}{\sqrt{S_P\left(\frac{1}{n_{train}+1}+\frac{1}{n_{train}}\right)}} \quad (2.22)$$

Second t-test: x_j^{testA} is placed with B group training set, a new mean is calculated and tested versus $\bar{x}^{trainA}$

$$t_{2,j}^{testA} = \frac{\bar{x}^{trainA} - \frac{(n_{train}\bar{x}^{trainB}+x_j^{testA})}{n_{train}+1}}{\sqrt{S_P\left(\frac{1}{n_{train}+1}+\frac{1}{n_{train}}\right)}} \quad (2.23)$$

The p-values for test samples belonging to A and B are calculated:

$$\begin{cases} PV_1(x_j^{testA}) = 1 - p\left(T_1 < t_{1,j}^{testA}\right) \\ PV_2(x_j^{testA}) = 1 - p\left(T_1 < t_{2,j}^{testA}\right) \end{cases} \quad (2.24)$$

Here, T_1 is the student's T distribution with one degree of freedom. The same procedure is applied for samples belonging to the second test set (i.e. x_j^{testB}). t-tests are performed assuming that the prior probabilities of a sample being a member of each class are equal (Liao & Akritas, 2007). This assumption is founded on the notion that the prior probability of belonging to a particular class is equal to the proportion of samples in the original population (Brereton, 2009). Hence, it is assumed that the two prior probabilities are equal because the sizes of the two training and test sets are equal.

Finally, assuming that the membership of a sample to group A denotes that the test is concluded as negative while its membership to group B indicates that the test concluded as positive the following statistical conclusions can be made:

- If $PV_1(x_j^{testA}) > PV_2(x_j^{testA})$ the sample that belongs to A group is wrongly classified to B class and the univariate classifier has detected a false positive.
- If $PV_1\left(x_j^{testA}\right) < PV_2(x_j^{testA})$, a true negative is detected
- If $PV_1\left(x_j^{testB}\right) > PV_2(x_j^{testB})$ the sample that belongs to B group is correctly classified and the univariate classifier has detected a true positive.

- If $PV_1\left(x_{ij}^{testB}\right) < PV_2(x_{ij}^{testB})$ a false negative is detected

The results for each tested sample are monitored for false positive (FP), true positive (TP) and correctly classified (CC) counts. In such a way, FPR and accuracy can be calculated:

$$FPR = \frac{FP}{FP+TN} \tag{2.25}$$

$$Accuracy = \frac{TP+TN}{n_{test1}+n_{test2}} = \frac{CC}{n_{test1}+n_{test2}} \tag{2.26}$$

In order to obtain a robust estimation of the goodness of classification, the dataset is partitioned into training and test set for several iterations, and the calculated performance indices are averaged.

2.2 Mutivariate models

2.2.1 Data preprocessing

Prior to performing multivariate analyses, data were generally preprocessed through autoscaling (Brereton, 2009):

$$x_{ij}^{*} = \frac{x_{ij}^{raw} - \bar{x}_{j}^{raw}}{s_j} \tag{2.27}$$

Here, x_{ij}^{raw} represent raw data correspondent to *i*-th sample and *j*-th variable, $\bar{x}_{\bullet j}^{raw}$ represent the average value along the *j*-th variable and s_j denotes the standard deviation of sample measurement relative to the *j*-th variable, which was calculated from the square root of variance:

$$s_j^2 = \frac{1}{n-1}\sum\left(x_{ij}^{raw} - \bar{x}_{.j}^{raw}\right)^2 \tag{2.28}$$

2.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate technique that is used to reduce the complexity of a high-dimensional problems and identify the underlying trends in the data. The PCA model is based on the concept of transforming the original data matrix into a new set of variables, referred to as principal components, that capture the most significant variations in the data. This

transformation is achieved through the computation of two matrices, the score matrix $\underline{\underline{T}}$ $(I \times L)$ and the loading matrix $\underline{\underline{P}}$ $(L \times J)$, where I represents the observations number, L is the number of principal components employed in the model and J are the total variables in the original dataset (Brereton, 2009). The score matrix represents the projected data in the new coordinate system, while the loading matrix is an orthogonal matrix that represents the weights assigned to each original variable in the construction of the principal components. PCA model foresees the decomposition of data matrix into two matrices, as stated in the following equation (Bro & Smilde, 2014):

$$\underline{\underline{\hat{X}}} = \underline{\underline{T}}\,\underline{\underline{P}}^T \tag{2.29}$$

Which is the approximation of the original data matrix, $\underline{\underline{X}}$, defined as follows:

$$\underline{\underline{X}} = \underline{\underline{T}}\,\underline{\underline{P}}^T + \underline{\underline{E}} \tag{2.30}$$

Here, $\underline{\underline{\hat{X}}}$ is the data projection to the principal subspace, and $\underline{\underline{E}}$ is the residual matrix which provides information about the unexplained variance in the original data and the quality of the PCA model. Once the scores are obtained from the PCA model, they can be graphically represented, thus allowing to identify, through simple visual inspection, any differences between the analyzed samples, such as the presence of outliers. An important aspect relevant to PCA is the assessment of Principal Components (PCs) eigenvalues. Given a generical principal component C, the eigenvalue λ_C is defined according to the formula (Brereton, 2009):

$$\lambda_C = \sum_{i=1}^{I} t_{iC}^2 \tag{2.31}$$

Eigenvalues of the PCs are ordered in decreasing order starting from the first component. This means that the first PC has the largest eigenvalue, which represents the amount of variation captured by that component. The eigenvalue of the second PC is lower than the first and so on, for each subsequent PC. Thus, PCs with the largest eigenvalues contribute the most to the overall explanation of the data, while those with smaller eigenvalues contribute less. More in detail, given $\underline{\underline{S}}$ the covariance matrix $\underline{\underline{S}}$, eigenvalues are related to $\underline{\underline{S}}$ through the following covariance elaboration:

$$\underline{\underline{S}} = \underline{\underline{X}}^T \underline{\underline{X}} \tag{2.32}$$

By substituting eqs. (2.29) and (2.30) to (2.32) (neglecting the error contribute), and applying to inverse matrix of loadings the property of orthogonal matrices (i.e. $\underline{\underline{P}}^{-1} = \underline{\underline{P}}^{T}$) it can be derived:

$$\underline{\underline{S}} = \underline{\underline{P}}\,\underline{\underline{\Lambda}}\,\underline{\underline{P}}^{T} \tag{2.33}$$

Where $\underline{\underline{\Lambda}}$ is the diagonal eigenvalues matrix. Eigenvalues are often presented as percentages, relative to the sum of squares in the entire preprocessed data set (i.e. the total variance), referred to as the percentage variances, defined as in equation (Brereton, 2009):

$$V_C = \frac{\lambda_C}{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2} \tag{2.34}$$

The application of eq. 2.34 allows to determine the percentage of variance explained by each principal component retained in the model.

2.2.3 Partial Least Square

Partial Least Squares (PLS) is a multivariate statistical technique designed to model the relationships between two sets of variables: predictors (independent variables) and responses (dependent variables). Unlike traditional regression methods, PLS is particularly employed when the predictor variables are highly collinear and when the sample size is relatively small compared to the number of predictors. When there is only one dependent variable to be elaborated, PLS can be defined as the following system of two linear equations (Brereton, 2009):

$$\begin{cases} \underline{\underline{X}} = \underline{\underline{T}}\,\underline{\underline{P}}^{T} + \underline{\underline{E}} \\ \underline{y} = \underline{\underline{T}}\,\underline{q} + \underline{f} \end{cases} \tag{2.35}$$

In such a system, the dependent variable is represented by the vector $\underline{y}$ $(I \times 1)$, while $\underline{q}$ $(L \times 1)$ denotes the response loading vector and $\underline{f}$ $(L \times 1)$ are the residuals for the dependent variable. The PLS algorithm follows the NIPALS algorithm which describes the following procedure (Wold et al., 2001):

When the response variable is a vector (like the cases addressed in this dissertation), the algorithm is initialized by defining the response observation $\underline{u}$=$\underline{y}$ with the definition of l-th latent variable weight $\underline{w_l}$, defined as follows:

$$\underline{w_l} = \frac{\underline{\underline{X}}^T \underline{u}}{\underline{u}^T \underline{u}} \tag{2.36}$$

Weights vectors are normalized and used to build the score and loading vector related to l-th latent variable:

$$\underline{t_l} = \underline{\underline{X}}\, \underline{w_l} \tag{2.37a}$$

$$\underline{p_l} = \frac{\underline{\underline{X}}^T \underline{t_l}}{\underline{t_l}^T \underline{t_l}} \tag{2.37b}$$

The l-th response loading is then calculated:

$$q_l = \frac{\underline{y}^T \underline{t_l}}{\underline{t_l}^T \underline{t_l}} \tag{2.38}$$

And finally, $\underline{u}$ vector is updated:

$$\underline{u} = \frac{\underline{y}\, q}{q^T q} \tag{2.39}$$

The iterative procedure is repeated until $\underline{t}$ vector does not change within a tolerance value (10^{-8}). Therefore, PLS algorithm identifies the directions of maximum covariance between $\underline{\underline{X}}$ and $\underline{y}$. Finally, errors are computed, in order to define the quantity of variance that both independent and dependent variables are not able to explain for a given number of principal components:

$$\begin{cases} \underline{\underline{E}} = \underline{\underline{X}} - \sum_{l=1}^{L} \underline{t_l} \underline{p_l}^T = \underline{\underline{X}} - \underline{\underline{T}}\, \underline{\underline{P}}^T \\ \underline{f} = \underline{y} - \sum_{l=1}^{L} \underline{t_l} q_l = \underline{y} - \underline{\underline{T}}\, \underline{q} \end{cases} \tag{2.40}$$

Here, $\underline{\underline{E}}$ represents the outcome of $\underline{\underline{X}}$ matrix deflation. Deflation removes the variability already explained from the previous latent variables. The iterative procedure above-presented is repeated using the deflated matrix if further latent variables are introduced in the PLS model, which means that $\underline{\underline{E}}$ substitutes $\underline{\underline{X}}$ in 2.36 and 2.38. This means that after concatenating weight vectors $\underline{\underline{W}} = [\underline{w_1}, \underline{w_2}, \dots, \underline{w_L}]$, the score matrix can be derived:

$$\underline{t_l} = \underline{\underline{X}} \prod_{k=1}^{l} (\underline{\underline{I}} - \underline{w_k} \underline{p_k}^T) \underline{w_l} \tag{2.41a}$$

$$\underline{\underline{T}} = \underline{\underline{X}} \prod_{k=1}^{l} (\underline{\underline{I}} - \underline{w_k} \underline{p_k}^T) \underline{\underline{W}} = \underline{\underline{X}}\, \underline{\underline{\widetilde{W}}} \tag{2.41b}$$

Once a model is built, the relationship between $\underline{y}$ and $\underline{\underline{X}}$ can be explicated by means of a coefficient vector $\underline{b}$ which can be obtained substituting eq. 2.41 to the second equation of system 2.35, thus landing in the following expressions:

$$\underline{y} = \underline{\underline{X}}\,\underline{\underline{\widetilde{W}}}\,\underline{q} + \underline{f} = \underline{y} = \underline{\underline{X}}\,\underline{b} + \underline{f} \tag{2.42a}$$

$$\underline{\hat{y}} = \underline{\underline{X}}\,\underline{b} \tag{2.42b}$$

Where $\underline{b} = \underline{\underline{W}}\,\underline{q}$, and $\underline{\hat{y}}$ denotes the estimation of response variable by the PLS model.

2.2.4 Partial Least Square Discriminant Analysis

Partial Least Squares Discriminant Analysis (PLS-DA) is a supervised multivariate statistical method used for sample classification and feature selection. It is an extension of Partial Least Squares Regression, adapted specifically for categorical response variables. In this model, the dependent variable is defined as a $\underline{\underline{Y}}$ $(I \times G)$ matrix, where G is the number of classes under examination. Every column of $\underline{\underline{Y}}$ is a dummy vector where the attribution "one" means that the i-th sample belongs to the g-th group, while "zero" denotes that the sample does not belong to the g-th class. In PLS-DA, samples can be classified using two different approaches: the "maximum probability" approach or the "Bayes' theorem-based threshold value" approach. In the former procedure a probability value is calculated for each class, meaning that samples are classified by choosing the class with the highest probability. This approach ensures that samples are always classified into one of the two classes. In the latter strategy, a threshold value is defined for each class: if the estimated response is greater than the threshold value defined for a particular class, the corresponding sample is assigned to that class, otherwise it is not allocated to that group (Ballabio & Consonni, 2013). The threshold is selected at the point where the number of false positives and false negatives is minimized. When the Bayesian approach is used, it may happen that the estimated class value for a sample is greater or less than all the threshold values defined for the classes. If greater, the sample will be assigned to all classes and thus, would be a confused sample whilst if less, the sample will not be recognized as a member of any class and is referred to as "unassigned". Both the approaches assume that the estimated values follow a normal distribution (Ballabio & Consonni, 2013). When calibrating a PLS model, it is crucial to select the proper number of LVs in order to avoid data overfitting. Such a task is generally accomplished by means of

validation techniques (Ballabio & Consonni, 2013). In the present dissertation various cross-validation strategies are considered to select the number LVs that maximizes the performance indices (e.g, classification accuracy for PLS-DA). The validation approach is here tailored to each specific case study. Consequently, the details regarding the LVs selection are provided within the corresponding chapters.

2.2.5 Probabilistic Principal Component Analysis

Probabilistic Principal Component Analysis seeks to relate a p-dimensional single observation vector $\underline{x}$ to a q-dimensional LVs vector $\underline{t}$ where q < p, thus mapping the original variable into a new space which offers a more parsimonious explanation of the dependences between observations. In such a model, LVs are assumed to be i.i.d and to follow a normal distribution. The p-dimensional error $\underline{\varepsilon}$ term is also considered, and it is assumed to behave like a gaussian noise $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{\underline{I}})$ where $\underline{\underline{I}}$ denotes the identity matrix and σ^2 is the variance parameter. Error is isotropic due to the variances being assumed equal for all the p variables. Hence, the PPCA model is a PCA formulation which defines a linear relationship between the observation vector and LVs, where noise term is taken into account as well (Tipping & Bishop, 1999):

$$\underline{x} = \underline{\underline{W}}\,\underline{t} + \underline{\mu} + \underline{\varepsilon} \tag{2.43}$$

In Eq. 2.43, $\underline{\underline{W}}$ is the loading matrix, $\underline{\mu}$ is the mean value for each observed variable while $\underline{t}$ is a score vector which is assumed to follow a normal behavior $\underline{t} \sim N(\underline{0}, \underline{\underline{I}})$. In PPCA modification a linear combination of gaussian variables is employed to model the multivariate dataset, giving rise to a gaussian distribution for the observation vector. Specifically, the expected value of the observation vector $E(\underline{x})$ corresponds to the mean value $\underline{\mu}$ (the expected value of $\underline{t}$ and $\underline{\varepsilon}$ is $\underline{0}$), while the variance matrix can be derived out by evaluating the expectation of linear transformation of a random variable:

$$Var(\underline{x}) = E\left[(\underline{x} - E(\underline{x}))(\underline{x} - E(\underline{x}))^T\right] = E\left[\left(\underline{\underline{W}}\,\underline{t} + \underline{\varepsilon}\right)\left(\underline{\underline{W}}\,\underline{t} + \underline{\varepsilon}\right)^T\right] \tag{2.44}$$

By assuming that $\underline{t}$ and $\underline{\varepsilon}$ are independent, eq. 2.44 can be rewritten as:

$$Var(\underline{x}) = E\left[\left(\underline{\underline{W}}\,\underline{t}\right)\left(\underline{\underline{W}}\,\underline{t}\right)^T\right] + E\left[\underline{\underline{W}}\,\underline{t}\underline{\varepsilon}^T\right] + E\left[\underline{\varepsilon}\left(\underline{\underline{W}}\,\underline{t}\right)^T\right] + E\left[\underline{\varepsilon}\,\underline{\varepsilon}^T\right] =$$

$$= E\left[\underline{\underline{W}}\,\underline{t}\,\underline{t}^T\underline{\underline{W}}^T\right] + E\left[\underline{\varepsilon}\,\underline{\varepsilon}^T\right] = \underline{\underline{W}}\,\underline{\underline{\mathrm{I}}}\,\underline{\underline{W}}^T + \sigma^2\underline{\underline{\mathrm{I}}} = \underline{\underline{W}}\,\underline{\underline{W}}^T + \sigma^2\underline{\underline{\mathrm{I}}} \tag{2.45}$$

The conditional expectation of $\underline{x}$ given the occurrence of $\underline{t}$ is straightforwardly equal to $\underline{\underline{W}}\,\underline{t} + \underline{\mu}$. More attention must be taken for computing the conditional variance of $\underline{x}$ given the occurrence of $\underline{t}$. For this problem, the application of the conditional variance definition leads to the following formula:

$$Var(\underline{x}|\underline{t}) = E\left[\left(\underline{x} - E(\underline{x}|\underline{t})\right)\left(\underline{x} - E(\underline{x}|\underline{t})\right)^T|\underline{t}\right] \tag{2.46}$$

By substituting the conditional expectation the equation 2.46 becomes:

$$Var(\underline{x}|\underline{t}) = E\left[\left(\underline{\underline{W}}\,\underline{t} + \underline{\mu} + \underline{\varepsilon} - \underline{\underline{W}}\,\underline{t} + \underline{\mu}\right)\left(\underline{\underline{W}}\,\underline{t} + \underline{\mu} + \underline{\varepsilon} - \underline{\underline{W}}\,\underline{t} + \underline{\mu}\right)^T|\underline{t}\right] \tag{2.47}$$

After simple calculations it is possible to obtain:

$$Var(\underline{x}|\underline{t}) = E\left[\underline{\varepsilon}\,\underline{\varepsilon}^T\right] = \sigma^2\underline{\underline{\mathrm{I}}} \tag{2.48}$$

Therefore, the conditional probability distribution of the observation vector $\underline{x}$ given $\underline{t}$ is Gaussian, as indicated in eq. 2.49:

$$\underline{x}|\underline{t} \sim N(\underline{\underline{W}}\,\underline{t} + \underline{\mu}, \sigma^2\underline{\underline{I}}). \tag{2.49}$$

Nonetheless, it may become necessary to obtain model scores from experimental data (for instance, for classification purposes). This task can be achieved by applying Bayes' Theorem on $\underline{x}|\underline{t}$, $\underline{x}$ and $\underline{t}$ distributions, which yields to the following expression (Tipping & Bishop, 1999):

$$\underline{t}|\underline{x} \sim N(\underline{\underline{M}}^{-1}\underline{\underline{W}}^T\,(\underline{x} - \underline{\mu}), \sigma^2\underline{\underline{M}}^{-1}). \tag{2.50}$$

Where $\underline{\underline{M}}$ is a qxq matrix evaluated as follows:

$$\underline{\underline{M}} = \underline{\underline{W}}^T\underline{\underline{W}} + \sigma^2\underline{\underline{I}}. \tag{2.51}$$

The demonstration of eq. 2.50 is reported in appendix A. The inference of the optimal loading matrix is performed by means of the EM algorithm, which maximizes the log-likelihood function with respect to model parameters (i.e., σ^2 and $\underline{\underline{W}}$). In the E-step, the expected values of the $\underline{t}$ given the observed data $\underline{x}$ are estimated by substituting the value of model parameters ($\underline{\underline{W}}$ and σ^2) at the current iteration into eq. 2.50. In the M-step, the model parameters are re-estimated by

maximising the log likelihood function using the expected values of $\underline{t}|\underline{x}$ derived in the previous E-step. These two steps are repeated iteratively until the algorithm converges (Tipping & Bishop, 1999). Several convergence criteria are available in the literature (Nyamundanda et al., 2010). In this work, the convergence criterion requires that the EM algorithm be executed until the difference between two log-likelihood functions computed in two successive iterations is less than a tolerance value set to 0.001 or until the maximum number of iterations, set equal to 1000, is exceeded. It is important to specify that the maximum likelihood estimate of the loading matrix corresponds to the loading matrix given by conventional PCA, with the addition that the assessment of model uncertainty is available in such a probabilistic formulation.

Once the optimal loading values of the PPCA model have been identified, the next step is to project the data containing missing values onto the space of latent variables and then reconstruct the samples affected by partial information (missing values), given the scores obtained from the probabilistic model. Firstly, observation vector $\underline{x}$ and the loading matrix for which the likelihood function is maximized $\underline{\underline{W}}_{ML}$ are split in two contributions:

$$\underline{x} = \begin{pmatrix} \underline{x}_k \\ \underline{NaN} \end{pmatrix} \tag{2.52}$$

$$\underline{\underline{W}}_{ML} = \begin{pmatrix} \underline{\underline{W}}_{k,ML} \\ \underline{\underline{W}}_{m,ML} \end{pmatrix} \tag{2.53}$$

Here, $\underline{NaN}$ corresponds to a rx1 vector correspondent to missing elements of the observation vector, which are identified as non-numerical values, whereas $\underline{\underline{W}}_{k,ML}$ is a (p-r)xq matrix which denotes the loading matrix partition related to the known variables of the observation vector to project. The quantiy r is the number of variables containing a missing value for a given sample. $\underline{\underline{W}}_{m,ML}$ denotes the loading matrix block associated with the missing information. Therefore, a new $\underline{\underline{M}}_k$ matrix computed from $\underline{\underline{W}}_{k,ML}$:

$$\underline{\underline{M}}_k = \underline{\underline{W}}_{k,ML}^T \underline{\underline{W}}_{k,ML} + \sigma^2 \underline{\underline{I}} \tag{2.54}$$

$\underline{\underline{M}}$ matrix dimensionality is preserved when $\underline{\underline{W}}_k$ is used for $\underline{\underline{M}}_k$ calculation because it only depends on the number of selected LVs. The following step foresees the computation of the scores vecor $\underline{t}_k$ made up of the contribution of the sole loading related to known variables for each specific sample:

$$\underline{t_k} = \underline{\underline{M_k^{-1}}}\,\underline{\underline{W_{k,ML}^T}}\left(\underline{x_k} - \underline{\mu_k}\right) \tag{2.55}$$

Where $\underline{\mu_k}$is a (p-r)x1 vector of variables average values. The expression reported in eq. 2.55 is not the final form of score vector, as it only considers the contribution of non-missing variables on the projection onto the latent space. The reconstruction of missing data can be achieved by multiplying both hands of 2.55 by $\underline{\underline{M}}$ which considers all the p variables. In such a way, the of reconstruction is reduced to a ordinary least square problem, where $\underline{\underline{M}}\,\underline{t_k}$is the dependent variable, and $\underline{\hat{x}}$ is the vector to be estimated (Tipping & Bishop, 1999):

$$\underline{\hat{x}} = \underline{\underline{W_{ML}}}\left(\underline{\underline{W_{ML}^T}}\,\underline{\underline{W_{ML}}}\right)^{-1}\underline{\underline{M}}\,\underline{t_k} + \underline{\mu} \tag{2.56}$$

It is important to specify that $\underline{\underline{W_{ML}}}$is not orthogonal and thus not optimal (Tipping & Bishop, 1999). Since $\underline{\underline{W_{ML}}}$ is a real matrix, it can be expressed in terms of the following singular value decomposition:

$$\underline{\underline{W_{ML}}} = \underline{\underline{U}}\,\underline{\underline{L}}\,\underline{\underline{V}}^T \tag{2.57}$$

Where $\underline{\underline{U}}$ is a pxq matrix of orthonormal column vectors, $\underline{\underline{L}}$ is the qxq diagonal matrix of singular values, and $\underline{\underline{V}}$ is a qxq orthogonal matrix. $\underline{\underline{U}}$ matrix represents the orthogonalised form of loading matrix obtained from maximum likelihood estimation and corresponds to the loadings obtained by conventional PCA. Hence the final expression of score matrix obtained from the overall contribution of both known and reconstructed information is given as follows:

$$\underline{t} = \underline{\underline{U}}^T\left(\underline{\hat{x}} - \underline{\mu}\right) \tag{2.58}$$

The values of $\underline{t}$ score can be used for classification purposes.

2.3 State Estimation

In the field of Process Analytical Technology it is crucial to keep track of process state, which refers to the set of variables that fully describe the dynamic condition of a system at a given time. These variables capture the internal evolution of the process and are essential for the process control and monitoring. Part of the key process variables are sometimes inaccessible to online

measurement due to issues such as measurement delays, intrusive sampling, or sensor limitations, just to name a few. In such cases, a state estimator can be designed to provide a real-time estimation of the values for the unmeasurable process variables, inferred by relying on the measurements at hand. In this dissertation, only linear state estimators are considered. Therefore, in the absence of any control input, the system can be represented by the following linear model:

$$\begin{cases} \frac{d\underline{x}}{dt} = \underline{\underline{A}}\underline{x} \\ \underline{y} = \underline{\underline{C}}\underline{x} \end{cases} \tag{2.59}$$

In order to any conflicts of notation, in this subsection $\underline{x} \in \mathbb{R}^v$ represents the set of v state variables, $\underline{\underline{A}} \in \mathbb{R}^{v \times v}$ is the matrix containing the coefficients of state variables, $\underline{\underline{C}} \in \mathbb{R}^{u \times v}$ is the matrix that relates the v state variables with u measured outputs, denoted by $\underline{y} \in \mathbb{R}^u$ (Kalman, 1960). An important property to be assessed before designing a state estimator lays in the system observability. Specifically, A system is said to be observable if and only if it is possible to determine any (arbitrary initial) state in a finite time by using the available measurements and measured inputs (Kalman, 1960). Given the observability matrix $\mathbf{L_O}$:

$$\underline{\underline{L_O}} = \left[\underline{\underline{C}}\ \underline{\underline{CA}} \ldots \underline{\underline{CA^{v-1}}}\right]^T \tag{2.60}$$

The system 2.59 meets the observability criterion if rank($\underline{\underline{L_O}}$) = v, meaning that $\underline{\underline{L_O}}$ has to be full rank.

2.3.1 Kalman filter

In the field of dynamic systems, filtering refers to estimating the state vector $\underline{x}$ at the current measurement time. Modelling errors and measurement noise represent an inherent characteristic of every real system. Kalman filter considers the randomness of $\underline{x}$ and $\underline{y}$, meaning that Eq. 2.59 has to be rewritten as follows:

$$\begin{cases} \frac{d\underline{x}}{dt} = \underline{\underline{A}}\underline{x} + \underline{\xi}(t),\ x(0) = \underline{x}_0 + \underline{\xi}_0 \\ \underline{y} = \underline{\underline{C}}\underline{x} + \underline{\eta}(t) \end{cases} \tag{2.61}$$

Here, $\underline{\xi}(t)$ and $\underline{\eta}(t)$ represent the modelling and measurement errors, respectively. The errors are assumed to behave as white and gaussian, meaning that the following equations are satisfied:

$$E(\underline{\xi}(t))=0;\ E(\underline{\xi}(t_i)\ \underline{\xi}(t_j)^T)=\underline{\underline{Q}}(t)\delta(t_i-t_j) \quad (2.62a)$$

$$E(\underline{\eta}(t))=0;\ E(\underline{\eta}(t_i)\ \underline{\eta}(t_j)^T)=\underline{\underline{R}}(t)\delta(t_i-t_j) \quad (2.62b)$$

$$E(\underline{x}(0))=\underline{x}_0;\ E([\underline{x}(0)-\underline{x}_0][\underline{x}(0)-\underline{x}_0]^T)=\underline{\underline{P}}_0 \quad (2.62c)$$

$$E(\underline{\eta}(t_i)\underline{\xi}(t_j)^T)=0;\ \underline{\underline{P}}=\ E([\underline{x}(t)-\underline{\hat{x}}][\underline{x}(t)-\underline{\hat{x}}]^T) \quad (2.62d)$$

Where $\underline{\underline{Q}}(t)$ is the model error covariance matrix, $\underline{\underline{R}}(t)$ is the measurement covariance matrix, $\underline{\underline{P}}$ is the estimated state error covariance matrix, and δ is the dirac's delta. It is important to emphasize that under the abovementioned assumptions there is neither self-correlation of error variables (Eq. 2.62a,b), nor correlation between modelling and error vectors (2.62d). The Kalman filter assumes the following form (2.60):

$$\begin{cases} \frac{d\underline{\hat{x}}}{dt}=\underline{\underline{A}}\underline{\hat{x}}+\underline{\underline{K}}(t)\ (\underline{y}-\underline{\underline{C}}\underline{\hat{x}}) \\ \underline{\underline{K}}=\underline{\underline{P}}\ \underline{\underline{C}}^T\ \underline{\underline{R}}^{-1} \\ \frac{d\underline{\underline{P}}}{dt}=\underline{\underline{P}}\ \underline{\underline{A}}^T+\underline{\underline{A}}\ \underline{\underline{P}}^T+\underline{\underline{Q}}-\underline{\underline{P}}\ \underline{\underline{C}}\underline{\underline{C}}^T\ \underline{\underline{R}}^{-1}\underline{\underline{C}}\ \underline{\underline{P}} \end{cases} \quad (2.63)$$

In Eq. 2.63, $\underline{\underline{K}}$ represents the Kalman filter gain, which represents a dynamic weighting factor that determines the relative influence of the measurement versus the model prediction in the state update. Therefore, it balances the uncertainty in the prediction and the observation. The matrix definition (second equation of 2.63) is the result of a least square optimization (Kalman, 1960). The third equation of the system (2.63) represents the Riccati Equation, which governs the time evolution of the error covariance matrix $\underline{\underline{P}}$. Specifically, it describes how the prediction error evolves over time in response to the system dynamics and measurement updates. The solution of the Riccati equation is used to compute the Kalman gain, ensuring that the filter remains optimal under the assumption of linearity and white, Gaussian noise (Çimen, 2008).

2.4 References

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical Methods, 5(16), 3790–3798. https://doi.org/10.1039/c3ay40582f

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brereton, R. G. (2009). Chemometrics for Pattern Recognition. In Chemometrics for Pattern Recognition. Wiley.

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical Methods, 6(9), 2812–2831. https://doi.org/10.1039/c3ay41907j

Çimen, T. (2008). State-Dependent Riccati Equation (SDRE) Control: A Survey. IFAC Proceedings Volumes, 41(2), 3761–3775. https://doi.org/10.3182/20080706-5-kr-1001.00635

Kalman, R. E. (1960). On the general theory of control systems. IFAC Proceedings Volumes, 1(1), 491–502. https://doi.org/10.1016/s1474-6670(17)70094-8

Liao, S. M., & Akritas, M. (2007). Test-based classification: A linkage between classification and statistical testing. Statistics and Probability Letters, 77(12), 1269–1281. https://doi.org/10.1016/j.spl.2007.03.012

Montgomery, D. (2013). Design and analysis of experiments. John Wiley \& Sons, Inc.

Nyamundanda, G., Brennan, L., & Gormley, I. C. (2010). Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics, 11. https://doi.org/10.1186/1471-2105-11-571

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 61(3), 611–622. https://doi.org/10.1111/1467-9868.00196

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Chapter 3: Online monitoring of milk rennet coagulation using Raman spectroscopy

In this chapter, the development of a novel system for online monitoring of milk coagulation is discussed. Specifically, this section presents the experimental results obtained from rheological and Raman measurements, along with their analysis using the relevant algorithms. This chapter refers to a previously published work by the author (Sibono et al 2025a; Sibono et al 2025b).

3.1 Milk rennet coagulation monitoring: Cutting time prediction and Chemical footprinting

Milk coagulation is the first stage of cheesemaking (Dufour, 2011). This process typically occurs through the acidification of milk or the enzymatic action of a coagulant, the rennet, which trigger the milk phase transition from liquid to a coagulated gel called curd. Among the possible solutions, rennet coagulation is the most studied process in this framework (Abdelgawad et al., 2014; Castillo, 2006). The renneting step concludes when the gel undergoes cutting, a mechanical operation that promotes the separation of the whey from the solid part of the gel matrix. A proper selection of curd cutting time after rennet addition is crucial to guarantee the desired product quality since a delay in curd cutting can cause the excessive water retainment, while premature cutting generates losses in final yield (Castillo, 2006). Several methods for predicting cutting time have been applied and developed. In traditional practice, the gel is cut after a predetermined reaction time or by knife test, based on the operator's evaluation of the textural and visual properties of the gel evolved during rennet addition (Derra et al., 2018). In the last decades, lactodynamography and low-amplitude dynamic oscillatory rheology have been employed to respectively keep track of curd firmness and elastic modulus dynamics during coagulation, as they are recognized as a direct and reliable variable to assess gel physical properties variations during coagulation (Esteves et al., 2002; Miloradovic et al., 2020; Stocco et al., 2015). Nonetheless, such methods require the introduction of an external body into the coagulation system, which is invasive and unsuitable in an industrial dairy plant due to hygienic concerns (O'Callaghan et al., 2002). From this exigency, the use of optical sensor such as NIR backscattering arose as a non-destructive tool to monitor the evolution of gel firmness (Abdelgawad et al., 2014; Nicolau et al., 2015; O'Callaghan et al., 1999). Specifically, optical data were correlated with the experimental cutting time determined using reference methods, including lactodynamography, rheological analysis, or the expert judgment of the cheesemaker (Arango & Castillo, 2018; Nicolau et al., 2015; Villaquiran et al., 2024). Only in the last ten years few works applied multivariate statistics and chemometrics to analyse the evolution of NIR spectrum over time after rennet addition to study the milk gelation dynamics, thus offering a broader understanding of the process behavior (Lyndgaard et al., 2012; Strani et al., 2021). However, the use of NIR spectroscopy, despite being effective in understanding physical properties progress during renneting, has many limitations in detect the chemical footprint of coagulating milk because of the high sensitivity of NIR signal to water, which hinders the spectroscopic information provided by organic molecules (De Beer et al., 2011). To this concern, Raman

spectroscopy has found high utilization in several fields of dairy such as milk composition analysis and safety assessment, thanks to its high sensitivity to fats and proteins and weak water scattering (He et al., 2019; Khan et al., 2023). Conversely, Raman techniques have not found any substantial application in the field of milk coagulation monitoring prior to the research activity presented in this dissertation, with the exception of a preliminary study, in which the times at the occurrence of reference optical measurements (e.g. signal area inflection) from Raman signal were correlated with the curd cutting time (Puerta et al., 2017). However, the research was limited to one experimental condition, lacking a detailed description of the wavenumbers involved in coagulation. In this regard, only few individuals molecules such as tryptophan and riboflavin have been employed as biotracers of milk coagulation and syneresis (Fagan et al., 2011; Felfoul et al., 2022; Panikuttira et al., 2019).

The research effort presented herein aims to establish a basis for the integration the quantitative prediction of curd cutting time with the qualitative identification of potential biotracers, through the use of a new analytical technique and chemometrics for a broader understanding to investigate the renneting process under different operating conditions. Specifically, Raman information is used as an indirect information of gel physical properties. Additionally, the wavenumbers that significantly change during coagulation are identified, providing a deeper understanding of the chemical evolution of milk constituents during renneting. The results reported in this work align to the current demand of dairy industry of implementing new techniques for monitoring dairy food processes (Hayes et al., 2023).

3.1.1 Materials and methods

3.1.1.1 Sample preparation and experimental design

For the analysis, 12 samples of low-heat bovine skimmed milk powder (Bongiovanni S.r.l, Cuneo, Italy) with 0.6% (w/w) of fat and 34% (w/w) of protein were dissolved in 180 g of distilled water to yield a reconstituted milk sample with 10 % (w/w) of milk powder, corresponding to a protein content of 3.4% (w/w) (Strani et al., 2021). The solution was mixed for 20 minutes at room temperature using a magnetic stirrer at a frequency of 350 rpm (Grassi et al., 2022). Subsequently, the milk sample was stored at 4°C for 30 minutes to enhance the casein hydration process (Sandra & Dalgleish, 2005). Milk pH after reconstitution was 6.71±0.01 (at 4 °C). The milk solution pH was

adjusted to 6.5 at 4 °C using the procedure reported by Fagan et al. (2007) to determine the amount of citric acid (≥99.5 %, Sigma-Aldrich) to add. $CaCl_2$ (≥99 %, Sigma-Aldrich) was then added to the solution to improve the casein micelles aggregation, reaching a concentration of added salt equal to 3 mM.

The experimental data was collected for all combinations of three different temperature levels (34 °C, 36 °C, 38 °C) and two rennet concentrations (0.0145 IMCU mL^{-1} of milk, 0.029 IMCU mL^{-1} of milk). The proposed renneting temperatures are consistent with those employed for several cheese varieties (Britten & Giroux, 2022), while the operative rennet concentration values were recommended by the supplier. Samples were heated to the operative temperature by means of a water bath. When the thermal equilibrium was reached, the operative amount of Chymosin 85% ± 3% and Pepsin 15% ± 3% calf rennet solution (Hjemmeriet, Denmark, 145 IMCU mL^{-1} of rennet) were added to the milk, which was stirred with a spatula for 30 s, thus starting the renneting reaction. An aliquot of 6 mL was extracted through a Pasteur pipette for the rheological measurement. Raman and rheological measurements were initiated simultaneously. Each experimental condition was performed in duplicates.

3.1.1.2 Rheological analysis

Milk renneting experiments were performed through time sweep tests using an Anton Paar rheometer (MCR 102 Anton Paar GmbH, Austria) equipped with a pre-heated double gap concentric cylinder geometry (1 mm). Tests were performed for 1 h, applying a constant strain (1%) at a constant frequency (1 Hz). Elastic modulus (G´) values were registered every 30 seconds to measure the variations of gel viscoelastic behavior during its coagulation. Cutting time was defined as the point when G' reached 30 Pa, while gelation time was defined as the point when G' reached 1 Pa (Abdelgawad et al., 2014).

3.1.1.3 Raman measurements

Raman acquisitions were performed through Raman immersion probe (Ocean Insights, FL, United States) using a 785nm fiber coupled laser (B&W TEK Inc., Newark, United States) with an output power of 200mW at the sample. This was coupled to a QE Pro-Raman+ (Ocean Insights, FL, United States) operating in a range of 350-3033 cm^{-1} and a resolution of 10-14 cm^{-1}. Every acquisition was obtained by averaging 30 spectra, with an integration time of 1 *s*, resulting in a sampling interval of 30 seconds. This allows the overlap of Raman acquisitions with the rheological sampling. Each time run provided 953 Raman intensities for 120 temporal acquisitions.

3.1.1.4 Data preprocessing

Data were preprocessed through the Savitzky-Golay filter to smooth the data by fitting successive windows of data points to a polynomial of a specified degree while preserving the original shape of the signal. A second-order polynomial and a frame length of 15 were used as this preprocessing gave sufficient smoothing without removing the relevant information. Raman data were gathered in a $\tilde{I} \times J \times K$ structure, where $\tilde{I}$ denotes the 120 time acquisitions, J the 953 Raman shifts, and K indicates the 12 experimental runs. For each experiment, the initial time acquisition was excluded from the analysis as its score value resulted to be an outlier. The removal of the first spectral acquisition gives rise to 119 acquisitions over time, resulting in a $\mathbf{X}$ ($I \times J \times K$) three-dimensional matrix, where the symbol I identifies the new number of time points for each batch. Therefore, two preprocessing protocols are proposed:

1) Mean-centered and smoothed data were analyzed to study the evolution of the physical properties of coagulating milk, to predict the curd cutting time under different operative conditions.
2) Mean-centered, smoothed and baseline corrected data were used to investigate the signal variation of various Raman shifts, with the goal of identifying which chemical bonds is most relevant for the coagulation process.

For the sake of cutting time prediction, Baseline correction was not applied to the data because it is a clear indicator of the phenomenon of interest, originating from the textural change from liquid to gel. The preprocessing protocol (1) was used to predict cutting time, accounting for changes in physical properties after rennet addition. The baseline variations observed over time are a spectroscopic result of sample scattering.

To better illustrate the chemical changes in the system over time, Raman spectra were subsequently baseline-corrected (Protocol 2) using a Whittaker filter (Eilers, 2003), with a penalty of 10^4 and a weight value equal to 0.1. For a robust assessment of bonds' signal variation during renneting, the spectra for each acquisition time were averaged across all batches, thus obtaining a single $I \times J$ matrix. The rationale behind this operation is that the truly descriptive features of the coagulation process are expected to vary consistently across all trials, while less important features could be mistakenly identified as significant if the experiments were analyzed individually. This approach helps to reduce the presence of any PCA loadings that are not informative for chemical variations of renneting and removes the undesired presence of intra-batch baseline differences that affect the interpretation of loading values. Raman spectra were also subjected to a one-

dimensional 3-order median filtering along renneting times to dampen the effect of eventual outliers among the score values from baseline corrected data, thus removing cosmic spikes (Pratt, 2007). Noteworthy, Savitzky-Golay filtering was applied along the Raman shift dimension, while median filtering was performed across the coagulation times for each individual Raman shift, thus avoiding overlapping effects of the two techniques. In summary, in the whole work Raman information is utilized for two purposes: 1) the downwards shift of signal is used to monitor the variation of physical properties as intended for NIR measurements, and 2) the changes in specific Raman shifts, as a result of the redistribution of milk constituents during its coagulation. For the latter purpose, only 500-710 cm^{-1} and 760-2483 cm^{-1} wavenumber bands were considered in order to remove part of the low-informative wavenumbers, for which the spectrum resulted not to significantly change (e.g. fats). The first two PC scores were analyzed for this purpose. By way of example, figure 3.1 shows the typical evolution of Raman spectrum along time after rennet addition and the relevant baseline variation over time.

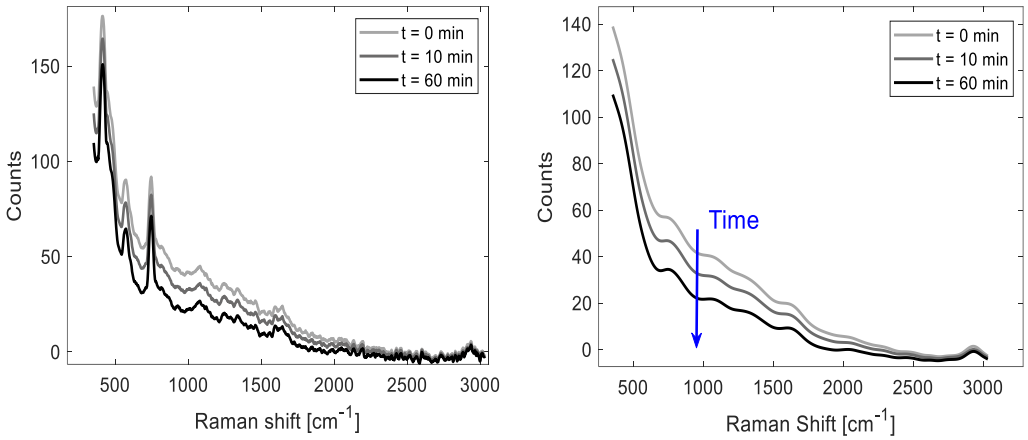


Figure 3.1 Unbaselined-Savitzki-Golay filtered Raman spectra captured over time during milk renneting of reconstituted skim milk powder (left), and their relevant baseline values (right). The plot was obtained by renneting milk at 38 °C with a rennet concentration of 0.0145 IMCU ml^{-1} of milk.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

3.1.1.5 Cutting time Prediction model

Each experiment was characterized by an apparent monotonically decreasing Raman spectra baseline during coagulation (Figure 3.1). A three-way data structure ($I \times J \times K$) was unfolded along times, thus giving rise to a $IK \times J$ data matrix, denoted as $\underline{\underline{X}}_u$, Here, IK was equal to 1428. Figure 3.2 shows the unfolding procedure for the three-way data structure. The spectral signal was then analyzed through PCA to extract the relevant patterns from spectra and to observe them in a lower-dimensional subspace (Brereton, 2009). The first PC, which accounts for the main variability in the data (87% when all batches were considered in a single PCA model), was employed to investigate the dynamic of baseline variations during renneting, as it was the only score showing a clear pattern (Lyndgaard et al., 2012). Therefore, for cutting time prediction concerns, the data matrix was decomposed using one-dimensional scores and loadings according to Eq. 3.1:

$$\underline{\underline{X}}_u, = \underline{t}\,\underline{p}^T + \underline{\underline{E}} \tag{3.1}$$

Where $\underline{t}$ is the $IK \times 1$ score vector, $\underline{p}$ denotes the $J \times 1$ loading vector and $\underline{\underline{E}}$ the residuals matrix. The prediction of rheological cutting time is typically performed using linear models that correlate this time with specific characteristic times experimentally derived from light backscatter and fluorescence, such as time at the inflection point, time at the maximum and minimum of the second derivative, to name a few (Abdelgawad et al., 2014; Castillo et al., 2006; Panikuttira et al., 2020). Eq. 3.2 was consistently used to correlate the rheological cutting time to the time at inflection point of the first PC:

$$\hat{t}_{cut} = \theta t_{max} \tag{3.2}$$

Where the time at inflection point (t_{max}) is defined as the time at which the second order difference of PC-score is zero, and rheological cutting time is the time at which G' becomes greater than 30 Pa. Such characteristic times were experimentally determined. The Pearson correlation coefficients, denoted as *r*, were computed to investigate the mutual correlation among different time parameters. Additionally, RMSE was normalized to the average experimental cutting time to calculate the coefficient variation RMSE. Normalized Mean Bias Error (NMBE) was also calculated to assess potential overestimation or underestimation in the model. In this study, t_{max} from Raman signal was analyzed in order to directly compare the model's performance metrics like R^2 and root mean square error (RMSE) resulting from the application of the methodology presented here with those obtained from more consolidated techniques. Consequently, in order to evaluate

the predictive reliability of $\hat{t}_{cut}$ estimations, cross-validation was performed using a 6-fold cross-validation approach, where each experimental condition (consisting of two replicates) was iteratively excluded from the calibration set. This method allows for a robust assessment of the model's predictive capability under experimental conditions other than those used for calibration. All spectra associated with the validation set data were projected into the subspace of the first principal component. The estimated cutting times for each iteration were then compared with the rheological cutting times to compute the R2 in cross-validation and the Root Mean Square Error in Cross-validation (RMSECV).

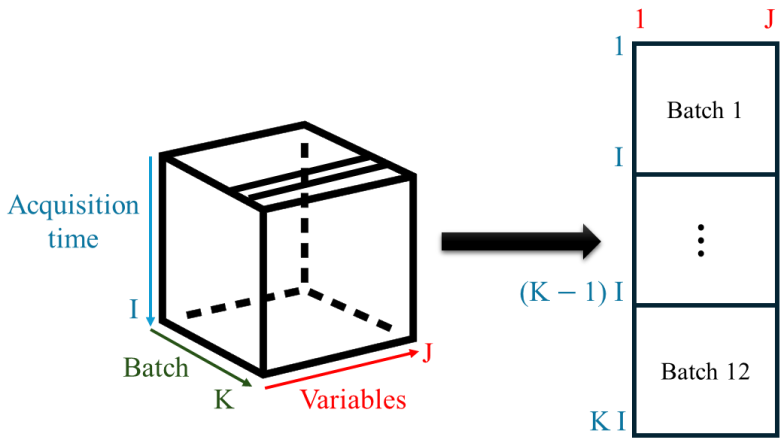


Figure 3.2 Unfolding procedure of the three-way data structure obtained from Raman measurement.

To further validate the mathematical model, the Eq. (3.2) derived from fitting the 12 experiments was tested on two additional replicated coagulation runs at 37 °C with a rennet concentration of 0.0218 IMCU mL^{-1} of milk, constituting an external validation, since such condition was not included in the DoE.

3.1.1.6 Software

Every computations related to PCA, cutting time predictions, and model validation were carried out through Matlab®2023a. Rheological analysis was supported with RheoCompass™ software (Version 1.32, Anton Paar GmbH, Austria), while OceanView 2.0.14 (Ocean Optics, USA) was used for the Raman measurements.

3.1.2 Preliminary analysis of milk coagulation using Raman Spectroscopy: Results and discussion

This section will report the results obtained from the preliminary analysis of milk coagulation using Raman spectroscopy. Since the study of the development of a Raman-based sensor to monitor milk coagulation was divided into two complementary protocols, the results will be shown in the following order: firstly, Raman spectra are analyzed to develop a prediction model of curd cutting time, for which baseline variation of signal is assessed. For this purpose, Raman information was coupled with the consolidated rheological technique, employed as a reference method to determine the actual cutting time. Secondly, the results related to the chemical foot printing of renneting milk are discussed. In this latter case, Raman signal is baseline corrected in order to remove the effect of physical variations which cover the chemical information. The results reported in this subchapter will serve as a foundation for the following study, which explores the process kinetics.

3.1.2.1 Gel cutting time prediction

The scores corresponding to the first principal component were experimentally determined and used to predict the curd cutting time. Particularly, the second order discrete difference of PC1 scores was computed. When such value attained zero, the time at inflection point was obtained and substituted to eq. 3.2 to estimate the cutting time. Twelve renneting processes under six different experimental conditions were carried out and analyzed for this aim. Rheological cutting time was consistently achieved in all experiments when the elastic modulus exceeded 30 Pa, as shown in figure 3.3. The evolution of G' along time exhibits a more rapid dynamic with increasing temperature and rennet concentration. Higher values of these factors resulted in earlier occurrences of both gelation (i.e. time at which G'>1 Pa) and cutting times as a result of improved enzymatic kinetics. Notably, after 45 minutes of renneting the G' value associated to the experiment conducted at 38 °C and a rennet concentration of 0.0145 IMCU mL^{-1} surpassed that of milk at 34 °C with a rennet concentration of 0.029 IMCU mL^{-1}. This result highlights that a four degrees higher temperature exerts a stronger influence than the lower enzyme concentration, thus yielding to the crossover of the rheological curves.

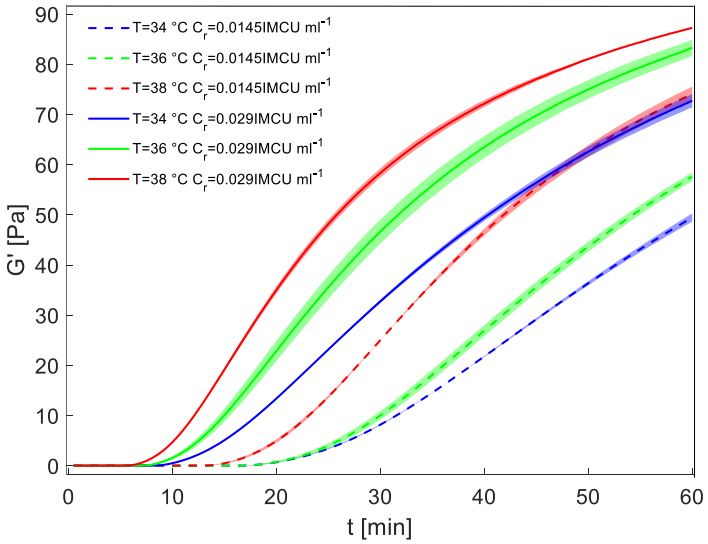


Figure 3.3 Evolution of G' value over time after rennet addition for different temperatures and rennet concentrations. The shaded area denotes the confidence interval (Confidence level of 95%).

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

Raman signal baseline variation was considered for predicting cutting time using a spectroscopic characteristic time (i.e. t_{max}). PCA was performed on smoothed Raman spectra to study the dependence of milk gel firmness on time, temperature, and rennet concentration. Figure 3.4 shows the experimental values of PC1 scores for each run. When PCA was conducted on the J×IK data matrix (Figure 3.4a), the first PC explained 87% of the total variance. PCA was also conducted on each I × J batch (Figure 3.4b), to better distinguish the time of the occurrence of PC1 inflection under varying temperature and rennet concentration. In this case, mean centering preprocessing was carried out individually for each batch matrix, enhancing the clarity of differences across experimental conditions. Here, PC1 has approximately 99% of the data variance for all the batches. As can be inferred, steeper PC1 score curves were obtained for higher temperatures and rennet concentrations. It is important to note that the impact of temperature increase was more noticeable at lower enzyme concentrations indicating an interaction effect between the two

factors. Similar results were observed in an other study that employed light backscattering to estimate the cutting time (Abdelgawad et al., 2014).

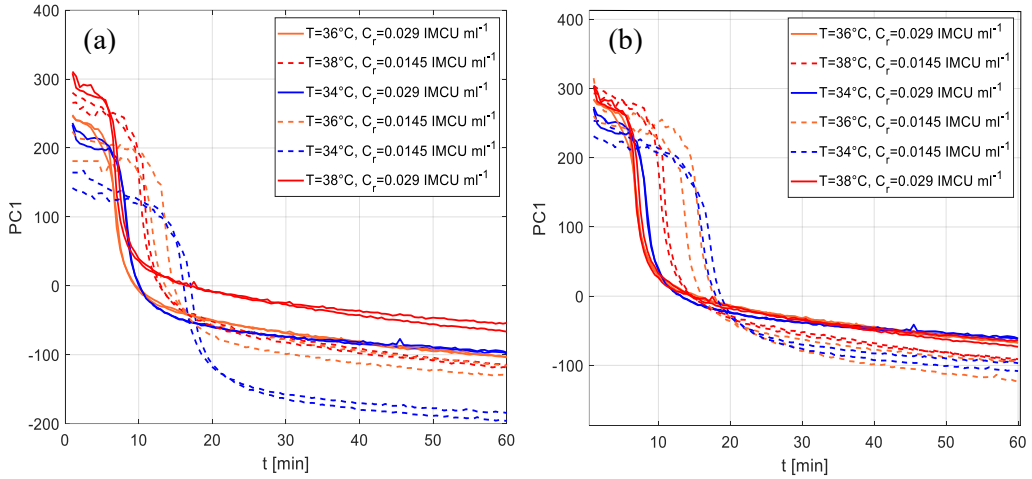


Figure 3.4 PC1 scores vs time during milk renneting. In plot (a) data were modelled as a time-wise unfolded three-way structure, while in plot (b) PCA model was executed on each individual batch data bidimensional matrix.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

Table 3.1 depicts the analysis of the experimental results, showing the values of cutting time, gelation time, time at an inflection point, and their Pearson correlation coefficient. Such metrics are used to examine the relationships between characteristic times observed during renneting as they provide more details of coagulation dynamics (Castillo et al., 2006; Nicolau et al., 2015; Villaquiran et al., 2024). It was observed a strong positive correlation between a baseline variation parameter like the inflection point (i.e., t_{max}) and the rheological characteristic times (t_{gel} and t_{cut}). This indicates that there is a clear relationship between the network formation phase, which denotes the elastic modulus variations, and the earlier aggregation phase, during which a sigmoidal behavior of PC1 score values can be observed. During this first stage no elastic modulus variation was detected.

Table 3.1 Rheological and spectroscopical characteristic times obtained from all the runs and the resulting correlation coefficients.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

T [°C]	C_{rennet} [IMCU mL^{-1}]	t_{gel} (G'>1 Pa)	t_{cut} (G'>30 Pa)	t_{max}
36	0.0290	9.5	22.5	6.5
36	0.0290	10.0	23.5	7
38	0.0145	16.5	32.5	10.5
38	0.0145	16.0	32.5	11.0
34	0.0290	11.0	28.5	8.0
34	0.0290	11.0	29.0	8.5
36	0.0145	21.0	42.5	15.5
36	0.0145	20.5	41.5	13.5
34	0.0145	21.0	46	17.0
34	0.0145	21.0	45.5	16.0
38	0.0290	7.5	18.5	7.0
38	0.0290	9.5	18.5	7.5

r	t_{gel}	t_{cut}	t_{max}
t_{gel}	1	0.978	0.973
t_{cut}	0.978	1	0.972
t_{max}	0.973	0.972	1

By referencing the cutting time derived from the CoAguLite light backscatter probe (Castillo et al., 2006), regression analysis was performed using parameters obtained from the Raman laser sensor for that prediction. Similarly, the cutting time was estimated using a linear relationship with the experimental time at the inflection point, obtained by analyzing PC1 scores during renneting. When using the whole dataset (12 runs) for estimating t_{cut} the obtained linear equation was $\hat{t}_{cut} = 2.93\ t_{max}$ and the parametric confidence interval was equal to 0.18 obtained from a two-sided t-test with a confidence level of 95%. For the presented equation, a coefficient of determination in calibration R^2 resulted to be equal to 0.901 with an RMSE of 3.2 min, as it can be observed in figure 3.5. Moreover, the model exhibited a coefficient variation RMSE of 0.1 and a NMBE of -1.7%, indicating a negligible model underestimation of rheological cutting times. Such a result is comparable to the one reported in another study where Raman spectroscopy was employed for similar purposes (Puerta et al., 2017). Conversely, such performance result seems to be lower than

the ones reported by Nicolau et al. (2015), where an R^2 0.977 was obtained when calibrating a linear model with intercept using NIR light backscattering. However, it should be noted that the model's performance in the present study was calculated based on the validation data, providing a more robust performance evaluation, as it assesses the model's predictive capabilities for cutting time. Each replicated experiment was iteratively removed from the calibration set, and the model's predictions were compared to the true values. Overall, an R^2_{cv} of 0.872 was obtained from the validation procedure, while the RMSECV was equal to 3.6 min (Figure 3.6). Such results indicate a satisfactory prediction of the rheological cutting time when using the linear model presented in Eq. 3.2, providing a strong quantitative basis for monitoring and controlling the coagulation process. It is worth noting that the model's prediction errors were not significantly affected by the experimental conditions under validation. Specifically, the predictions of rheological cutting time for batches belonging to the boundaries of the design space were no less accurate than those within the design space (Figure 3.6). Finally, the time at inflection point was used to estimate the rheological cutting time of two additional test samples at a temperature of 37 °C and a rennet concentration of 0.0218 IMCU mL-1 of milk. In this case, by multiplying the model's slope (i.e. 2.93) by t_{max}, the estimated cutting time was 27.8 min and 30.7 min while the respective rheological cutting time was 26.5 and 28 minutes, corresponding to a RMSE in external validation of 2.1 min. The high performance in both cross-validation and test procedures indicates that the model captures the majority of the variance in the cutting time when varying the operating condition, and demonstrate that the signal baseline changing consistently responded to the kinetic changes given by rheological assays under different temperatures and rennet concentrations.

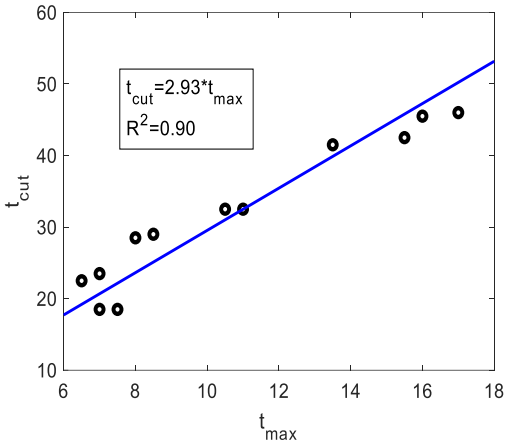


Figure 3.5 Cutting time vs time at inflection point for all the milk coagulation runs.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

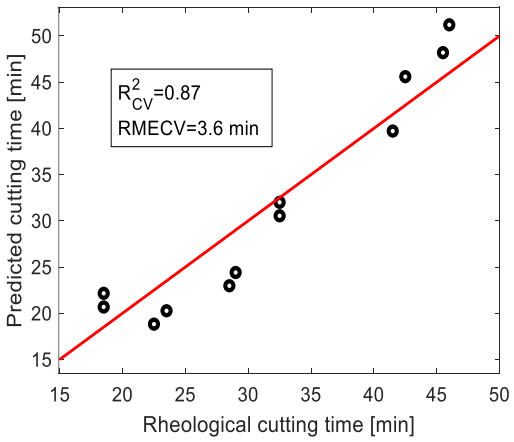


Figure 3.6 Predicted vs Rheological cutting times of the validation batches from the 6-fold cross validation procedure.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

3.1.2.2 Chemical variations of renneting milk

Milk renneting involves the chemical composition variation of the food matrix due to the hydrolysis reactions triggering the redistribution of milk constituents during the process. Therefore, different spectra dynamics caused by chemical changes are expected (Svendsen et al., 2016). The baseline spectra variations reported in this work are mainly attributed to the physical modification of liquid to gel process (Figure 3.1). Baseline correction filters out the main influence of the liquid-gel transition, allowing for the observation of other behaviors that the dominant phase-transition phenomenon would otherwise overshadow. As shown in figure 3.7, baseline correction allows to observe the main molecular bonds of milk. Similar prominent Raman peaks were reported in other studies (El-Abassy et al., 2011; Khan et al., 2023; Mendes et al., 2016). Specifically, the spectrum displays the CH_3 bending at 1081 cm^{-1}, the C=C stretching at 1654 cm^{-1}, and the fatty acids C-H asymmetric stretching at 2922 cm^{-1} (Khan et al., 2023; Mendes et al., 2016). A low Raman signal was observed in the neighborhood of 2900 cm^{-1} due to the low fat content of skim milk.

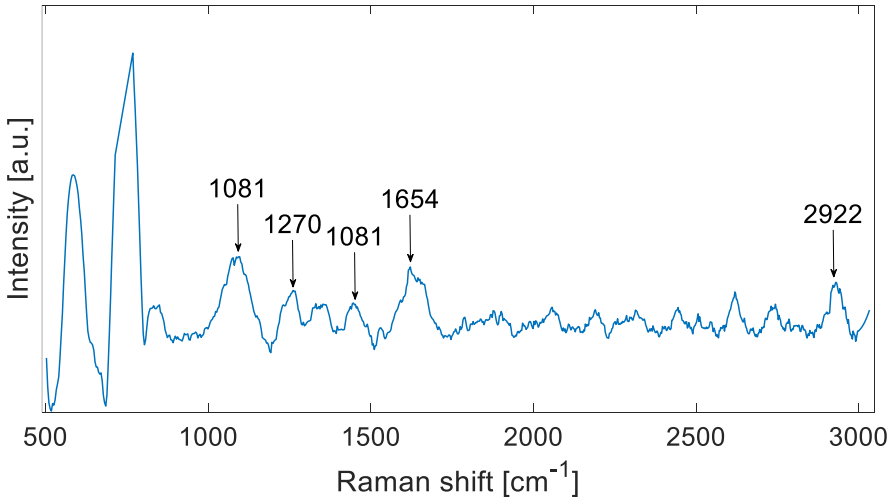


Figure 3.7 Baseline corrected spectra obtained by renneting milk at 38 °C with a rennet concentration of 0.0145 IMCU ml^{-1} of milk.

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

Raman spectra were preprocessed to identify the chemical variations during renneting. Figure 3.8a shows the temporal variation of baselined spectrum. The gradient of color depicts the time at which the spectrum was collected. PCA of data, after smoothing, baseline correction, and batch-wise averaging, was conducted for the analysis. As reported in figure 3.8b-c, loading values and Raman spectra variations showed the dominance of specific spectral features throughout the process. Concerning the preprocessed spectra, the most prominent peak was found at 577 cm^{-1}, which corresponds to C–C–C deformation and C–O twisting, generally associated with carbohydrates (de Almeida et al., 2012; Z. Y. Zhang et al., 2022). The corresponding PC1 loading was found to be the dominating one. Such a result could be attributed either to the complex structural rearrangement of milk constituents during milk coagulation or to a baseline variation of Raman intensity in the lower wavenumbers that was not fully corrected during the data preprocessing. Furthermore, three significantly negative loadings were individuated for 544 cm^{-1}, 775 cm^{-1}, and 1002 cm^{-1}, which were respectively assigned to S-S stretching, C–C–O deformation or C–C–H deformation bonds of tryptophan and ring breathing structure of phenylalanine (Panthi et al., 2023; Walther et al., 2022), meaning that a signal decreasing was observed for such Raman shifts. Particularly, S-S bond reactivity is increased due to their exposure to casein structural changes during coagulation (Hashizume & Sato, 1988). Tryptophan has emerged as a prominent biomarker for milk coagulation due to its strong response to spectrofluorimetric excitation (Panikuttira et al., 2019), related to the modification of protein environment during coagulation. Furthermore, Phenylalanine is directly involved in milk coagulation as it represents the specific hydrolysis site attacked by rennet in the earliest times of the reaction, which cleaves the C–N Phenylalanine105–Methionine106 bond in κ-casein molecule, slicing it into para-κ casein and soluble glycomacropeptide fractions (Banks & Horne, 2003). This is consistent with the fact that the aggregation rate depends on the concentration of free paracasein sites, which implies that the dynamic behavior of coagulating milk during this phase is dependent on rate and extent of κ-casein proteolysis (Lyndgaard et al., 2012). A high positive loading value was found at 1107 cm^{-1}, which is assigned to PO_4^{3-} stretch modes (Walther et al., 2022; Zhao et al., 2020). In this case, the increasing signal of phosphate stretch Raman signal can be ascribed to the dissolution of micellar calcium phosphate during network formation (Lamothe & Britten, 2023). This means that the steady decreasing/increasing trend for such compounds' signal during renneting is consistent to the monotonical behavior of PC1 scores as shown in figure 3.8d. The second principal component showed informative results as well (Figure 3.8c). The highest PC2 loadings were obtained for

wavenumbers of 1077–1086, generally attributed to C–O–H deformation, C–O vibration and C–C vibration of aspartic acid (Zhao et al., 2020). At these Raman shifts, a non-monotonic behavior was observed as the intensity values reached a minimum value and then increased during the network formation phase, which explains the dynamic behavior of PC2 scores (Figure 3.8e). It is known that aspartic proteinases cover an important fraction in chymosin composition (Chitpinityol & Crabbe, 1998). The change in signal in correspondence of these Raman shifts may be due to the dynamic of the lysis process performed by the rennet: the enzymatic attack towards the κ-casein active site and the subsequent return to the non-bonded state of the enzyme could account for the presence of a extremum in the signal intensity trend across renneting. Finally, The second highest PC2 loading was found at 1654 cm^{-1} corresponding to the stretching of C=C bond (Gallier et al., 2011). For the chemical variation analysis, where the averaged spectra were studied, PC1 and PC2 explained 30% and 12% of total variance, respectively. The remaining amount of variance was dispersedly distributed in the higher dimensions without showing an intelligible pattern. After 1700 cm^{-1}, low loading values were observed because the compounds related to such Raman shifts (e.g. water bonds) are not involved in rennet-induced coagulation of milk.

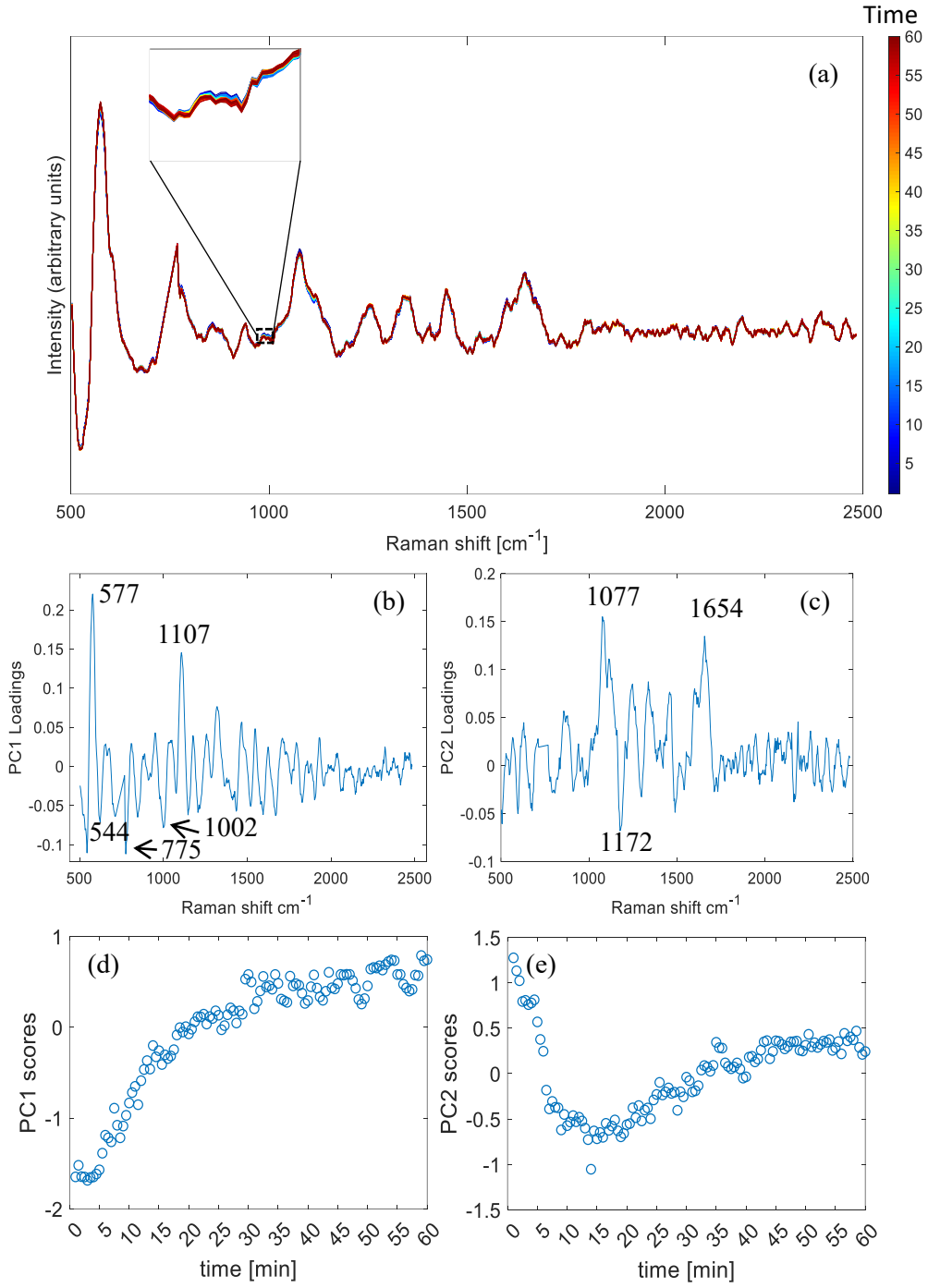


Figure 3.8 Results from the analysis of baseline, smoothed, and batchwise averaged spectra: comparison of spectra at different acquisition times (a) and the resulting PC1 and PC2 loading values (b, c). The PC1 and PC2

scores evolution over time is also reported (d, e).

Reproduced from: Leonardo Sibono, Stefania Tronci, Martin Aage Barsøe Hedegaard, Massimiliano Errico, Massimiliano Grosso, Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy, Applied Food Research, Vol. 5, Issue 1, 2025, 100701.

In conclusion of this subsection, averaging step was intentionally adopted to identify Raman shifts that contributed consistently across experiments, rather than focusing on batch-specific fluctuations. To support the validity of this choice, Figure 3.9 shows the temporal profiles of the signals corresponding to part of the Raman shifts highlighted in Figures 3.8b and 3.8c. These trajectories illustrate how the identified Raman shift signals exhibit a behavior that approximately aligns with those observed from the PCA (Figure 3.8), providing an additional check on the robustness of the features retained for interpretation. Nevertheless, this approach carries the limitation that the noise becomes substantial when examining individual experiments. For this reason, the analysis was used primarily as an exploratory tool, aimed at identifying general spectral trends rather than for practical monitoring tool.

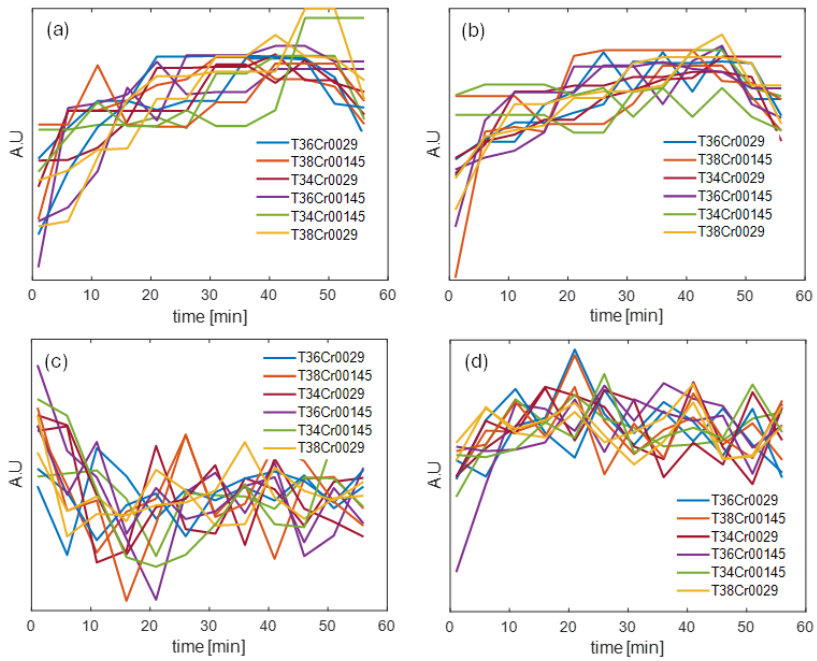


Figure 3.9 Evolution of intensity of different Raman shifts over time for each experiment. The arbitrary units for 577 cm^{-1} (a), 1107 cm^{-1} (b), 1077 cm^{-1} (c), 1176 cm^{-1} (d) signals show consistent behavior to the loading values reported in figure 3.8.

By way of example, the score plot obtained from the experiment conducted at T = 36 °C and Cr = 0.029 IMCU mL−1 is report in figure 3.10. As can be observed, the plot can be subdivided into three main regions that reflect the evolution of the coagulation process over time. In region 1, the samples progressively shift from the lower-right area of the plot (blue points) towards the upper-central region as time advances, indicating a rapid variation of the spectral profile. Upon reaching a maximum, PC2 scores begin to decline (region 2), while PC1 values exhibit a gradual decrease. Finally, in region 3, the PC1 scores continue to decrease, but the rate of change increasingly reduces, suggesting the approach to a steady state. The trend of PC1 scores reflects the enzymatic hydrolysis which mainly takes place in the early stages of the process and the subsequent micelles aggregation captured by PC2.

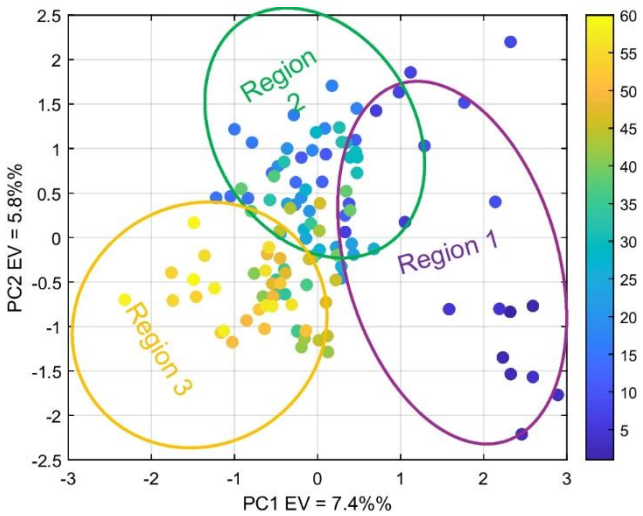


Figure 3.10 Score plot obtained from performing PCA on Raman spectra for a milk coagulation assay conducted at T = 36 °C and Cr = 0.029 IMCU mL−1.

3.2 Kinetic modeling of milk rennet coagulation

One of the most critical aspects of milk gelation is the development of a kinetic model of the process, as it plays a crucial role in the proper design of the coagulation reactor. Specifically, a reliable mathematical description of milk renneting allows both to predict the key parameters of plant operations and to continuously keep track of the system's state. The relationship between

optical and technological times (Discussed in the previous section of this chapter) can serve as groundwork to achieve a deeper knowledge of temporal behavior of hydrolysis reaction and visco-elastic properties during coagulation. In particular, the methodology presented in the previous subsection was employed to assess whether Raman spectroscopy could serve as a viable alternative to the well-established NIR-based spectroscopy in successfully estimating the cutting time. Such a methodology can be extended to develop an inline monitoring system for the milk constituents involved in the gelation process. A time by time understanding of coagulation evolution has the potential to aid cheese manufacturers in making better informed decisions aimed at ensuring product quality and standardizing functional properties.

This section of chapter 3 intends to expand the current knowledge in the field of PAT applied to dairy processes, by employing non-invasive measuring techniques coupled with kinetic modelling of reconstituted Skimmed Milk Powder coagulation to monitor the system's evolution during renneting. A pseudo-kinetic model is proposed to keep track of the dynamic behavior of milk constituents involved into renneting process, by investigating the PC scores and elastic modulus, which consistently vary with the organic molecules' concentration. A broader knowledge of the process chemistry allowed for the application of a real-time monitoring tool for real-time G' estimation. To assess the operational applicability of the model, the parameters were evaluated under different operative conditions.

3.2.1 Development of the kinetic model

The overall kinetics of milk coagulation consist on the combination of the enzymatic reaction and the micelle aggregation reaction (Lyndgaard et al., 2012). In the first step, which occurs in the earliest times of reaction, the specific hydrolysis site attacked by rennet cleaves the C–N Phenylalanine105–Methionine106 bond in κ-casein molecule, slicing it into non-soluble para-casein and soluble glycomacropeptide fractions (Banks & Horne, 2003). The evolution of hydrolysis reaction promotes the subsequent formation of hydrophobic para-casein micelles cross-linking sites (Carlson et al., 1987). During the aggregation process, when the calcium ion concentration is sufficient, the number of available cross-linking sites decreases due to the formation of permanent crosslinks among micelles. The reaction pathway can be therefore summarized as follows (Carlson et al., 1987):

$$n_\kappa \kappa \xrightarrow{R} \upsilon^* \rightarrow \upsilon \tag{3.3}$$

Here, symbol κ represents the κ-casein site, which is hydrolyzed by means of rennet enzyme, denoted as R, while n_κ is the number of κ-casein molecules needed to generate one crosslinking site. The cross-linking site is denoted as υ^*, of which consumption generates the permanent crosslink υ. It was assumed that the permanent crosslink number is proportional to the elastic modulus G' retrieved by means of rheological analysis (Carlson et al., 1987), meaning that the protein redistribution during renneting is strongly correlated to milk rheological properties. It has been documented that micelle aggregation only occurs after a certain conversion of κ-casein is obtained, which is attributed to the restricted number of reactive sites on the micelle surface that reduces the number of effective collisions in the early times of reaction (Van Hooydonk & Walstra, 1987). This phenomenon probably justifies the presence of a rennet-dependent lag phase in the storage modulus evolution (Y. Zhang et al., 2019).

For the coagulating system it is also assumed that κ-casein hydrolysis and cross-linking reactions follow a first order reaction kinetic. Such assumption is experimentally supported by other studies (Carlson et al., 1987; Yang et al., 2022). Given the scheme presented in Eq. 3.3, Carlson et al. proposed a kinetic model for κ-casein hydrolysis and micelles aggregation (Carlson et al., 1987):

$$\begin{cases} \frac{d[\kappa]}{dt} = -k_1[\kappa] \\ \frac{d[\upsilon^*]}{dt} = \frac{1}{n_\kappa}k_1[\kappa] - k_2[\upsilon^*] \\ \frac{d[\upsilon]}{dt} = k_2[\upsilon^*] \end{cases} \tag{3.4}$$

In Eq. 3.4, [κ] denotes the κ-casein site concentration, while $[\upsilon^*]$ and [υ] respectively represent the concentration of free cross-linking sites and permanent crosslinks. In this investigation, the original formulation provided in Eq. 3.4 is refined to model the coagulation process through Raman and rheological information. In the previous section, it was reported that after employing PCA on baselined spectra, PC1 and PC2 resulted to account for the variability in the hydrolysis reaction extent. Specifically, a signal decreasing of S-S stretching, C–C–O deformation or C–C–H deformation bonds of tryptophan and ring breathing structure of phenylalanine was observed, indicating that these compounds, known to be directly related to the binding sites of κ-casein (i.e. [κ]) , were captured from the multivariate analysis of Raman spectra. Such components' variability was captured by the first PC. On the other hand, PC2 loadings were attributed to the rennet aspartic acid proteinases and other organic molecules' bonds (C=C stretching and CH_2 bending) which resulted to significantly change during coagulation. These outcomes suggest that PC2 can be exploited as a tool for keeping track of cross-linking sites (i.e. $[\upsilon^*]$) evolution during renneting. This

means that PC1 and PC2 score values are here assumed to be respectively proportional to $[\kappa]$ and $[\upsilon^*]$ variables proposed in Eq. 3.4, while the elastic modulus G' obtained from rheological assay is considered to be proportional to the number of formed crosslinks υ, accordingly to the assumption made by Carlson et al. Therefore, S_1 denotes the PC1 score variable which describes the κ-casein hydrolysis evolution, whereas S_2 defines the PC2 score variable related to the cross-linking stage. The micelles aggregation is finally reflected by G' increasing. Here follows the pseudo-kinetic modification of the coagulation model provided by Carlson et al.:

$$\begin{cases} \frac{dS_1}{dt} = -\alpha_1 S_1 \\ \frac{dS_2}{dt} = n\,\alpha_1 S_1 - \alpha_2 S_2 \\ \frac{dG'}{dt} = m\,\alpha_2 S_2 \end{cases} \tag{3.5}$$

In this notation, the generic α_j is identified as a pseudo-kinetic constant for the j-th reaction which refers to the score values dynamic evolution. Moreover, n is a model coefficient that accounts for the proportional relationship between S_1 and S_2 for the hydrolysis reaction, while m accounts for the corresponding relationship between S_2 and G' in the aggregation phase. It is reported that no apparent micelle aggregation is observed until a critical conversion of κ-casein is reached (Van Hooydonk & Walstra, 1987; Yang et al., 2022), explaining the clear delay in gelation showed by the elastic modulus. Indeed, the model developed by Carlson et al. only considers the renneting system's behavior after the occurrence of gelation, that is generally defined as the time at which G'=1 Pa (Castillo et al., 2006). The presented model introduces a lag time t^* after which the crosslinking sites consumption gives a noticeable variation of elastic modulus, due to the gelation time being higher than zero.

The achievement of the asymptotic concentration for κ-casein and cross-linking sites (theoretically zero) is described by means of the asymptotic values $S_{1,\infty}$ and $S_{2,\infty}$ for S_1 and S_2, respectively. Such parameters are introduced to account for the system's progression towards the respective hydrolysis and aggregation equilibrium.

Eq. 3.5 is thus modified as follows:

$$\begin{cases} \frac{dS_1}{dt} = -\alpha_1 (S_1 - S_{1,\infty}) \\ \frac{dS_2}{dt} = n\,\alpha_1 (S_1 - S_{1,\infty}) - \alpha_2 (S_2 - S_{2,\infty}) \phi(t - t^*) \\ \frac{dG'}{dt} = m\,\alpha_2 (S_2 - S_{2,\infty})\, \phi(t - t^*) \end{cases} \tag{3.6}$$

According to the present formulation of Eq. 3.6, it is a non-linear model, in which the non-linearity is introduced by the activation function $\phi : \mathbb{R} \rightarrow [0,1]$ In this study, the activation function is $\phi(t) = H(t - t^*)$, where H(x) is the Heaviside function used to activate/deactivate the reaction rate of the second reaction.

If the model is divided into two reaction stages, an analytical solution can be derived for the problem. Specifically, Eq. 3.6 was first solved by assuming as null the reaction rate of the aggregation reaction (Which is valid for $t < t^*$) and afterwards it was considered in the model (Where $t > t^*$). The two analytical solutions were then merged into a unique system of non-linear equations by using the activation function. Here follows the solution (Eq. 3.7):

$$\begin{cases} S_1(t) = S_{1,\infty} + \left(1 - S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2(t) = f_1(t)\,\phi(t^* - t) + f_2(t)\;\phi(t - t^*) \\ G'(t) = f_3(t)\,\phi(t - t^*) \end{cases} \tag{3.7}$$

Where $f_1(t)$ and $f_2(t)$ are the functions describing the time evolution of S_2 during hydrolysis and aggregation, respectively, and $f_3(t)$ the function associated to the temporal progression of G'. Such functions can be written as (Eq. 3.8):

$$\begin{cases} f_1(t) = S_{2,0} + n\left(S_{1,0} - S_{1,\infty}\right)\left(1 - e^{-\alpha_1 t}\right) \\ f_2(t) = S_{2,\infty} + S_{2,0}\left(1 - e^{-\alpha_1 t^*}\right)e^{-\alpha_2(t - t^*)} + e^{-\alpha_2 t}\left\{\frac{n(S_{1,0} - S_{1,\infty})\left(\alpha_2 e^{\alpha_1 t + \alpha_2 t^* - \alpha_1 e^{\alpha_2 t + \alpha_1 t^*}}\right)e^{-\alpha_1(t + t^*)}}{\alpha_1 - \alpha_2} + \left[n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty}\right]e^{\alpha_2 t^*}\right\} \\ f_3(t) = G'_0 + m\,S_{2,0}\left(1 - e^{-\alpha_1 t^*}\right)\left(1 - e^{-\alpha_2(t^* - t)}\right) + m\left\{n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty} + \frac{n\alpha_2(S_{1,0} - S_{1,\infty})e^{-\alpha_1 t}}{\alpha_1 - \alpha_2} - \frac{n\alpha_2(S_{1,0} - S_{1,\infty})e^{-\alpha_2(t - t^*) - \alpha_1 t^*}}{\alpha_1 - \alpha_2} - \left[n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty}\right]e^{-\alpha_2(t - t^*)}\right\} \end{cases} \tag{3.8}$$

All steps to obtain the final forms of the equations are given in the Appendix B.

The enzyme attack on κ-casein sites can be assumed to occur randomly (Payens, 1989). As a result, the number of unhydrolyzed κ-casein sites per micelle is expected to follow a bell-shaped distribution. Consequently, there will be a point in time when most micelles shift from having a number of cleaved sites below a critical threshold to a majority where this number exceeds the threshold. This smooth transition is here taken into account using the hyperbolic tangent function as activation function, in order to regulate the deactivation/activation of the generation term for G'. Noteworthy, the derivative of such a function gives rise to a bell-shaped function. Therefore, the hyperbolic tangent function provides a smoothing effect, initially emphasizing the hydrolysis stage before t = t* and progressively weighting the aggregation stage when time after rennet addition surpasses t*.This approach is supported by previous studies, where, despite the presence of a well-defined transition time, the increase in G' occurs in a relatively smooth manner around the

gelation time (Arango & Castillo, 2018; Sibono et al., 2025). The activation function is thus approximated through the expression $\phi(t) = \frac{1}{2}\left(1 + \tanh\left(\frac{t-t^*}{\gamma}\right)\right)$ where g denotes the characteristic time of transition and was set equal to 0.2, as inferred from visual inspection of G' profiles.

For the sake of representing the variables as a fraction of the highest value, S_1, S_2 and G' were mapped in a range between 0 and 1 where 0 was attributed to the smallest value, and 1 to the highest one (refer to the parameter estimation section). The initial condition set for this problem is $PC_{1,0}$=1, $PC_{2,0}$=0 and G'_0=$G'_{0,exp}$. To model the whole process, the parameters $\underline{\theta} = [n, m, \alpha_1, \alpha_2, t^*, S_{1,\infty}, S_{2,\infty}]$ are used, where $\underline{\theta} \in \mathbb{R}^7$.

3.2.2 Pseudo-kinetic parameters estimation

PCA analysis was carried out on mean-centered, smoothed and baseline corrected data according to the preprocessing methodology above-reported. subjected to investigate the signal variation of various Raman shifts. Parameter estimation for the seven parameters was performed through Nelder-Mead algorithm (Nelder & Mead, 1965) through the minimization of a weighted cost function reported in eq. 3.9:

$$\min_{\underline{\theta}} \sum_{i=1}^{I} \left[\left(w_1 (S_1(\underline{\theta}, t_i) - S_{1,i}^{exp})\right)^2 + w_2 \left(S_2(\underline{\theta}, t_i) - S_{2,i}^{exp}\right)^2 + w_3 \left(G'(\underline{\theta}, t_i) - G'^{exp}_i\right)^2 \right] \quad (3.9)$$

Here, $\underline{\theta}$ represents the parameters' vector, while t_i denotes the i-th experimental sampling time out of a total of I samples for each experimental run. The vector $\underline{w} = [w_1, w_2, w_3]$ represents the weight values attributed to each term of the cost function and was arbitrarily set to [1,1,5]. A higher weight was attributed to the elastic modulus due to the lower noise in its measurements. The experimental values from two replicated runs were included in the calculation of the objective function. The calibration performance of the model was assessed using MSE and R^2 values.

A Monte Carlo simulation was employed to provide an estimation of the confidence intervals (Press et al., 2003). Briefly, the method involves the execution of regression algorithm multiple times to construct an empirical distribution of each parameter. For each iteration, Gaussian noise was introduced on all the experimental acquisitions for each measured variable, in order to assess

how the perturbation affects the parameters estimation. Specifically, the experimental variables were perturbated according to the following formula:

$$\underline{y}_k^{noise} = \underline{y}_k^{exp} + \sqrt{MSE_k}\,\underline{\varepsilon}, \qquad \underline{y}_k \in \mathbb{R}^I, \quad \forall\, k = 1,\ldots,K. \tag{3.10}$$

In Eq. 3.10, $\underline{y}_k^{exp} = \left(\underline{S}_{1,k}^{exp}, \underline{S}_{2,k}^{exp}, \underline{G'}_k^{exp}\right)$ refers to the experimental value of one of the three variables at the k-th experimental condition. MSE_k is the mean square error computed for the k-th experimental condition. Finally, $\boldsymbol{\varepsilon}$ is a vector of dimension I of random numbers drawn from a standard random variable, $\underline{\varepsilon} \sim \mathcal{N}\left(\underline{0}, \underline{\underline{I}}\right)$, where $\underline{\underline{I}}$ is the I × I identity matrix. The fitting process across resampled datasets with the addition of synthetic noise was repeated 1000 times. Confidence intervals were then calculated by determining the quantiles at α/2 and 1 − α/2 of the distributions of parameters estimates. The value of α was set equal to 0.05.

After estimating the parameters for each experimental condition, a stepwise regression analysis with interaction terms was performed to establish a relationship between the parameters and the operative conditions (i.e. T and C_r). The significance level for adding a term to the model was set at 0.05, whereas a threshold of 0.1 was used for removing a term. These linear relationships were subsequently employed to estimate the parameter values for an additional experimental condition at 37°C with a rennet concentration of 0.0218 IMCU mL^{-1} of milk which served as a test sample to validate the reliability of parameters estimation. The linear model employed for each parameter can be represented through Eq. 3.11:

$$\boldsymbol{\theta}(T, C_r) = \beta_0 + \beta_1 T + \beta_2 C_r + \beta_3 T\, C_r \tag{3.11}$$

3.2.3 Real time monitoring algorithm: Kalman filter

A state estimator has been developed for online monitoring of the elastic modulus G'. The $\underline{\underline{A}}$ matrix associated to Eq. 3.6 can be written as follows:

$$\underline{\underline{A}} = \begin{pmatrix} -\alpha_1 & 0 & 0 \\ n\alpha_1 & -\alpha_2 & 0 \\ 0 & m\,\alpha_2(S_2 - S_{1,\infty}) & 0 \end{pmatrix} \qquad \forall t > t^* \tag{3.12}$$

Given that the matrix that relates the state variables and measured output is $\underline{\underline{C}} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ it is possible to derive the observability matrix for the system presented in Eq. 3.6:

$$\underline{\underline{L_0}} = \begin{pmatrix} 1 & -\alpha_1 + n\alpha_1 & (-\alpha_1 + n\alpha_1)^2 \\ 1 & -\alpha_2\phi & (-\alpha_2\phi)^2 \\ 0 & 0 & 0 \end{pmatrix} \qquad \forall t > t^* \qquad (3.13)$$

Since $\mathbf{L_0}$ is rank deficient, the system is not fully observable (its observability matrix has a rank of two). However, it is possible to design of a reduced-order estimator in which the PC scores S_1 and S_2 are innovated, while G' (non-observable state) dynamics is not corrected (non-innovated) as described by Lisci et al. (2020). Therefore, the model and observability matrix can be rewritten as follows:

$$\underline{\underline{A}} = \begin{pmatrix} -\alpha_1 & 0 \\ n\,\alpha_1 & -\alpha_2 \end{pmatrix} \qquad \underline{\underline{L_0}} = \begin{pmatrix} 1 & -\alpha_1 + n\alpha_1 \\ 1 & -\alpha_2\phi \end{pmatrix} \qquad \forall t > t^* \qquad (3.14)$$

To enable real-time prediction of the elastic modulus G'(t) from Raman spectra, the ODE system presented in the kinetic model was coupled with the KF algorithm according to the following formulation (3.15):

$$\begin{cases} \frac{d\hat{S}_1}{dt} = -\alpha_1(\hat{S}_1 - S_{1,\infty}) + K_{11}^F(y_1^{exp} - \hat{S}_1) + K_{12}(S_2^{exp} - \hat{S}_2) \\ \frac{d\hat{S}_2}{dt} = n\,\alpha_1(\hat{S}_1 - S_{1,\infty}) - \alpha_2(\hat{S}_2 - S_{2,\infty})\phi(t - t^*) + K_{21}(S_1^{exp} - \hat{S}_1) + K_{22}^F(S_2^{exp} - \hat{S}_2) \\ \frac{dG'}{dt} = m\,\alpha_2(\hat{S}_2 - S_{2,\infty})\phi(t - t^*) \\ \underline{\underline{K}} = \underline{\underline{P}}\,\underline{\underline{C}}^T\,\underline{\underline{R}}^{-1} \\ \frac{d\underline{\underline{P}}}{dt} = \underline{\underline{P}}\,\underline{\underline{A}}^T + \underline{\underline{A}}\,\underline{\underline{P}}^T + \underline{\underline{Q}} - \underline{\underline{P}}\,\underline{\underline{C}}^T\,\underline{\underline{R}}^{-1}\underline{\underline{C}}\underline{\underline{P}} \end{cases} \qquad (3.15)$$

In Eq. 3.15, $\underline{\underline{K}}$ represents the Kalman filter gain, $\underline{\underline{P}}$ is the estimated state error covariance matrix, $\underline{\underline{Q}}$ is the model error covariance matrix, $\underline{\underline{R}}$ is the measurement covariance matrix and $\underline{\underline{C}}$ is the measurement matrix.

In this work, the calibration set experiments were used to tune the filter (i.e. estimate $\underline{\underline{Q}}$ and $\underline{\underline{R}}$) which were subsequently employed for the real-time prediction of G' from the test sample (i.e. T=37 °C and C_r=0.0218 IMCU mL^{-1}). The parameters derived from Eq. 3.11 were embedded within the Kalman filter during the recursive estimation.

3.2.4 Kinetic model calibration

The evolution of G' obtained from twelve rheological analysis along with S_1 and S_2 calculated from the Raman spectra analysis over time were experimentally determined to model the coagulation process. To this concern, figure 3.11a-c respectively show the behavior of S_1, S_2 and G' subjected to unitary normalization for the model fitting at T=36 °C and Cr=0.029 IMCU ml-1, while in figure 3.11d G' is mapped back to its original scale for a clearer visualization of milk rheological properties variation during renneting under the same experimental condition. Noteworthy, a monotonical increase was observed in the first principal component score, while a maximum value was detected in the second principal component score, in accordance with figure 3.8. The apparent inversion in the sign of principal components observed between figure 3.11a,b and 3.8d,e is a consequence of the inherent sign ambiguity of PCA. Since both the scores and the corresponding loadings can be multiplied by –1 without affecting the underlying variance structure or the orthonormality of the components, the direction of PCs was arbitrary set to comply with the kinetic model. As it can be observed from figure 3.11a, the exponential decay is satisfactory for capturing the S_1 decreasing resulting from renneting. This observation is consistent with a first order reaction for κ-casein hydrolysis reported in other studies (Mateo et al., 2009; Yang et al., 2022). An asymptotical behavior is detected for both S_1 and S_2 scores as a possible consequence of complete κ-casein hydrolysis and cross-linking, respectively. When S_1 is close to unity, in the early stage of reaction, S_2 increases due to the high values of the production term. Afterwards, a slowing down effect is observed due to S_1 decreasing, until a trend inversion for S_2, promoted by its consumption term, which becomes significant in the proximity of t^*. Such an inversion of S_2 curve is observed at 16 minutes. Concerning the elastic modulus, G' showed a lag phase during which a constant value of liquid milk G' (approximately 10^{-5} Pa) was recorded. After approximately 9.5 minutes, G' resulted to begin its increasing as a result of the onset of micelle aggregation (i.e. gelation time). These outcomes underline the importance of cross-linking sites concentration on the viscoelastic properties of gel.

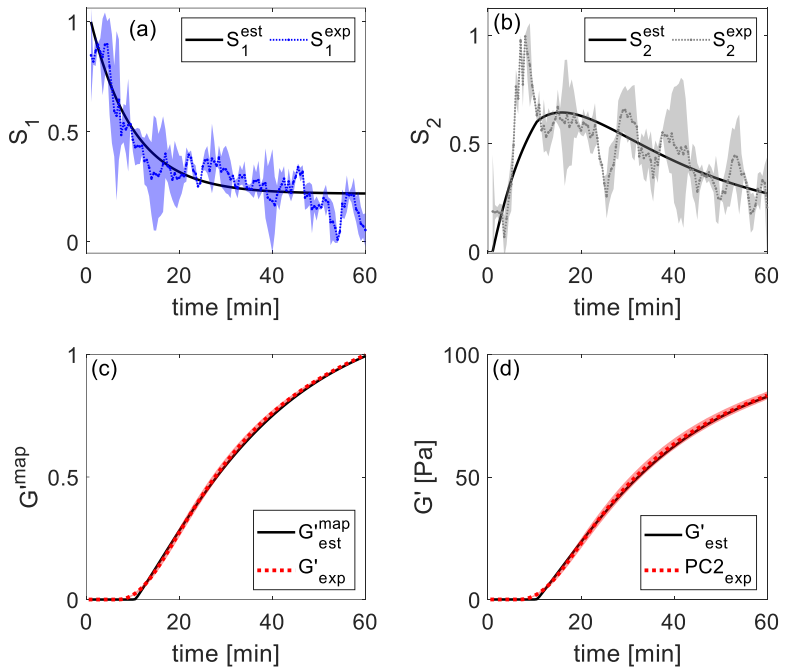


Figure 3.11 Experimental observation and model estimation of the dynamic evolution for all the three normalized variables: score values of the first PC S_1 (a) and second PC S_2 (b) obtained from applying PCA on Raman spectra and the Rheological elastic modulus mapped in a 0-1 space (c). Elastic modulus was also mapped back to its original scale (d). Such plots were obtained for T=36 °C and C_r=0.029 IMCU ml^{-1}. The average values of two replicates were used to display the standard deviation (shaded area).

Reproduced with permission from Springer Nature: Sibono, L., Tronci, S., Hedegaard, M.A.B., et al., PAT-driven dairy processing: a rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics, Analytical and Bioanalytical Chemistry, 417, 5691–5702 (2025). DOI: 10.1007/s00216-025-05965-2

The values of fitting parameters obtained from fitting the model on the experimental scores and elastic moduli are reported in figure 3.12. Such model coefficients were obtained from the application of a Monte Carlo simulation based on 1000 iteration. The pseudo-kinetic parameters estimation showed different behaviors under diverse experimental conditions. The rate constant α_1 showed a slight increase with temperature under high rennet concentrations, while no evident variation was observed for the lower amount of enzyme (Figure 3.12a). Such an outcome is probably due to the high noise in S_1 and S_2 observed after performing PCA on Raman spectra. On the other hand, α_2 monotonically increased with temperature for both higher and lower level of rennet concentration (Figure 3.12b), confirming that the kinetics are favored at higher temperatures. For the same reason, the characteristic time t^* resulted to reasonably decrease

when increasing both temperature and rennet concentration (Figure 3.12c). The estimated values of t^* resulted to be numerically close to the gel clotting time. Finally, no clear pattern was observed in the model coefficients n and m, other than a slight increase for higher rennet concentrations.

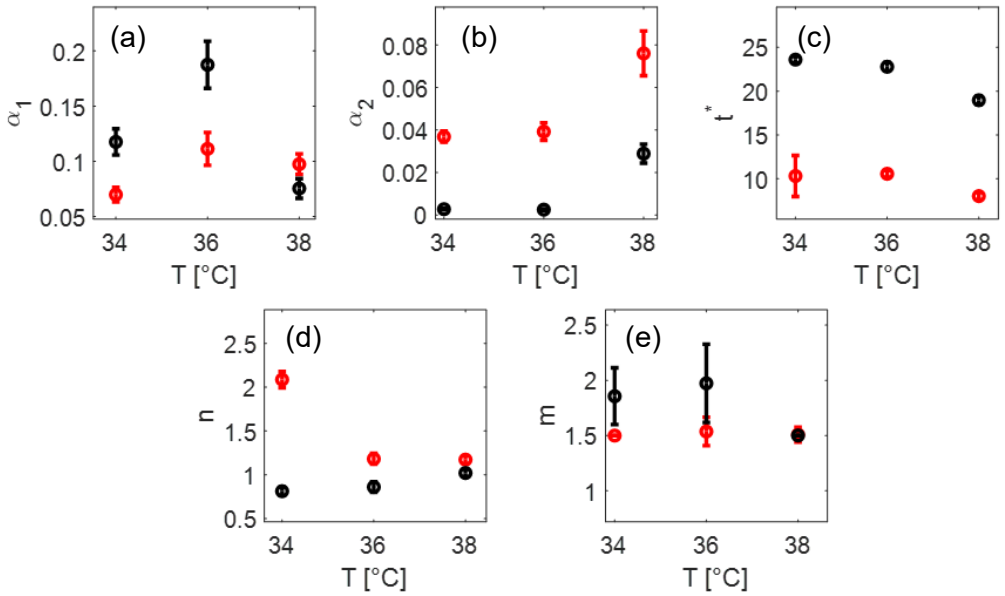


Figure 3.12 Parameter estimation for each experimental run and the relevant error bars resulting from fitting S_1, S and G' to the corresponding experimental data. Red dots denote the parameters obtained from Cr = 0.029 IMCU ml^{-1} of milk, while black dots indicate those from Cr = 0.0145 IMCU ml^{-1} of milk.

Reproduced with permission from Springer Nature: Sibono, L., Tronci, S., Hedegaard, M.A.B., et al., PAT-driven dairy processing: a rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics, Analytical and Bioanalytical Chemistry, 417, 5691–5702 (2025). DOI: 10.1007/s00216-025-05965-2

Table 3.2 shows the performance metrics obtained from the model presented in this work. High values $R^2_{G'}$ were obtained from the regression, suggesting good model performances for all the runs. Similar results were obtained in a previous study, where VIS-NIR was used to predict the elastic modulus dynamics when varying protein and calcium concentrations (Villaquiran et al., 2024).

Table 3.2 Parameters estimation and performance metrics obtained by fitting the model reported in Eq. 3.8 to experimental data for each experimental condition. Both replications were used for regression purposes.

Reproduced with permission from Springer Nature: Sibono, L., Tronci, S., Hedegaard, M.A.B., et al., PAT-driven dairy processing: a

rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics, Analytical and Bioanalytical Chemistry, 417, 5691–5702 (2025). DOI: 10.1007/s00216-025-05965-2

T [°C]	C_r [IMCU ml^{-1}]	$MSE_{G'}$	$R^2_{G'}$
36	0.029	1.12	0.9987
38	0.0145	1.45	0.9981
34	0.029	0.39	0.9994
36	0.0145	1.16	0.9971
34	0.0145	0.58	0.9979
38	0.029	2.53	0.9974

Stepwise linear regression was performed to estimate the relationships between the coagulation model parameters and the operative conditions (i.e., temperature and rennet concentration). Table 3.3 shows a summary of stepwise regression, displaying the coefficient values and the significance of the linear equations. No interaction term resulted to be significant after the analysis.

Table 3.3 Results of stepwise regression analysis. The parameters were expressed as linear functions of temperature and rennet concentrations but only significant effects were retained.

Coefficient	Model	Square error			p value			R^2
		β_0	β_1	β_2	β_0	β_1	β_2	
α_1 [min^{-1}]	0.110	0.017			0.002			
α_2 [min^{-1}]	−0.32+0.008 T+2.712 C_r	0.099	0.003	0.61	0.048	0.058	0.021	0.905
t^* [min]	65.11−0.866 T− 837.9C_r	10.750	0.30	66.56	0.009	0.061	0.001	0.986
n [-]	1.189	0.190			0.002			
m [-]	1.646	0.087			7.4 10^{-6}			

3.2.5 Kinetic model validation

Given the linear equations reported in Table 3.3, parameters can be estimated for the testing condition (T=37 °C and C_r=0.0218 IMCU mL^{-1} of milk). Under this condition, it is possible to derive out that α_1 = 0.11 min^{-1}, α_2=0.039 min^{-1}, t^*=14.80 min, n=1.189 and m=1.646. As well, the test experiment presents a value of t^* for activating G' generation term that is numerically close to the corresponding gel clotting time (t_{gel}=13 min), denoting a strong consistency between the model prediction and the experimental determination of rheological properties. The integration of the ODE showed in the present coagulation model using these parameters values allowed for the estimation of transient behavior of S_1, S_2 and G' for the test experiment. Upon inspecting figure 3.13a, it can be observed that the kinetic model presented in this study well describes the time evolution of the variables under investigation despite S_1 and S_2 are highly scattered. The analysis of the test sample resulted in a R^2_{S1}=0.843, R^2_{S2}=0.500, and $R^2_{G'}$=0.997, indicating a reliable prediction

of the elastic modulus. Lower performances in S_1 and S_2 prediction are due to the high scattering attributed to such variables. However, the overall dynamics of score values are satisfactorily explained by the model, along with excellent prediction of elastic modulus (Figure 3.13b).

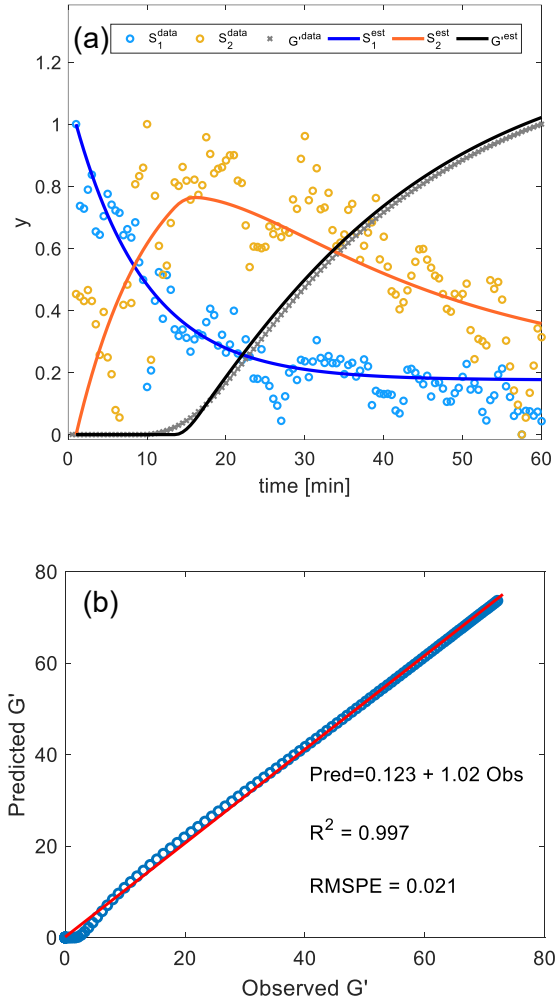


Figure 3.13 Time profile of S_1, S_2 and G' reconstructed from the integration of kinetic model for milk rennet coagulation reported in Eq. 3.8 (a) and predicted vs observed plot (b). The symbol y in figure 3.13a refers to the generic response variable. The experimental acquisitions were retrieved from the coagulation of reconstituted milk powder at a temperature of 37 °C and a rennet concentration of 0.0218 IMCU mL^{-1} of milk.

3.2.6 Real time prediction of Elastic Modulus

The performance of the KF is illustrated in figures 3.14a and 3.14b which compare experimental data for S_1, S_2 and G' against KF estimates. To assess the robustness of the KF, modeling errors were introduced into the system: parameter α_1 was adjusted by 50% relative to its reference value, and α_2 by 10%. Notably, larger errors in α_2 pose a significant challenge to the state estimator. The results highlight KF's effectiveness, showing that it substantially enhances state estimation accuracy compared to relying solely on the model when parameters are affected by error. Furthermore, the KF successfully filters the two observable states, mitigating the effects of noise and modeling errors. Indeed an R^2 of prediction for G' resulted to be equal to 0.995. This means that the presented Raman-based inline system can successfully provide a reliable estimation of system's state.

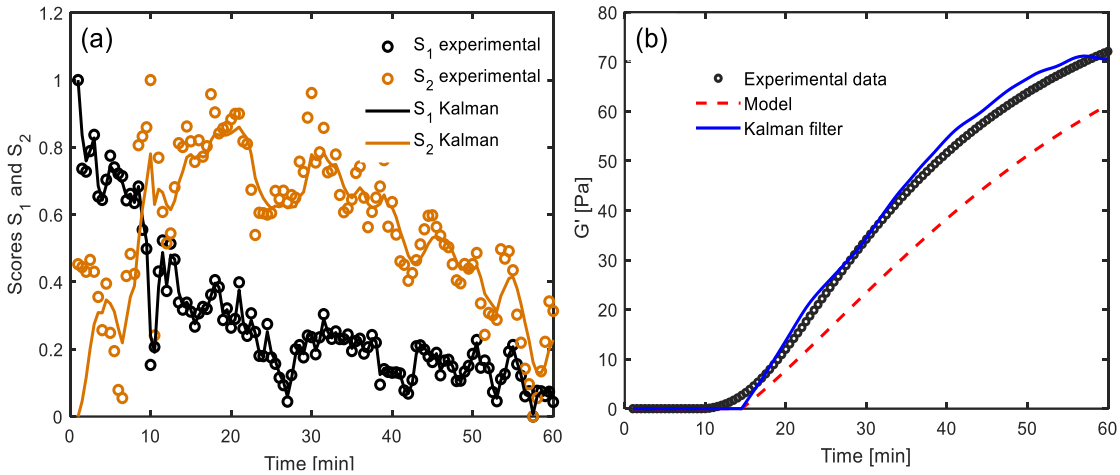


Figure 3.14 Application of the kinetic model and Kalman Filter to the real-time monitoring problem: comparison between the experimental data and Kalman filter for S_1 and S_2 (a) along with the real-time G' estimates from the filter and the model compared with the experimental data (b).

Reproduced with permission from Springer Nature: Sibono, L., Tronci, S., Hedegaard, M.A.B., et al., PAT-driven dairy processing: a rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics, Analytical and Bioanalytical Chemistry, 417, 5691–5702 (2025). DOI: 10.1007/s00216-025-05965-2

3.3 Conclusions

Spectroscopic signal baseline variation is a reliable information for curd cutting time prediction. First, the evolution of rheological properties during coagulation was used as a reference analytical method for the determination of the real cutting time. Time at the inflection point of PC1 scores obtained from the spectra matrix was employed to calibrate a linear equation for estimating the cutting time. Results showed good model performances in both calibration and validation, thus proving the robustness of Raman information to predict the curd cutting time. Afterwards, baseline corrected Raman signals were analyzed for the identification of significant wavenumbers through PCA loadings assessment. Some compounds already recognized as coagulation tracers, like tryptophan, were identified as significant from the analysis, while others like aspartic and glutamic acid were first individuated in this study. The identification of the biomarkers here presented is supported by studies that have analyzed these compounds individually. However, Raman spectroscopy enables a simultaneous analysis of these molecular bonds of interest, offering a more comprehensive approach. Two different behaviors of dynamic variation of system composition were revealed by the first two PCs, demonstrating a high complexity of milk compounds redistribution during its coagulation. Therefore, it has been shown that Raman spectroscopy can be effectively employed for various purposes in investigating the dynamic behavior of enzymatic milk coagulation. The application of baseline correction enables the identification of molecular bonds that exhibit changes during the hydrolysis process, offering an additional benefit over previously studied methodologies. Afterwards, such spectroscopical information was linked to a coagulation technological parameter (i.e. elastic modulus) by means of a kinetic model. Specifically, a system of ODEs was embedded within a Kalman Filter to provide a real time estimation of the PC scores and G', thus providing a thorough description of milk constituents evolution along renneting, which has not been achieved by the consolidated traditional methods. However, a few limitations of the methodology reported herein must be acknowledged. First, further validation and replication of these results are necessary to establish the method's reliability using different source of variability (e.g. milk thermal treatment, protein and fat concentration etc.). Moreover, the spectroscopic data, despite showing consistency with other studies, are affected by noisy signal, which can compromise the robustness of prediction when extending the operational variability. To this concern, future research can explore the integration of Raman spectroscopy in data fusion frameworks, where the combination of multiple analytical methods can further extend the understanding of the process and increase the robustness of prediction.

3.4 References

Abdelgawad, A. R., Guamis, B., & Castillo, M. (2014). Using a fiber optic sensor for cutting time prediction in cheese manufacture from a mixture of cow, sheep and goat milk. Journal of Food Engineering, 125(1), 157–168. https://doi.org/10.1016/j.jfoodeng.2013.10.001

Arango, O., & Castillo, M. (2018). A method for the inline measurement of milk gel firmness using an optical sensor. Journal of Dairy Science, 101(5), 3910–3917. https://doi.org/10.3168/jds.2017-13595

Banks, J. M., & Horne, D. S. (2003). CHEESES | Chemistry of Gel Formation. Encyclopedia of Food Sciences and Nutrition, 1056–1062. https://doi.org/10.1016/b0-12-227055-x/00195-4

Brereton, R. G. (2009). Chemometrics for Pattern Recognition. In Chemometrics for Pattern Recognition. Wiley.

Britten, M., & Giroux, H. J. (2022). Rennet coagulation of heated milk: A review. International Dairy Journal, 124. https://doi.org/10.1016/j.idairyj.2021.105179

Carlson, A., Hill, C. G., & Olson, N. F. (1987). The kinetics of milk coagulation: IV. The kinetics of the gel-firming process. Biotechnology and Bioengineering, 29(5), 612–624. https://doi.org/10.1002/bit.260290510

Castillo, M. (2006). Cutting time prediction methods in cheese making. Encyclopedia of Agricultural, Food, and Biological Engineering, 1, 1–7.

Castillo, M., Lucey, J. A., & Payne, F. A. (2006). The effect of temperature and inoculum concentration on rheological and light scatter properties of milk coagulated by a combination of bacterial fermentation and chymosin. Cottage cheese-type gels. International Dairy Journal, 16(2), 131–146. https://doi.org/10.1016/j.idairyj.2005.02.005

Chitpinityol, S., & Crabbe, M. J. C. (1998). Chymosin and aspartic proteinases. Food Chemistry, 61(4), 395–418. https://doi.org/10.1016/S0308-8146(97)00090-3

de Almeida, M. R., de Sá Oliveira, K., Stephani, R., & Cappa de Oliveira, L. F. (2012). Application of FT-Raman Spectroscopy and Chemometric Analysis for Determination of Adulteration in Milk Powder. Analytical Letters, 45(17), 2589–2602. https://doi.org/10.1080/00032719.2012.698672

De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. International Journal of Pharmaceutics, 417(1–2), 32–47. https://doi.org/10.1016/j.ijpharm.2010.12.012

Derra, M., Bakkali, F., Amghar, A., & Sahsah, H. (2018). Estimation of coagulation time in cheese manufacture using an ultrasonic pulse-echo technique. Journal of Food Engineering, 216, 65–71. https://doi.org/10.1016/j.jfoodeng.2017.08.003

Dufour, É. (2011). Recent advances in the analysis of dairy product quality using methods based on the interactions of light with matter. International Journal of Dairy Technology, 64(2), 153–165. https://doi.org/10.1111/j.1471-0307.2010.00665.x

Eilers, P. H. C. (2003). A perfect smoother. Analytical Chemistry, 75(14), 3631–3636. https://doi.org/10.1021/ac034173t

El-Abassy, R. M., Eravuchira, P. J., Donfack, P., Von Der Kammer, B., & Materny, A. (2011). Fast determination of milk fat content using Raman spectroscopy. Vibrational Spectroscopy, 56(1), 3–8. https://doi.org/10.1016/j.vibspec.2010.07.001

Esteves, C. L. C., Lucey, J. A., & Pires, E. M. V. (2002). Rheological properties of milk gels made with coagulants of plant origin and chymosin. International Dairy Journal, 12(5), 427–434. https://doi.org/10.1016/S0958-6946(01)00170-4

Fagan, C. C., Ferreira, T. G., Payne, F. A., O'Donnell, C. P., O'Callaghan, D. J., & Castillo, M. (2011). Preliminary evaluation of endogenous milk fluorophores as tracer molecules for curd syneresis. Journal of Dairy Science, 94(11), 5350–5358. https://doi.org/10.3168/jds.2011-4399

Fagan, C. C., Leedy, M., Castillo, M., Payne, F. A., O'Donnell, C. P., & O'Callaghan, D. J. (2007). Development of a light scatter sensor technology for on-line monitoring of milk coagulation and whey separation. Journal of Food Engineering, 83(1), 61–67. https://doi.org/10.1016/j.jfoodeng.2006.12.014

Felfoul, I., Bouazizi, A., Attia, H., & Karoui, R. (2022). Monitoring of acid-induced coagulation of dromedary and cows' milk by untargeted and targeted techniques. International Dairy Journal, 127. https://doi.org/10.1016/j.idairyj.2021.105300

Gallier, S., Gordon, K. C., Jiménez-Flores, R., & Everett, D. W. (2011). Composition of bovine milk fat globules by confocal Raman microscopy. International Dairy Journal, 21(6), 402–412. https://doi.org/10.1016/j.idairyj.2011.01.008

Grassi, S., Strani, L., Alamprese, C., Pricca, N., Casiraghi, E., & Cabassi, G. (2022). A FT-NIR process analytical technology approach for milk renneting control. Foods, 11(1). https://doi.org/10.3390/foods11010033

Hashizume, K., & Sato, T. (1988). Gel-Forming Characteristics of Milk Proteins. 1. Effect of Heat Treatment. Journal of Dairy Science, 71(6), 1439–1446. https://doi.org/10.3168/jds.S0022-0302(88)79706-4

Hayes, E., Greene, D., O'Donnell, C., O'Shea, N., & Fenelon, M. A. (2023). Spectroscopic technologies and data fusion: Applications for the dairy industry. Frontiers in Nutrition, 9. https://doi.org/10.3389/fnut.2022.1074688

He, H., Sun, D. W., Pu, H., Chen, L., & Lin, L. (2019). Applications of Raman spectroscopic techniques for quality and safety evaluation of milk: A review of recent developments. Critical Reviews in Food Science and Nutrition, 59(5), 770–793. https://doi.org/10.1080/10408398.2018.1528436

Khan, H. M. H., McCarthy, U., Esmonde-White, K., Casey, I., & O'Shea, N. (2023). Potential of Raman spectroscopy for in-line measurement of raw milk composition. Food Control, 152. https://doi.org/10.1016/j.foodcont.2023.109862

Lamothe, S., & Britten, M. (2023). Coagulation and proteolysis of high-protein milks in the gastric environment. Journal of Dairy Science, 106(8), 5242–5252. https://doi.org/10.3168/jds.2022-22714

Lisci, S., Grosso, M., & Tronci, S. (2020). A geometric observer-assisted approach to tailor state estimation in a bioreactor for ethanol production. Processes, 8(4). https://doi.org/10.3390/PR8040480

Lyndgaard, C. B., Engelsen, S. B., & Van Den Berg, F. W. J. (2012). Real-time modeling of milk coagulation using in-line near infrared spectroscopy. Journal of Food Engineering, 108(2), 345–352. https://doi.org/10.1016/j.jfoodeng.2011.07.029

Mateo, M. J., O'Callaghan, D. J., Everard, C. D., Fagan, C. C., Castillo, M., Payne, F. A., & O'Donnell, C. P. (2009). Influence of curd cutting programme and stirring speed on the prediction of syneresis indices in cheese-making using NIR light backscatter. Lwt, 42(5), 950–955. https://doi.org/10.1016/j.lwt.2008.12.014

Mendes, T. O., Junqueira, G. M. A., Porto, B. L. S., Brito, C. D., Sato, F., de Oliveira, M. A. L., Anjos, V., & Bell, M. J. V. (2016). Vibrational spectroscopy for milk fat quantification: line shape analysis of the Raman and infrared spectra. Journal of Raman Spectroscopy, 47(6), 692–698. https://doi.org/10.1002/jrs.4878

Miloradovic, Z., Kljajevic, N., Miocinovic, J., Levic, S., Pavlovic, V. B., Blažić, M., & Pudja, P. (2020). Rheology and microstructures of rennet gels from differently heated goat milk. Foods, 9(3). https://doi.org/10.3390/foods9030283

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. The Computer Journal, 7(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nicolau, N., Buffa, M., O'Callaghan, D. J., Guamis, B., & Castillo, M. (2015). Estimation of clotting and cutting times in sheep cheese manufacture using NIR light backscatter. Dairy Science and Technology, 95(4), 495–507. https://doi.org/10.1007/s13594-015-0232-7

O'Callaghan, D. J., O'Donnell, C. P., & Payne, F. A. (1999). Comparison of on-line techniques for determination of curd setting time using cheesemilks under different rates of coagulation. Journal of Food Engineering, 41(1), 43–54. https://doi.org/10.1016/S0260-8774(99)00072-2

O'Callaghan, D. J., O'Donnell, C. P., & Payne, F. A. (2002). Review of systems for monitoring curd setting during cheesemaking. International Journal of Dairy Technology, 55(2), 65–74. https://doi.org/10.1046/j.1471-0307.2002.00043.x

Panikuttira, B., Payne, F. A., O'Shea, N., Tobin, J. T., O'Callaghan, D. J., & O'Donnell, C. P. (2020). Investigation of an in-line prototype fluorescence and infrared backscatter sensor to monitor rennet-induced coagulation of skim milk at different protein concentrations. International Journal of Food Science and Technology, 55(1), 175–182. https://doi.org/10.1111/ijfs.14267

Panikuttira, B., Payne, F. A., O'Shea, N., Tobin, J. T., & O'Donnell, C. P. (2019). Evaluation of a fluorescence and infrared backscatter sensor to monitor acid induced coagulation of skim milk.

Innovative Food Science and Emerging Technologies, 54, 219–224. https://doi.org/10.1016/j.ifset.2019.04.011

Panthi, R. R., Shibu, S. N., Ochalski, T. J., & O'Mahony, J. A. (2023). Raman spectra of micellar casein powders prepared with wet blending of glycomacropeptide and micellar casein concentrate. International Journal of Dairy Technology, 76(2), 429–435. https://doi.org/10.1111/1471-0307.12920

Payens, T. A. (1989). The enzyme-triggered coagulation of casein micelles. Advances in Colloid and Interface Science, 30(C), 31–69. https://doi.org/10.1016/0001-8686(89)80003-X

Pratt, W. K. (2007). Digital Image Processing (J. W. & Sons (ed.); 4th ed.).

Press, W. H., Teukolsky, S. A., Vettering, W. T., & Flannery, B. P. (2003). Numerical Recipes in C++: The Art of Scientific Computing (2nd edn) 1 Numerical Recipes Example Book (C++) (2nd edn) 2 Numerical Recipes Multi-Language Code CD ROM with LINUX or UNIX Single-Screen License Revised Version 3 . European Journal of Physics, 24(3), 329–330. https://doi.org/10.1088/0143-0807/24/3/701

Puerta, F., Wessel, W., García, V., López, E., Rovira, S., Alvarez, D., Roa, I., & López, M. B. (2017). A preliminary study of the use of a Raman laser sensor to monitor coagulation and syneresis for the online control of cheesemaking. International Journal of Dairy Technology, 70(4), 607–609. https://doi.org/10.1111/1471-0307.12439

Sandra, S., & Dalgleish, D. G. (2005). Effects of ultra-high-pressure homogenization and heating on structural properties of casein micelles in reconstituted skim milk powder. International Dairy Journal, 15(11), 1095–1104. https://doi.org/10.1016/j.idairyj.2004.11.015

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. (2025). Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy. Applied Food Research, 5(1). https://doi.org/10.1016/j.afres.2025.100701

Stocco, G., Cipolat-Gotet, C., Cecchinato, A., Calamari, L., & Bittante, G. (2015). Milk skimming, heating, acidification, lysozyme, and rennet affect the pattern, repeatability, and predictability of milk coagulation properties and of curd-firming model parameters: A case study of Grana Padano. Journal of Dairy Science, 98(8), 5052–5067. https://doi.org/10.3168/jds.2014-9146

Strani, L., Grassi, S., Alamprese, C., Casiraghi, E., Ghiglietti, R., Locci, F., Pricca, N., & De Juan, A. (2021). Effect of physicochemical factors and use of milk powder on milk rennet-coagulation: Process understanding by near infrared spectroscopy and chemometrics. Food Control, 119(April 2020). https://doi.org/10.1016/j.foodcont.2020.107494

Svendsen, C., Cieplak, T., & Van Den Berg, F. W. J. (2016). Exploring process dynamics by near infrared spectroscopy in lactic fermentations. Journal of Near Infrared Spectroscopy, 24(5), 443–451. https://doi.org/10.1255/jnirs.1244

Van Hooydonk, A. C. ., & Walstra, P. (1987). Interpretation of the kinetics of the renneting reaction in milk. Netherlands Milk Dairy Journal, 41, 19–47.

Villaquiran, Z., Zamora, A., Arango, O., & Castillo, M. (2024). Inline Determination of the Gel Elastic Modulus During Milk Coagulation Using a Multifiber Optical Probe. Food and Bioprocess Technology. https://doi.org/10.1007/s11947-023-03294-9

Walther, A. R., Ditzel, N., Kassem, M., Andersen, M. Ø., & Hedegaard, M. A. B. (2022). In vivo non-invasive monitoring of tissue development in 3D printed subcutaneous bone scaffolds using fibre-optic Raman spectroscopy. Biomaterials and Biosystems, 7. https://doi.org/10.1016/j.bbiosy.2022.100059

Yang, M., Ye, A., Yang, Z., Everett, D. W., Gilbert, E. P., & Singh, H. (2022). Kinetics of pepsin-induced hydrolysis and the coagulation of milk proteins. Journal of Dairy Science, 105(2), 990–1003. https://doi.org/10.3168/jds.2021-21177

Zhang, Y., Li, Y., Wang, P., Liang, Q., Zhang, Y., & Ren, F. (2019). The factors influencing rennet-induced coagulation properties of yak milk: The importance of micellar calcium during gelation. Lwt, 111, 500–505. https://doi.org/10.1016/j.lwt.2019.05.015

Zhang, Z. Y., Shi, X. J., Zhao, Y. J., Zhang, Y. S., & Wang, H. Y. (2022). Brand Identification of Soybean Milk Powder based on Raman Spectroscopy Combined with Random Forest Algorithm. Journal of Analytical Chemistry, 77(10), 1282–1286. https://doi.org/10.1134/S1061934822100173

Zhao, M., Markiewicz-Keszycka, M., Beattie, R. J., Casado-Gavalda, M. P., Cama-Moncunill, X., O'Donnell, C. P., Cullen, P. J., & Sullivan, C. (2020). Quantification of calcium in infant formula using laser-induced breakdown spectroscopy (LIBS), Fourier transform mid-infrared (FT-IR) and Raman

spectroscopy combined with chemometrics including data fusion. Food Chemistry, 320. https://doi.org/10.1016/j.foodchem.2020.126639

Chapter 4: Raman spectroscopy as screening tool for detecting anomalies during rennet coagulation of milk

The analytical techniques presented in the previous chapter are here proposed for fault detection purposes. Typical anomalies encountered in food industry are considered and identified using multivariate statistical process control strategies, aiming at providing a wider perspective of process monitoring within the dairy framework.

4.1 The problem of fault occurrence within the dairy environment: goal of the study

In the context of dairy processes like cheesemaking, final product's quality is highly sensitive to process variations in the manufacturing line, where even minor deviations can significantly affect the texture, flavor, and safety of the final product (Castillo et al., 2006). Particularly, several critical quality attributes of cheese (e.g. fat and whey content, cheese yield, and firmness) are heavily dependent on how milk is processed along the production line. The adoption of PAT philosophy for continuous monitoring and control of CPP in dairy manufacturing holds the benefits to significantly reduce the production of off-spec products, while simultaneously enhancing process efficiency, productivity, and overall profitability (Tajammal Munir et al., 2015).

Therefore, every stage of manufacturing need the application of monitoring strategies that enable the identification of deviations of CPP from the so-called "golden batch" trajectory, that is the set of optimal process conditions known to consistently produce in-specification product (Duran-Villalobos et al., 2020). Ensuring product quality requires more than simply adhering to a final product quality perspective; it is equally essential to maintain CPPs within their optimal ranges at each individual stage of production. In the specific case of milk coagulation, the improper start-up and conduction of the process can result in several defects, including undesirable moisture levels in the curd, which subsequently compromise the ripening process (Castillo, 2006). The phase transition occurring during milk coagulation represents a particularly critical step, where the CPP role is conventionally covered up by viscoelastic properties measured by means of intrusive techniques such as manual operator's assessment, lactodynamography or rheological measurements (Derra et al., 2018; Pretto et al., 2011; Villaquiran et al., 2024). To date, process monitoring of milk enzymatic coagulation has been extensively investigated using NIR spectroscopy (Grassi et al., 2022; Lyndgaard et al., 2012; Strani et al., 2021). In the study reported by Grassi et al. (2022), the authors proposed a novel application of Multivariate Statistical Process Control (MSPC) to milk coagulation. However, achieving comprehensive fault diagnosis (i.e. identifying the source of a failure) can be challenging when relying on NIR spectroscopy, due to its high sensitivity to water content (De Beer et al., 2011). Despite the current research in this field is largely focused on the identification of coagulation time and optimal curd cutting time, which are technical parameters of high relevance in the dairy industry, low attention has been paid to how process monitoring can support corrective actions. Particularly, the evaluation of how early a fault is detected can offer timely and effective information for prompt intervention. Indeed, it is well

known that the phase transition induced by casein coagulation occurs only after a certain lag time following enzyme addition, due to the enzymatic hydrolysis step (Nicolau et al., 2015; Yang et al., 2022). At this early stage, the system allows corrective actions, such as temperature adjustment, enzyme concentration correction, or pH modulation, to be effectively implemented before the occurrence of the irreversible colloid destabilization. This means that the ability to detect process faults prior to the onset of the phase transition during milk coagulation represents a critical advantage for process control in dairy production. In this regard, Raman spectroscopy benefits from a higher sensitivity to organic molecular structures, opening up the path for multiple applications such as process monitoring and chemical characterization (Hayes et al., 2023; Sibono et al., 2025a). The occurrence of deviation from Normal Operating Conditions (NOC) is here simulated by mimicking two types of faults: reduction of enzyme concentration and temperature control shutdown. To this regard, the present investigation intends to address the following tasks:

1) Provide a critical analysis of global and local MSPC algorithms to identify the optimal fault detection configuration for correctly discriminating NOC experiments from the faulty ones.
2) Carry out a fault diagnosis procedure that allows to identify the specific source of a fault.
3) To assess whether the detection time allows for corrective action by comparing the alarm time with the time of the onset of coagulation.

This methodology aims to establish the basis for a time-efficient fault detection and diagnosis strategy that can be further exploited to keep the CPP inside the optimal operational window, with the potential to reduce product losses and improve the consistency of the final product.

4.1.1 Materials and methods

4.1.1.1 Sample preparation and experimental design

Reconstituted skimmed milk powder was prepared according to the procedure reported in the previous chapter. Here, the NOC experiments set consists of three temperature levels (34 °C, 36 °C, 38 °C) and a rennet concentration of 0.029 IMCU mL^{-1} of milk. The experiments were quadruplicated, yielding to 12 NOC assays. Five additional experiments were conducted to simulate faulty conditions. Four of these were carried out at 34 °C (in duplicate), 36 °C, and 38 °C, each with only 0.007 IMCU mL^{-1} of milk. Such experiments were denoted as rennet concentration (C_r) faults. The fifth fault experiment was conducted at an initial temperature of 38°C after which the water bath temperature was abruptly reduced to room temperature, counting as a temperature control system failure. Table 4.1 summarizes the experimental design for the coagulation assays.

Table 4.1 Operative conditions for the milk coagulation batches explored in this work.

Label	Temperature [°C]	Rennet concentration [IMCU/mL of milk]	Replications	Operative category
NOC_1	34	0.029	4	NOC
NOC_2	36	0.029	4	NOC
NOC_3	38	0.029	4	NOC
F1	34	0.007	2	Fault
F2	36	0.007	1	Fault
F3	38	0.007	1	Fault
F4	T_0=38 °C	0.0145	1	Fault

4.1.1.2 Software

Every computation shown in this work was performed using Matlab®2023a. The fault detection algorithms and model validation were carried out through Matlab codes developed by the authors. Here follows the description of the algorithms and data preprocessing procedures.

4.1.1.3 Preprocessing

The spectral acquisitions were assembled into a three-way data structure (N×P×K), where N is the number of acquisitions per experiment (119), P is the number of Raman shifts (953) and K is the number of experiments (18). Data were preprocessed through the Savitzky-Golay filter to smooth the data by fitting successive windows of data points to a polynomial of a specified degree while preserving the original shape of the signal (Schmid et al., 2022). A second-order polynomial and a frame length of 15 were used as this preprocessing gave sufficient smoothing without removing the relevant information (Sibono et al., 2025b). Moreover, spectra were subjected to a one-dimensional 3-order median filtering along renneting times to dampen the effect of eventual outliers among the score values from baseline corrected data (Pratt, 2007), thus removing cosmic spikes.

4.1.1.4 Fault detection algorithms

In this study, MSPC principles are applied to a batch process, which differs from fault analysis in continuous processes, as it requires focusing on a desired dynamic behavior over time. In this study, PCA-based modeling was employed to extract relevant patterns from the Raman spectral dataset. Only one principal component was selected for calibrating the PCA model, since no apparent pattern was obtained from the exploratory analysis of higher-dimensional PCs. The model is represented through eq. 3.1. Six algorithms were employed for the assessment of NOC and fault experiments:

1) PC scores confidence interval (SCI) algorithm.
2) PC scores-Non-linear-regression-model-based fault detection.
3) T^2-based fault detection.
4) Squared Prediction Error (SPE)-based fault detection algorithm.
5) SPE or T^2-based algorithm.
6) SPE and T^2-based algorithm

For all the algorithms, the fault detection quality was assessed through Fault Discovery Rate (FDR), False Alarm rate (FAR) and accuracy. FDR, FAR and accuracy were computed by means of a cross-validation scheme. Specifically, the calibration step involved the random selection of nine out of the twelve NOC samples, whereas the remaining three NOC samples were reserved as validation experiments for FAR assessment and five fault samples were used for FDR calculation. Before performing PCA decomposition, the three-way data tensor made of calibration experiments was

unfolded to form a (P × N K_{cal}) dimensional structure. Here, K_{cal} denotes the number of calibration experiments. Subsequently, the specific fault detection algorithms described below were applied. This procedure was repeated 100 times out of which the average accuracy, FAR and FDR were calculated. The validation protocol was used to tune the hyperparameters (i.e. parameters that are set before training and adjusted to improve fault detection performance). Unlike the conventional model parameters, which are directly derived from data, hyperparameters selection is driven by the model's accuracy.

Algorithm 1 and 2 belong to the *global-PCA* category, since the model and the confidence limits are built on all the N acquisitions associated with the calibration samples (Overall, N K_{cal}). Conversely, the algorithms 3-6 are associated to *local-PCA*-based models, wherein the model statistics evolve over time. This results in fluctuating control limits that dynamically adapt to the variability inherent in each individual data acquisition for the calibration samples. Among the investigated methodologies, only two methods are here described, as they outperformed the others in terms of classification accuracy. These include a global PCA-based model (PC scores-Non-linear-regression-model-based fault detection) and a local PCA-based model (SPE or T^2-based algorithm). A detailed explanation of algorithms 1,3,4,6 is reported in Appendix C, Section C.1.

4.1.1.5 Global PCA algorithm

In PC scores-Non-linear-regression-model-based fault detection, a PCA model is trained using NOC calibration samples and employed to compute the loading values. PC1 scores (denoted as S_1) are modelled through a Hill-type sigmoidal equation (Ford Versypt et al., 2017):

$$\hat{S}_1(t) = B + \frac{L}{1+\left(\frac{C}{t}\right)^D} \tag{4.1}$$

In Eq. 4.1, B is the offset parameter, L is a parameter such that $B + L$ corresponds to the asymptotic value, C is a displacement parameter along the x-axis, D is the Hill coefficient which defines the steepness of the function. For each iteration, after estimating the model's parameters, acquisitions from NOC and fault validation samples were sequentially projected into the PC space and compared to the model's Upper Control Limit (UCL) and Lower Control Limit (LCL) defined by the prediction confidence interval at the relevant acquisition time (Lane & Dumouchel, 1994). The sample is then classified as fault if all projected values in a preselected time window exceeded the UCL or fell below the LCL. In this algorithm the prediction confidence level (1-α) and the number of

consecutive points falling outside the NOC region required to trigger the alarm were considered as the hyperparameters of interest.

4.1.1.6 Local PCA algorithms

SPE or T^2-based algorithm is the local counterpart of the investigated MSPC model. The first PC was used for dimensionality reduction, and the T^2 statistic was calculated to evaluate deviations from the normal operational state (Alcala & Joe Qin, 2011):

$$T_i^2 = t_i \lambda_1^{-1} t_i^T \tag{4.2}$$

Where t_i is the PC1 score value for the i-th acquisition and λ_1 is the relevant eigenvalue.

The UCL for the T^2 statistic were computed using a χ^2-distribution-based threshold (Reis et al., 2021):

$$T_{lim}^2 = \frac{\sigma_{T^2}^2}{2\mu_{T^2}} \chi_\alpha^2 \left(\frac{2\mu_{T^2}^2}{\sigma_{T^2}^2} \right) \tag{4.3}$$

Where μ_{T^2} and $\sigma_{T^2}^2$ are the sample mean and variance for T^2 obtained from each individual acquisition of the calibration set. This implies a dynamic evolution of the limit statistics. Concurrently, SPE statistic was also computed (Jul-Jørgensen et al., 2024):

$$SPE_i = \underline{e}_i^T \underline{e}_i \tag{4.4}$$

Where $\underline{e}_i$ is the (P × 1) error vector related to the i-th acquisition.

The UCL for SPE was computed as (Taris et al., 2017):

$$SPE_{lim} = \theta_1 \left[1 + \theta_2 h_0 \left(\frac{h_0 - 1}{\theta_1^2} \right) + \frac{c_\alpha h_0 \sqrt{2\,\theta_2}}{\theta_1} \right]^{\frac{1}{h_0}} \tag{4.5}$$

For which it can be written that:

$$\theta_k = \sum_{r=2}^{P} \lambda_r^k \tag{4.6}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\,\theta_2^2} \tag{4.7}$$

In Eq. 4.5, c_α is the α-th upper percentile of a normal distribution with zero mean and unity variance. The fault detection algorithm assessed deviations from NOC conditions using an approach based on the number of consecutive points above the limit (Jul-Jørgensen et al., 2024).

Specifically, Raman spectra from the validation samples (3 NOCs and 5 faults) were iteratively projected onto PC space built from training spectra corresponding to a single, continuously updated experimental time point. Subsequently, the T^2 and SPE statistics for the validation acquisition were computed and compared to the corresponding limit value for a predefined number of consecutive points. If all consecutive T^2 or SPE values exceeded the limit, a fault was detected. The projection is performed until a fault is detected or all acquisitions of the same experiment are considered. Different values of consecutive points to raise the alarm and α values were assessed to maximize the accuracy. Once the T^2 or SPE-based model was tuned, the variables' contribution were assessed to perform the fault diagnosis, that is the identification of which Raman shifts mostly contribute to the deviation of T^2 and SPE from the NOC behavior. The Complete Decomposition Contributions (CDC) applied on a one-dimension PC subspace were employed for T^2 and SPE statistics (Alcala & Joe Qin, 2011):

$$c_i^{T^2} = \left(\frac{\underline{\underline{p}}\underline{p}^T\underline{x_i}}{\sqrt{\lambda_1}}\right)^2 \tag{4.8}$$

$$c_i^{SPE} = \underline{e_i}^2 = \left(\underline{x_i} - \underline{\hat{x}_i}\right)^2 = \underline{x_i} - t_i\underline{p}^T \tag{4.9}$$

Here, $\underline{x_i}$ denotes the i-th spectrum acquired experimentally, while $\underline{\hat{x}_i}$ is the corresponding PCA-model estimation. The use of local-PCA implies a loss of time correlation information between successive samples (Jul-Jørgensen et al., 2024), which although essential for investigating the technological parameters such as curd cutting time (Sibono et al., 2025a), it does not constitute a critical requirement for fault detection purposes. After tuning the models with the optimal combination of hyperparameters, the average time at fault was computed to assess how quickly the failure is detected. This time param-eter was ultimately used as a comparison metric for the selection of the best algorithm

4.1.2 Results and discussion: Global PCA-based algorithm

Before validating the algorithms, the Raman spectra from the entire dataset were explored using a PCA model. Such an operation was carried out to visually highlight differences between NOC and fault experiments. Figure 4.1 emphasize the divergence between NOC and fault experiments in terms of PC1 evolution over time. In particular, such a figure has been built by calibrating the PCA model using the NOC samples and projecting fault samples using the pre-derived loading values.

As can be observed from the sigmoidal shape of PC1 vs time profile, the scores related to the first PC account for the dynamic evolution of curd firmness during milk coagulation explaining 43% of the total variance. On the other hand, the second and third PCs, which respectively explained 12% and 11% of the variance, did not exhibit any additional information, and were therefore discarded from the analysis together with higher-dimensional PCs. The fault experiments F1, F2 and F3 led to a delay of milk gelation, reflected by the sluggish dynamics of the relevant PC1 scores. In addition, the temperature drop induced by switching off the water bath (fault 4) results in the thermal limitation on coagulation kinetics. In such a case the PC1 values fell inside the NOC area for up to 8 minutes, after which the curve trajectory dropped below the lower limit. This delay is likely due to the probe position near the surface and close to the center of the cylindrical beaker: under quiescent conditions, heat transfer occurs predominantly by conduction, so that the effect of thermal disturbance required time to establish a significant radial temperature gradient within the system and become detectable by the probe.

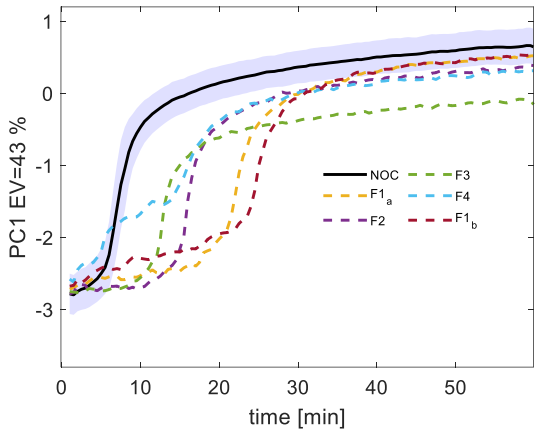


Figure 4.1 Time evolution of PC1 during milk coagulation. The NOC curve (solid black) represents normal operating conditions, shown as the mean trajectory and the standard deviation, denoted with the shaded area. Dashed curves indicate fault scenarios, induced by variations in initial temperature and enzyme concentration.

Six different algorithms were subjected to cross-validation in order to evaluate their ability to detect faults in milk coagulation induced by temperature control failures and reduced rennet dosage. For each algorithm, the hyperparameters were tuned according to the respective accuracy maximization. To this concern, the non-linear-model-based fault detection (2) and T^2 or SPE

algorithm (5) are discussed in the following sections. The results obtained from the application of algorithms (1), (3), (4) and (6) are reported in Appendix C, sections C.2 and C.3.

4.1.2.1 PC scores-Non-linear-regression-model-based fault detection

Eq. 4.1 was used to model the PC1 score profiles evolution of NOC samples, establishing a reference trajectory of curd properties as monitored through Raman spectroscopy. Based on the global PCA principles, the Algorithm (2) provides an equation-based approach that exhibits low sensitivity to noise, as it performs data fitting on PC1 associated with each spectral acquisition. By fitting Eq. 4.1 to PC1 scores from NOC experiments the following parameters were obtained:

$$\hat{S}_1 = -2.69 + \frac{3.51}{1+\left(\frac{8.07}{t}\right)^{2.72}} \quad (4.10)$$

For this non-linear regression model, it resulted that R^2=0.91, indicating that the equation was capable of satisfactorily capturing the variability of S_1, along with narrow parameters confidence intervals (from 6 to 12% of the estimated value), denoting a low correlation among parameters.

The non-linear Eq. 4.1 depicting the overall behavior of NOC batches was supplemented with UCL and LCL to delineate the region of acceptability. The deviation from nominal condition was assessed for each individual acquisition through the difference between its PC1 value and the model's prediction confidence interval. To this concern, figure 4.2 shows the accuracy profile when varying the number of consecutive points to trigger the alarm for three confidence levels (1-α). As can be observed, by fixing the confidence level, it is possible to determine the number of consecutive acquisitions that maximizes the number of correctly classified batches, hence achieving the highest accuracy. From figure 4.2, it is possible to infer that the detection accuracy is strongly sensitive to the value of α. Indeed, when reducing α from 0.05 to 0.01, the optimal accuracy increased by 5 percentage points, and by 9 percentage points compared to the case with α = 0.09.

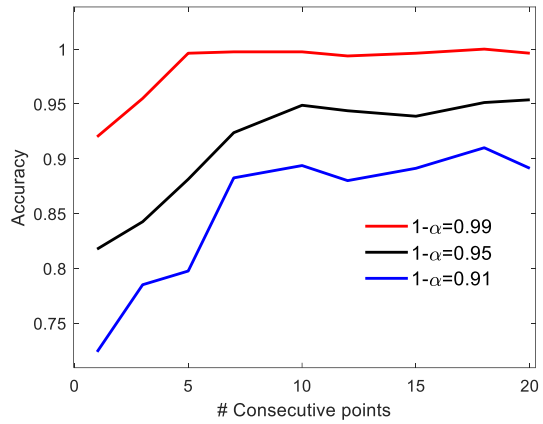


Figure 4.2 Accuracy as a function of the number of consecutive points for different values of confidence levels (1-α) when using a PC scores-Non-linear-regression-model-based fault detection algorithm.

In order to precisely find the optimal configuration of the detection algorithm, one has to decrease the mesh size of the hyperparameters grid. To this regard, the accuracy heatmap offers a clear visualization of the algorithm's performance profile. Figure 4.3 depicts the accuracy profile over the hyperparameter space associated with the application of algorithm (2) on the cross-validation scheme, highlighting that the optimal performance was obtained for a confidence level of 0.99 combined with 5 consecutive points outside the NOC region for which an accuracy of 99.63% and a unitary FDR were obtained.

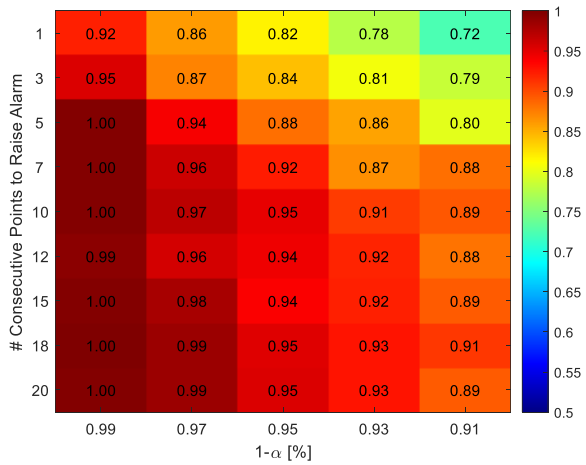


Figure 4.3 Accuracy as a function of the number of consecutive points and the confidence level for the non-linear-regression-model-based algorithm.

The hyperparameters combination that maximized the accuracy was employed on the whole dataset to graphically illustrate its fault detection capability (Figure 4.4). Under fault conditions, the PC1 trajectories systematically deviate from the nominal behavior, as they cross the control boundaries for at least 5 consecutive acquisitions. Overall, this model resulted in a fast and accurate identification of faults while retaining a high specificity.

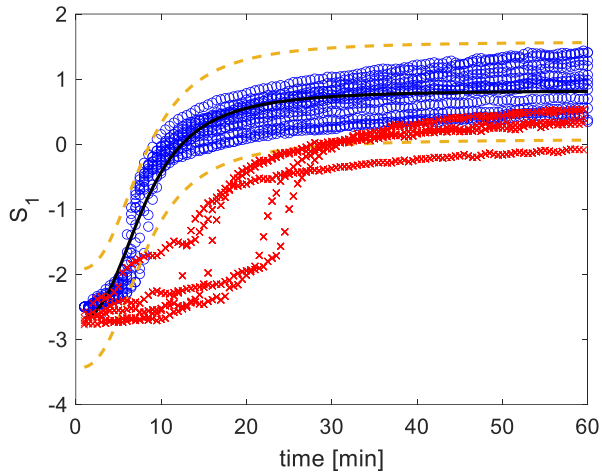


Figure 4.4 Application of the hill equation-based fault detection algorithm to the whole dataset under the selection of the optimal hyperparameters combination. The blue circles represent the NOC trajectories while the red crosses correspond to faulty conditions. The dashed orange lines denote UCL and LCL under the optimal α value (0.01).

For the sake of faults diagnosis, PC1 score prediction confidence interval values at the alarm time were projected back onto the original variable space using the loading values (Figure 4.5a). These reconstructed spectra represent the expected range of normal behavior and were compared to the actual spectrum recorded at the alarm time (8.5 minutes after rennet addition) for an illustrative fault batch (T=34 °C C_r=0.007 IMCU mL^{-1}). For reference, a spectrum from a NOC batch acquired at the same time was also included. As shown in Figure 4.5a, the fault spectrum lies predominantly above the control region, with the NOC spectrum remaining within expected bounds. Figure 4.5b displays the corresponding error plot, derived by calculating the difference between the fault spectrum and the spectrum reconstructed by projecting the model-predicted score back into the original variable. The strongly negative error value observed at 410 cm^{-1} is attributed to baseline effects, likely linked to changes in the curd's firmness (Sibono et al., 2025a). Other important deviations from the nominal spectrum appear between 1000 and 1600 cm^{-1},

corresponding to protein-associated Raman bands, suggesting a slower biochemical evolution due to lower enzyme concentration. This behavior is clearly reflected in the error plot (Figure 4.5b), where positive deviations at 1133, 1176, and 1552 cm^{-1}, which denote the vibrational modes associated with aspartic and glutamic acids, along with N-H deformation (Rodrigues Júnior et al., 2016; Stephani et al., 2017; Zhao et al., 2020).

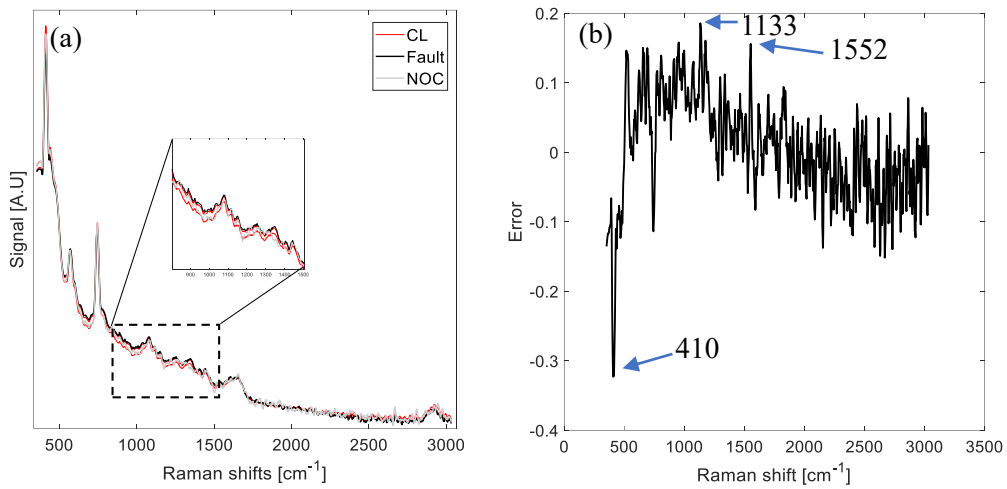


Figure 4.5 Comparison of NOC and fault spectra (T=34 °C C_r=0.007 IMCU mL^{-1}) along with the reconstructed spectra using the control limits values in correspondence of the fault detection time (a), and the associated error plot (b).

4.1.3 Results and discussion: Local PCA-based algorithm

In the following discussion, the hyperparameter tuning for the presented dataset is examined along with the diagnosis of fault batches using T^2 or SPE algorithm (5).

4.1.3.1 Hyperparameters tuning

Figure 4.6 shows the accuracy heatmaps resulting from tuning the T^2 or SPE model corresponding to algorithms (5), using different couples of hyperparameters. The x-axes are the 1-α percentile chosen for alarm limit. The y-axes indicate how many points should be above the UCL before the alarm triggers.

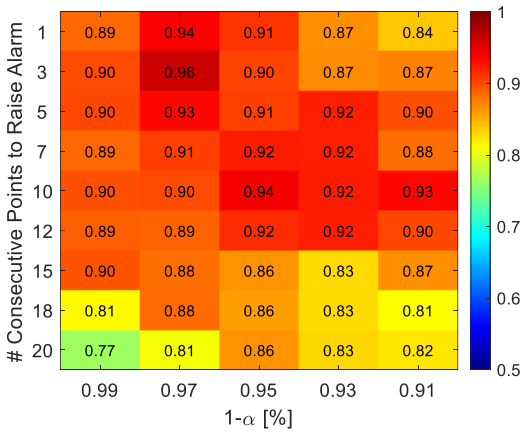


Figure 4.6 Heatmap for the accuracy as a function of the hyperparameters corresponding to T^2 and SPE fault detection algorithm.

The best performance for the T^2-SPE-based detection algorithm using the logical OR approach was achieved by three consecutive points above the UCL and α=3%, yielding an accuracy of 96%, meaning that this local PCA algorithm is a potential candidate for addressing the problem. Further details about the other local-based algorithms are reported in Appendix C, Section C.3. It is important to specify that these considerations are specific to the presented dataset, which nonetheless reflects reasonable operating conditions.

4.1.3.2 Fault diagnosis

After tuning the detection model, fault diagnosis enables to investigate the source of the process failures. The joint representation of SPE and T^2 in a single plot allows for a combined visualization of both the explained and residual data variance. Figure 4.7 shows the batches trajectories in the T^2–SPE space, including NOCs from both the calibration and validation sets, as well as batches affected by reduction in rennet concentration at two temperatures (34°C and 36°C), and the temperature control system fault. While NOC trajectories remain confined within a tight region near the origin, all fault batches deviate significantly. Particularly, the rennet concentration faults exhibited the most pronounced divergence from NOC conditions with the deviation being more substantial at 34 °C than at 36 °C. This behavior can be attributed to the greater kinetic limitation posed by the lower temperature, resulting in a slower dynamic response of the signal. Conversely, the temperature fault shows a moderate deviation from NOC compared to the other faults, particularly in terms of T^2. Regarding the validation NOC samples, their trajectories remain within the region delimited by NOC calibration samples.

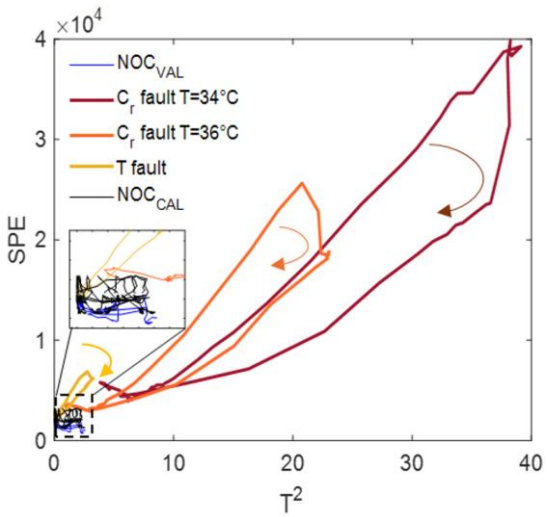


Figure 4.7 T^2 vs SPE plot showing fault trajectories compared to those from calibration and validation NOC samples, highlighting the deviation patterns under rennet concentration and temperature faults. The outcomes of this plot were obtained from a generic iteration of cross-validation scheme.

Figure 4.8 presents the evolution of statistical monitoring indices over time. For simplicity, only T=34 °C is displayed for NOC analysis, while temperature control fault (F4) and the F2 run are considered among the fault experiments. The plots are organized into three rows and two columns, with each row corresponding to a specific scenario. The left column shows the analysis for T^2 statistic and the right column displays the resulting curve for SPE. Both statistics show a time variation of the control limits resulting from a local PCA model. Concerning the T^2 statistic, a noticeable increase is generally observed around the PC1 inflection point across all cases. This behavior is primarily attributed to the high variability in the process dynamics in the proximity of inflection points, which probably arises from different temperatures used to calibrate the model. These sources of variability lead to the formation of T^2 peaks, which do not exceed the control limits in the case of NOC batches, but become significantly more pronounced in the presence of enzyme concentration faults. While the T^2 algorithm correctly classified the NOC batches and the rennet concentration faults, it failed in properly identifying the temperature control fault, due to the modest deviations in PC1 values with respect to the other types of faults. Such a miss probably justifies the smaller accuracy values observed when only using T^2 statistics (Appendix C, section C.3). Conversely SPE resulted in showing better sensitivity to all the types of faults, including the rennet concentration failures at 34 and 38 °C.

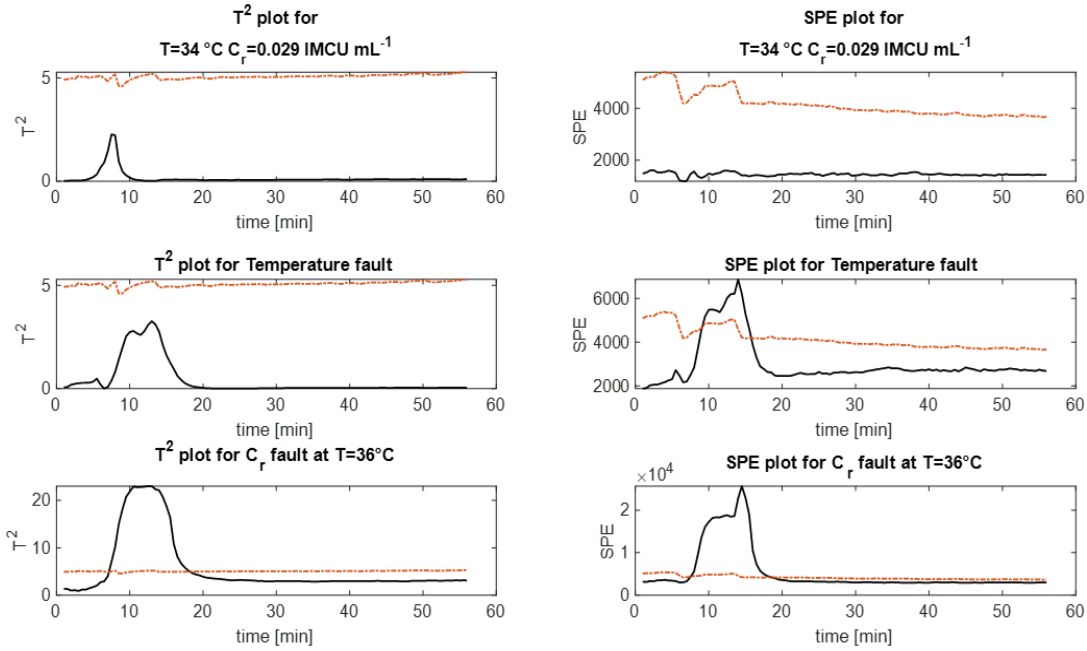


Figure 4.8 T^2 (left column) and SPE (Right column) statistics time evolution for a NOC (first row) and two faulty batches (second and third rows). The statistics values is indicate through the black line, while the control limits are denoted by the orange dashed lines.

The analysis of contribution plots derived from eqs. 4.8 and 4.9 allows for a deeper understanding of the Raman shifts involved in the fault occurrence. Particularly, this investigation opens up the possibility to discriminate the source of the faults. Figure 4.9 shows the contribution plots for both SPE and T^2 associated to the spectrum corresponding to the time at fault using the optimal combination of hyperparameters. Specifically, the alarm triggered after 9.5 minutes for fault F2, while 7 minutes were required to detect the fault for Fault $F1_a$. Interestingly, these two similar types of fault share high SPE contribution in correspondence of 1553 cm^{-1} which is known to belong to the bending mode of N-H bond and stretching of C-N corresponding to amide II (Rodrigues Júnior et al., 2016; Stephani et al., 2017; Zhao et al., 2020). Such an outcome is reasonably due to the involvement of proteins into the rennet-casein phenomena. Other remarkable, yet less intense SPE contribution values were observed between 900 and 1200 cm^{-1} region, which is known to take part of milk constituents redistribution during rennet coagulation (Sibono et al., 2025a). Regarding the contribution from Raman shifts below 900 cm^{-1}, their significance is likely associated with baseline variations, as the most intense signal was observed in the lower wavenumber region and was not corrected as baseline shift accounts for physical

changings of curd during renneting. In regard to the T^2 algorithm, the highest contribution values are predominantly located in the lower Raman shift region, along with a notable effect at 1437-1443 cm^{-1} which corresponds to C-H bending of -CH2 group (El-Abassy et al., 2011). In contrast, no specific vibrational assignment has yet been reported for 1286 cm^{-1} band.

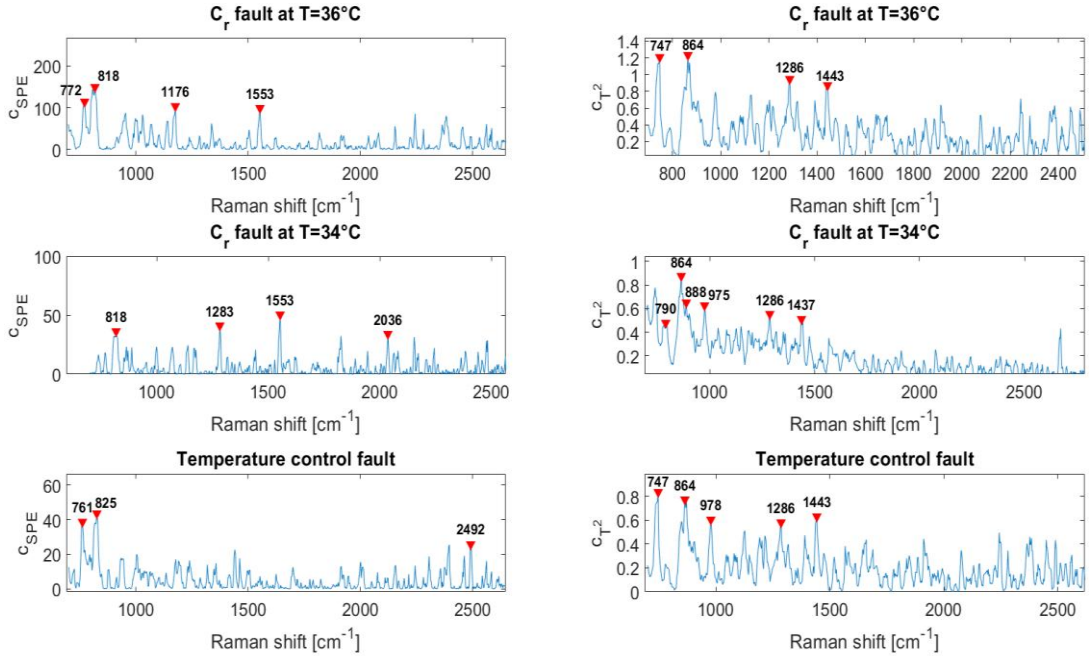


Figure 4.9 Variable contributions to the SPE (left column) and T^2 (right column) for three different fault types. The contribution plots were obtained at the corresponding time of fault.

4.1.4 Comparison of Detection Algorithm Performance

Six fault detection algorithms were explored to detect two common types of faults in food processing (temperature and reactant concentration) using the multivariate signal provided by Raman spectroscopy. This section outlines the practical implications of applying these algorithms, emphasizing the technical benefits of an accurate and fast fault detection within a dairy processing environment. Table 4.2 summarizes the results obtained by tuning the hyperparameters associated with the optima settings, including those reported in Appendix C. It is worth mentioning that the PCA-based models were trained over 100 iterations to identify the optimal configuration, and the resulting maximum performance was further confirmed using 1000 iterations. The consistency of the outcomes across iterations demonstrates the stability of the PCs, even when different process disturbances were introduced.

Table 4.2. Fault detection results for each model obtained under the optimal combination of hyperparameters.

Parameter	# consecutive points	α	Accuracy	FDR	FAR	$\overline{t}_{fault}$ [min]
SCI	10	0.01	0.90	0.98	0.18	13.3
Non-linear model	5	0.01	0.996	1	0.01	8.49
T^2	1	0.07	0.92	0.84	0	5.2
SPE	3	0.03	0.96	1	0.09	6.9
SPE and T^2	1	0.01	0.91	0.82	0	8.4
SPE or T^2	3	0.03	0.96	0.99	0.07	6.0

It is worth noting that T^2 resulted to be the fastest detection algorithm and SCI the slowest. The fast detection of T^2 is probably related to its incapability of detecting temperature control fault, which resulted in occurring later than the rennet concentration faults, and therefore was not accounted in the average alarm time computation. The most accurate algorithm (i.e. non-linear model) demonstrated an intermediate fault detection time, constituting a possible tool for this application.

Importantly, the characteristic time of fault detection for the presented algorithms well fits with the industrial practice of milk coagulation. To this concern SPE or T2 algorithm showed an accuracy

of three percentage points lower than the sigmoidal model while achieving fault detection 2.5 minutes faster (i.e. 6 minutes). Several studies demonstrated that, the coagulating milk stands in a pre-gelation state where milk has not changed phase yet (Abdelgawad et al., 2014; Castillo et al., 2006; Sibono et al., 2025a). This means that the window for time at fault has to occur earlier than the gelation point, which is the time at which the elastic modulus G' = 1 Pa, a practical threshold for deeming the initialization of the phase transition triggered by rennet introduction (Abdelgawad et al., 2014). As reported in the previous chapter, where rheological analyses were conducted under the same experimental conditions as the NOC batches presented in this work, gelation was found to occur between 7.5 and 21 minutes after rennet addition. This indicates that the SPE or T^2 fault detection system effectively provides an early warning, which could potentially allow for timely intervention to prevent process deviations before the phase transition of milk. In conclusion, the SPE or T^2 algorithm was selected as the best candidate for this study. The optimal algorithms reported herein exhibited similar performance to those reported in a similar work where FT-NIR was used as a fault detection tool (Grassi et al., 2022). However, it should be noted that the latter employed fault batches that slightly differed from those considered here. The proposed methodology offers the additional advantage by enabling the diagnosis of the contribution of each investigated variable to the onset of the fault, which can be used as a biomarker of quality. Although the investigated operative condition window reflects the one typically employed in industrial settings, this study is limited by the relatively small size and controlled nature of the available datasets. Further validation is therefore required to confirm the robustness and generalizability of the pro-posed approach, ideally by extending the dataset and introducing additional sources of variability that better represent real process fluctuations. In addition, future studies should explore the integration of control strategies to assess whether the system can actively respond to detected faults and bring potential failure batches back to normal operating conditions.

4.2 Conclusions

This chapter presents different fault detection algorithms to investigate the use of Raman spectroscopy as a screening tool for food processing applications. Different nominal and failure conditions were investigated to assess the fault detection rate and false alarm rate. Local and global PCA-based algorithms were employed to find the best options in terms of detection speed and accuracy. High accuracies were obtained by each investigated algorithm, with the SPE or T^2 algorithm emerging as particularly effective, reaching up to 96% accuracy and enabling the detection of deviations from NOC before the occurrence of colloid destabilization. The analysis of variable contributions further demonstrated the capability of Raman spectral data, when combined with multivariate statistical modeling, to identify and interpret the source of faults within the process, thus offering additional benefits over previous studies. In particular, spectral bands below 900 cm^{-1} or ranging between 1100 and 1600 cm^{-1} were found to be consistently involved in fault scenarios, suggesting a close relationship between process failures and specific molecular vibrations. The presented methodology holds the potential of achieving a correct diagnosis of process failures prior to the curd gelation point, thus providing a timely informed decision-making strategy to dairy manufacturers in the industrial settings. While the results obtained here provide a basis for fault detection process monitoring, future work may consider the application of this methodology to test the sensor capability of tracking any corrective actions. Subsequent investigations should also aim to extend this framework to evaluate how such early interventions affect the critical quality attributes of the product. Nonetheless, the integration of Raman spectroscopy-based monitoring with MSPC opens promising perspectives for enhancing product quality, reducing manufacturing costs and avoiding waste.

4.3 References

Abdelgawad, A. R., Guamis, B., & Castillo, M. (2014). Using a fiber optic sensor for cutting time prediction in cheese manufacture from a mixture of cow, sheep and goat milk. Journal of Food Engineering, 125(1), 157–168. https://doi.org/10.1016/j.jfoodeng.2013.10.001

Alcala, C. F., & Joe Qin, S. (2011). Analysis and generalization of fault diagnosis methods for process monitoring. Journal of Process Control, 21(3), 322–330. https://doi.org/10.1016/j.jprocont.2010.10.005

Castillo, M. (2006). Cutting time prediction methods in cheese making. Encyclopedia of Agricultural, Food, and Biological Engineering, 1, 1–7. http://manuelcastillo.es/pdfs-publications/Cutting Time Prediction Methods in Cheese Making.htm.pdf

Castillo, M., Lucey, J. A., & Payne, F. A. (2006). The effect of temperature and inoculum concentration on rheological and light scatter properties of milk coagulated by a combination of bacterial fermentation and chymosin. Cottage cheese-type gels. International Dairy Journal, 16(2), 131–146. https://doi.org/10.1016/j.idairyj.2005.02.005

De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. International Journal of Pharmaceutics, 417(1–2), 32–47. https://doi.org/10.1016/j.ijpharm.2010.12.012

Derra, M., Bakkali, F., Amghar, A., & Sahsah, H. (2018). Estimation of coagulation time in cheese manufacture using an ultrasonic pulse-echo technique. Journal of Food Engineering, 216, 65–71. https://doi.org/10.1016/j.jfoodeng.2017.08.003

Duran-Villalobos, C. A., Goldrick, S., & Lennox, B. (2020). Multivariate statistical process control of an industrial-scale fed-batch simulator. Computers and Chemical Engineering, 132. https://doi.org/10.1016/j.compchemeng.2019.106620

El-Abassy, R. M., Eravuchira, P. J., Donfack, P., Von Der Kammer, B., & Materny, A. (2011). Fast determination of milk fat content using Raman spectroscopy. Vibrational Spectroscopy, 56(1), 3–8. https://doi.org/10.1016/j.vibspec.2010.07.001

Ford Versypt, A. N., Harrell, G. K., & McPeak, A. N. (2017). A pharmacokinetic/pharmacodynamic model of ACE inhibition of the renin-angiotensin system for normal and impaired renal function. Computers and Chemical Engineering, 104, 311–322. https://doi.org/10.1016/j.compchemeng.2017.03.027

Grassi, S., Strani, L., Alamprese, C., Pricca, N., Casiraghi, E., & Cabassi, G. (2022). A FT-NIR process analytical technology approach for milk renneting control. Foods, 11(1). https://doi.org/10.3390/foods11010033

Hayes, E., Greene, D., O'Donnell, C., O'Shea, N., & Fenelon, M. A. (2023). Spectroscopic technologies and data fusion: Applications for the dairy industry. Frontiers in Nutrition, 9. https://doi.org/10.3389/fnut.2022.1074688

Jul-Jørgensen, I., Facco, P., Gernaey, K. V., Barolo, M., & Hundahl, C. A. (2024). Data fusion of Raman spectra in MSPC for fault detection and diagnosis in pharmaceutical manufacturing. Computers and Chemical Engineering, 184. https://doi.org/10.1016/j.compchemeng.2024.108647

Lane, T. P., & Dumouchel, W. H. (1994). Simultaneous confidence intervals in multiple regression. American Statistician, 48(4), 315–321. https://doi.org/10.1080/00031305.1994.10476090

Lyndgaard, C. B., Engelsen, S. B., & Van Den Berg, F. W. J. (2012). Real-time modeling of milk coagulation using in-line near infrared spectroscopy. Journal of Food Engineering, 108(2), 345–352. https://doi.org/10.1016/j.jfoodeng.2011.07.029

Nicolau, N., Buffa, M., O'Callaghan, D. J., Guamis, B., & Castillo, M. (2015). Estimation of clotting and cutting times in sheep cheese manufacture using NIR light backscatter. Dairy Science and Technology, 95(4), 495–507. https://doi.org/10.1007/s13594-015-0232-7

Pratt, W. K. (2007). Digital Image Processing (J. W. & Sons (ed.); 4th ed.).

Pretto, D., Kaart, T., Vallas, M., Jõudu, I., Henno, M., Ancilotto, L., Cassandro, M., & Pärna, E. (2011). Relationships between milk coagulation property traits analyzed with different methodologies. Journal of Dairy Science, 94(9), 4336–4346. https://doi.org/10.3168/jds.2011-4267

Reis, M. S., Rendall, R., Rato, T. J., Martins, C., & Delgado, P. (2021). Improving the sensitivity of statistical process monitoring of manifolds embedded in high-dimensional spaces: The truncated-Q

statistic. Chemometrics and Intelligent Laboratory Systems, 215. https://doi.org/10.1016/j.chemolab.2021.104369

Rodrigues Júnior, P. H., De Sá Oliveira, K., Almeida, C. E. R. De, De Oliveira, L. F. C., Stephani, R., Pinto, M. D. S., Carvalho, A. F. De, & Perrone, Í. T. (2016). FT-Raman and chemometric tools for rapid determination of quality parameters in milk powder: Classification of samples for the presence of lactose and fraud detection by addition of maltodextrin. Food Chemistry, 196, 584–588. https://doi.org/10.1016/j.foodchem.2015.09.055

Schmid, M., Rath, D., & Diebold, U. (2022). Why and How Savitzky-Golay Filters Should Be Replaced. ACS Measurement Science Au, 2(2), 185–196. https://doi.org/10.1021/acsmeasuresciau.1c00054

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. (2025a). Monitoring of milk rennet coagulation: Chemical and physical perspective using Raman spectroscopy. Applied Food Research, 5(1). https://doi.org/10.1016/j.afres.2025.100701

Sibono, L., Tronci, S., Hedegaard, M. A. B., Errico, M., & Grosso, M. (2025b). PAT-driven dairy processing: a rheo-Raman-based kinetic model for in-line prediction of milk coagulation dynamics. Analytical and Bioanalytical Chemistry. https://doi.org/10.1007/s00216-025-05965-2

Stephani, R., Oliveira, K. de S., de Almeida, C. E. R., Perrone, Í. T., de Carvalho, A. F., de Oliveira, L. F. C., & Almeida, M. R. (2017). Raman spectroscopy as a tool to identify modification of whey protein concentrate (WPC) during shelf life. Food Packaging and Shelf Life, 11, 1–9. https://doi.org/10.1016/j.fpsl.2016.10.001

Strani, L., Grassi, S., Alamprese, C., Casiraghi, E., Ghiglietti, R., Locci, F., Pricca, N., & De Juan, A. (2021). Effect of physicochemical factors and use of milk powder on milk rennet-coagulation: Process understanding by near infrared spectroscopy and chemometrics. Food Control, 119. https://doi.org/10.1016/j.foodcont.2020.107494

Tajammal Munir, M., Yu, W., Young, B. R., & Wilson, D. I. (2015). The current status of process analytical technologies in the dairy industry. Trends in Food Science and Technology, 43(2), 205–218. https://doi.org/10.1016/j.tifs.2015.02.010

Taris, A., Hansen, T. B., Rong, B. G., Grosso, M., & Qu, H. (2017). Detection of Nucleation during Cooling Crystallization through Moving Window PCA Applied to in Situ Infrared Data. Organic Process Research and Development, 21(7), 966–975. https://doi.org/10.1021/acs.oprd.7b00076

Villaquiran, Z., Zamora, A., Arango, O., & Castillo, M. (2024). Inline Determination of the Gel Elastic Modulus During Milk Coagulation Using a Multifiber Optical Probe. Food and Bioprocess Technology. https://doi.org/10.1007/s11947-023-03294-9

Yang, M., Ye, A., Yang, Z., Everett, D. W., Gilbert, E. P., & Singh, H. (2022). Kinetics of pepsin-induced hydrolysis and the coagulation of milk proteins. Journal of Dairy Science, 105(2), 990–1003. https://doi.org/10.3168/jds.2021-21177

Zhao, M., Markiewicz-Keszycka, M., Beattie, R. J., Casado-Gavalda, M. P., Cama-Moncunill, X., O'Donnell, C. P., Cullen, P. J., & Sullivan, C. (2020). Quantification of calcium in infant formula using laser-induced breakdown spectroscopy (LIBS), Fourier transform mid-infrared (FT-IR) and Raman spectroscopy combined with chemometrics including data fusion. Food Chemistry, 320. https://doi.org/10.1016/j.foodchem.2020.126639

Chapter 5: Raman spectroscopy as a non-invasive tool for monitoring combined acid-rennet coagulation of milk

In this chapter, Raman spectroscopy is applied to monitor milk coagulation, with the inclusion of a combined coagulation strategy, which contemplates the use of both bacteria and rennet. This work represents a preliminary study, and the results are still under evaluation for potential publication.

5.1 The problem of monitoring lactic acid fermentation in dairy products: an overview

Milk coagulation can be induced by acidification, renneting or by a combined effect of acid and enzyme. These coagulation methods offer a diverse range of curd properties with practical implications to the final product quality, such as texture, taste and composition (Tarapata et al., 2021). Despite gel firmness in acid-rennet type cheese primarily depends on pH variation and to a lesser extent on rennet action, the combined use of rennet and starter cultures in milk fermentation confer unique gel properties that cannot be achieved by using either rennet or starter alone, as the their cooperative action exerts a major influence on the rheological properties (Castillo, Lucey, & Payne, 2006). Therefore, in the industrial environment, combined technology is often preferred over other coagulation strategies. Several works investigated the effect of different process parameters including temperature and starter culture concentration on the optical and rheological properties of coagulated milk using combined fermentation (Abdelgawad et al., 2014; Castillo, Lucey, & Payne, 2006; Castillo, Lucey, Wang, et al., 2006; Tarapata et al., 2021). Such studies represent a benchmark for monitoring the technological properties of acid-rennet milk coagulation, which embrace clotting time, cutting time and somatic cell count. Despite these technological parameters are important for ensuring product quality, the increasing safety demand requires a broader understanding of curd processing (Hayes et al., 2023). In particular, it is crucial to track the temporal concentration of chemical species during fermentation, because small off-specification in the chemical properties unavoidably change the characteristics of the food along its production chain (Rodriguez et al., 2013). For instance, monitoring lactic acid content is essential in curd production, as it reflects the progression of fermentation, and serves as an indicator of strain efficiency (Bouteille et al., 2013). Traditionally, lactic acid is quantified by acid–base titration (Torrestiana et al., 1994). However, this method is not easily applicable to production lines and requires the use of chemical reagents. To address this limitation, nuclear magnetic resonance spectroscopy has been employed to provide a picture of metabolites involved within the fermentation system, although it relies on sample extraction and is characterized by acquisition times ranging from 20 minutes to few hours (Bouteille et al., 2013).

Measuring other sugars such as glucose and galactose is also important, as their consumption patterns provide insights into microbial metabolism, influence acidification kinetics and affect cheese sensorial properties (Sibono et al., 2024). Within this framework, Raman spectroscopy been employed as a tool for process monitoring of lactic acid fermentation (Rodriguez et al., 2013;

Sivakesava et al., 2001). Its main strengths lie in the minimal sample preparation required, the rapid acquisition of spectra that enables real-time tracking of chemical transformations, and its suitability for aqueous environments where other spectroscopic techniques are often hindered by broad signal of OH group (De Beer et al., 2011; Rodriguez et al., 2013). Moreover, the growing availability of affordable instrumentation has expanded its applications, making it increasingly valuable for studying microbial activity and the kinetics of biochemical changes (Rodriguez et al., 2013). In general, despite Raman spectroscopy has been extensively studied for several fermentation process applications, its use as a real-time monitoring tool of different analytes concentration during the combined action of starter culture and rennet remains unexplored. It is worth mentioning that in the study presented by Sivakesava et al (2001) FT-Raman spectroscopy has been employed to estimate sugars and lactic acid during bacterial fermentation. However, in the same study rather low predictive performance was achieved for lactic acid concentration ($Q^2 < 0.4$). These gaps provide the basis for the present study, which intends to extend the current knowledge in quality monitoring of dairy food processing. In particular, this work aims to address the following tasks:

1) Couple Raman spectroscopy with supervised multivariate models (PLS regression) to provide a real time estimation of four metabolites involved in lactic acid fermentation (Glucose, galactose, lactose and lactic acid)
2) Use Raman spectroscopy as a non-invasive tool for inline estimation of pH, which stands as an important technological parameter for predicting gelation time, cutting time and product safety. However, this study only considers the real time pH estimation.

5.2 Materials and methods

5.2.1 Materials

Coagulation experiments were performed on a Mettler Toledo EasyMax 102 crystallization system equipped with temperature control and 100 mL of maximum operative capacity. The pH value was determined by a VWR pHenomenal pH 1100 L pHmeter (Mettler Toledo, USA). In this study, CaCl2 (≥99 %, Sigma-Aldrich) was used to precondition the system. The calf rennet was a blended solution of Chymosin 85% ± 3% and Pepsin 15% ± 3% (Hjemmeriet, Denmark, 145 IMCU mL-1 of rennet), while the acidification was performed through direct vat set starter cultures made of Thermophilic lactic acid bacteria (*Streptococcus Thermophilus*, STI-12/13/14, Hjemmeriet, Denmark).

5.2.2 Sample preparation and DoE

The same experimental procedure reported in chapter 3 and 4 was followed to set the reconstituted SMP for conducting the experiment. The operative volume was 100 mL. In this case, temperature and inoculum concentration were the factors under investigation. Six experimental conditions were explored: two levels for temperatures (36 °C and 44 °C) and three starter culture concentration (0.1, 0.15 and 0.2 g L^{-1} of milk). For each experiment, after the sample reached the test temperature, the operative amount of starter culture was inoculate to the milk, which was stirred for 9 minutes at 100 RPM to promote bacterial acclimatation. Rennet was then added to the samples (0.00073 IMCU mL^{-1} of milk) and the system was stirred for 1 additional minute at the same speed. Afterwards, the stirrer was switched off and Raman and pH measurements were initiated simultaneously. Acquisition interval was 5 min for Raman and 15 min for pH measurements. Hourly, 2 mL of samples were gently extracted using a micropipette and stored in 2 mL vials for LC-MS/MS analysis. Sodium azide was added to each vial at a final concentration of 0.04% (w/v) to prevent microbial growth (Sibono et al., 2023). Each experimental condition was performed in duplicate. Two additional experiments at a temperature of 40°C and a starter culture concentration of 0.125 g L^{-1} and 0.175 g L^{-1} were performed and used as test samples. Overall, 12 experiments were used to calibrate and cross-validate the multivariate models, while 2 external experiments were included in the test set.

5.2.3 Raman measurements

Raman acquisitions were performed through Raman probe (Ocean Insights, FL, United States) using a 785 nm fiber coupled laser (B&W TEK Inc., Newark, United States) with an output power of 200 mW at the sample. This was coupled to a QE Pro-Raman+ (Ocean Insights, FL, United States) operating in a range of 350–3033 cm-1 and a resolution of 10–14 cm-1. Every acquisition was obtained by averaging 300 spectra, with an integration time of 1 s. Each time run provided 953 Raman intensities for 85 temporal acquisitions.

5.2.4 LC-MS/MS analysis

The Lc-MS/MS analysis here presented were carried out by external collaborators, whose contributions are acknowledged at the end of the dissertation. Milk samples were stored at 4 °C prior to processing. An aliquot of 300 µL of each milk sample was subjected to cold centrifugation at 17,700 rcf and 4 °C for 10 min. 100 µL of the supernatant was transferred to clean tubes and dried under nitrogen stream. The dried phase was then reconstituted with 400 µL of ultrapure water and 600 µL of acetonitrile. Finally, the extracts were filtered through 0.2 µm membranes and transferred into autosampler vials for LC–MS analysis.

Lactic acid, glucose and galactose, and lactose levels were measured using chromatographic techniques with a 1260 Agilent Infinity II Multisampler BIO binary pump. This chromatographic system was coupled to an Agilent Ultivo triple quadrupole mass spectrometer (Agilent Technologies, Santa Clara, USA). A BEH Amide chromatographic column (1.7 µm, 150 mm x 2.1 mm) from Waters, Ireland, maintained at 40 °C, was used for the analysis. The mobile phase consisted of water containing 10 mM ammonium acetate (A) and acetonitrile containing ammonium acetate (B) with a flow rate of 0.2 mL/min. The gradient began at 98 % (B) and linearly decreased to 40 % (B) over 0.5–3 min, then decreased to 40 % (B) from 3 to 5 min, and finally returned to the initial conditions for 4 min to re-equilibrate. Mass spectrometric analysis was conducted using a jetstream (ESI) source in both positive and negative ionization modes. Nitrogen served as the sheath gas, drying gas, and collision gas, with the sheath gas flow rate set at 11 L/min at 375 °C and the drying gas flow rate at 7 L/min at 300 °C. The nebulizer pressure was adjusted to 30 psi, with a capillary voltage of 3500 V and a nozzle voltage of 1500 V. The fragmentor voltage was set to 81 V, and the cell acceleration voltage to 4 V. Data acquisition was performed in multiple reaction monitoring (MRM) mode, controlled by Agilent MassHunter Workstation.

5.2.5 Chemometric techniques

Raman spectra were pre-processed with third order median filtering to remove cosmic spikes and Savitzky–Golay filtering with 2nd order polynomial and window width of 15 for signal smoothing. Spectra were then baselined using Whittaker baseline filtering with a penalty of 10^4 and a weight value equal to 0.1 (Eilers, 2003). Finally, spectra were subjected to SNV. Calibration matrix was unfolded variable wise to obtain a 2-D matrix for multivariate analysis. Prior to using Raman spectroscopy as a tool for real-time monitoring of analyte concentration, PCA was conducted to explore the data structure and relationships among variables. For the sake of a better representation, score values were mapped in a [0-1] space.

Afterwards, PLS regression was used to estimate glucose, galactose, lactose, lactic acid concentration and pH, given the Raman spectra as input. Separate response-specific models were developed in MATLAB, meaning that one model was constructed for each response variable. This approach enabled the optimization of the number of latent variables for each response, as followed by other studies (Chandrasekaran et al., 2025; Rhiel et al., 2002; Sivakesava et al., 2001). For the sake of a complete analysis of PLS models, their predictive performance was then assessed in terms of Cross-Validation R^2 (Q^2_{CV}) and external validation R^2 (Q^2_{EV}). In the latter case, the response variables values were estimated using Raman spectra obtained from test experiments. The number of LVs that maximized the average between Q^2_{CV} and Q^2_{EV} was selected as optimal. The correlation between the Raman intensity at every shift in the training data spectra and the response variables (pH and concentrations of glucose, galactose, lactose and lactic acid) was computed for variable selection purposes (Sivakesava et al., 2001). Regions with high correlations (Typically, r>0.45) were selected, while the regions that show low or no correlation were discarded from the analysis (Sivakesava et al., 2001). However, if prediction quality degradation was observed, the entire set of Raman shift was used for calibrating the PLS model. The chemometric analysis was performed through MATLAB® 2024b environment.

5.3 Results and Discussion

5.3.1 Exploratory data Analysis

Figure 5.1 displays the pH profile experimentally determined from milk acidification. As it can be seen, the acidification profile is kinetically consistent with the experimental conditions. Indeed, the most sluggish curve was obtained from the lowest temperature and starter concentration levels. A greater incubation temperature and bacterial concentration resulted in an increasing of pH curve steepness as a consequence of acidification rate increasing. The initial pH value was the same for all the experiments and equal to ~ 6.41. The slight differences in the pH values at the first acquisitions arise from different acclimatization temperature and inoculum concentration, which correspond to the relevant operative conditions. To this concern, a one-way ANOVA test was conducted to assess the homogeneity of the initial pH values across the replicated experiments. No statistically significant differences were observed ($p = 0.84$), confirming that all experiments approximately began with an equivalent pH value.

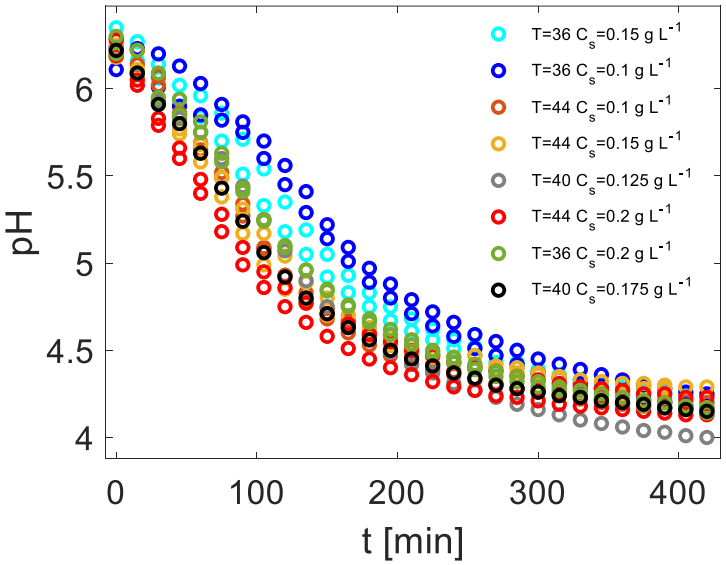


Figure 5.1 Acidification curve for each experimental condition, including the test experiments.

PCA model was calibrated using Raman spectra retrieved from coagulation assays to explore data trend and to assess the relationships among Raman intensities. The simultaneous recording of Raman signal and pH, allowed to better understand the variation of chemical composition during the acidification process. By way of example, the time evolution of pre-processed Raman spectra obtained from T=36 °C and C_s=0.15 g L^{-1} is shown in figure 5.2. Some recognizable vibrational

modes show signal variations within the Raman shift region comprising 300-1700 cm^{-1}. Specifically, Glycoside bond of alpha and beta lactose at 443-447 cm^{-1} shows a significant decrease in the intensity signal (Panthi et al., 2023), along with intensity increasing for C-O stretching vibration of the carboxyl group at 1078 cm^{-1} (Rodriguez et al., 2013) and superposition of C=O stretching vibration for lactic acid and amide I at 1654 cm^{-1} (Czaja et al., 2023). Such a signal intensity increasing is probably linked to lactic acid production. The band at 1350 cm^{-1} could be suggested to C–O stretching and C–O–H deformation of sugar molecules (Zhang et al., 2017), which explains the signal decreasing. Similar conclusions can be drawn when analyzing other experiments. No clear assignment was found for the signal at 740 cm^{-1}, which resulted in a strong decrease over time.

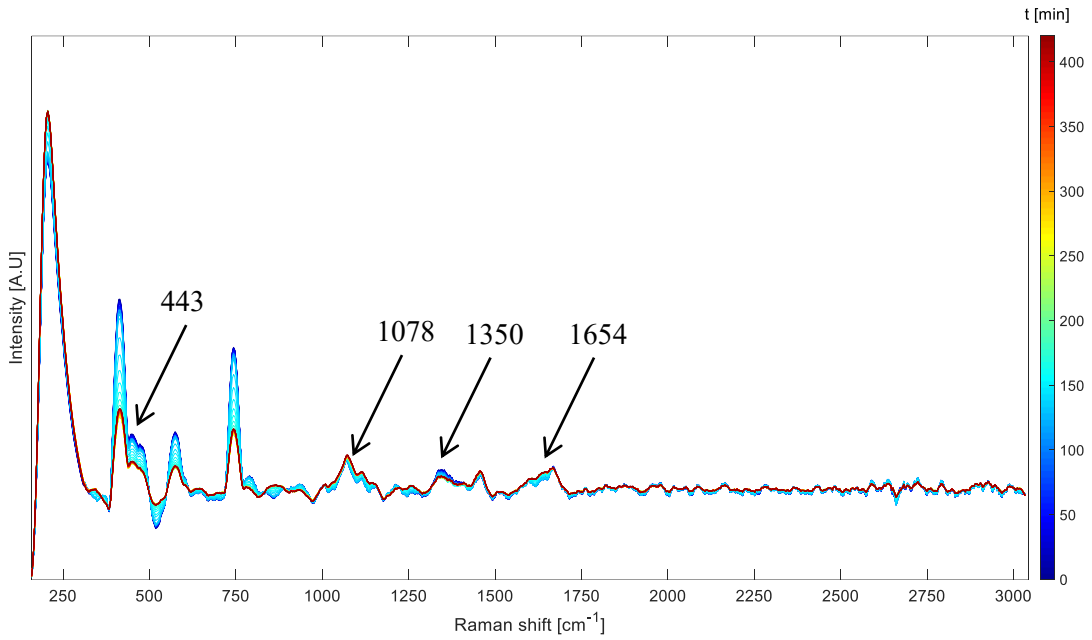


Figure 5.2 Evolution of Raman spectra over acidification time [min] at T=36 °C and C_s=0.15 g L^{-1}

Data were analyzed through PCA. The first PC, which explained 87% of total variance, was used to represent the data trend, as higher dimensional PCs didn't show clear pattern. Figure 5.3a shows the normalized PC1 as a function of acidification time under different temperatures and starter concentrations. Overall, the profiles exhibit a sigmoidal-like trend: an initial lag phase with low PC1 values, followed by a steep increase, and finally a maximum region approaching 1.0, followed by a slight decrease. Similarly to pH profiles, the rate of the transition vary depending on temperature and substrate concentration. Overall, both temperature and starter concentration play a key role in shaping the dynamics of the process. The fastest dynamics are observed for the experiment

conducted at 44 °C, since the corresponding PC values tend to cluster at earlier times, reaching the maximum value earlier than those performed at 36 °C. Secondly, experiments performed at the same temperature showed steeper dynamics when higher starter concentrations were inoculated. At 40 °C, the curves exhibited an intermediate behavior, with the test at 0.175 g L^{-1} being faster than that at 0.125 g L^{-1}. Figure 5.3b displays the loading values, which corroborate the behavior of Raman signal intensities.

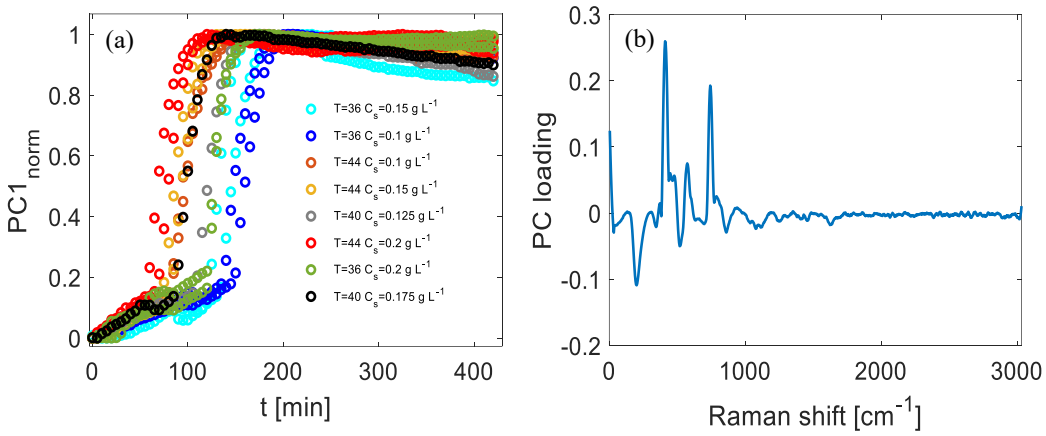


Figure 5.3 Normalized PC1 score values over acidification time under different experimental conditions (a) and the relevant loading values (b).

5.3.2 Quantification of Multiple Components Using Raman Spectroscopy

Raman spectrum provides a composite information of the components in the analyzed material that exhibits detectable inelastic scattering when excited from the laser probe (De Beer et al., 2011). The components set include glucose, galactose, lactose and lactic acid. Sugars are generally consumed by *Streptococcus thermophilus* cultures to produce lactic acid, which promotes system acidification. After calibration, PLS regression models were employed to predict analytes concentration using data that were not included in the training phase. For the sake of a deeper understanding of the system, Pearson correlation coefficients between metabolites' concentrations and signal intensity at every Raman shift were computed and reported in figure 5.4. By visually inspecting coefficients values, results are consistent with the loading analysis. Correlation coefficients ranged from -0.55 to 0.55 for monosaccharides, whereas stronger interdependence was observed for lactose and lactic acid ($-0.8 < r < 0.8$). Noteworthy, strong correlation (positive for

lactic acid and negative for lactose) was recorded in the 2800-3000 cm^{-1} region, which corresponds with C-H vibration (Czaja et al., 2023). Pearson coefficients between pH and Raman intensities mirrored that of lactic acid, but with opposite sign due to pH decrease during acidification.

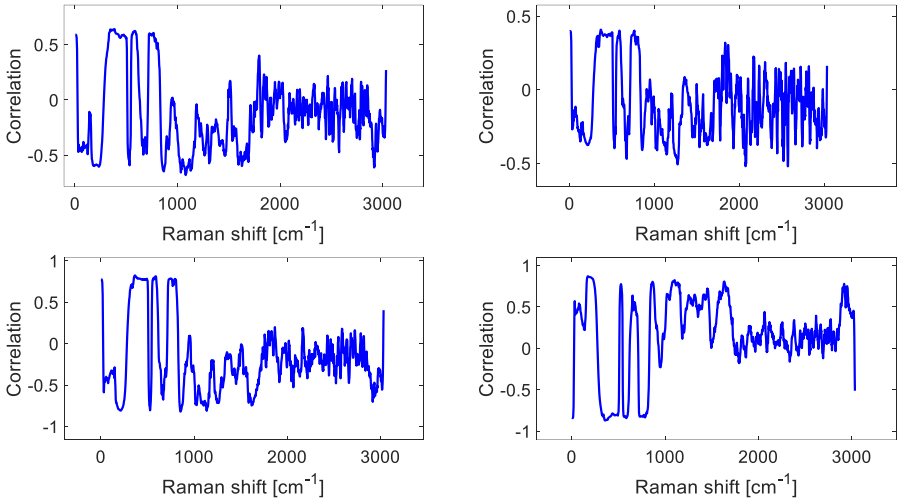


Figure 5.4 correlation plot for glucose (a), galactose (b), lactose (c), lactic acid (d) concentration with Raman shift intensities.

Table 5.1 summarizes the performance of the PLS models both using the whole set of Raman shifts and the high correlation windows for each variable of interest. Concerning glucose, it can be observed a slight improvement of the model's predictive ability when selecting a Raman shift window (177-878 cm^{-1}) that includes C-C-O, C-O-C and C-C-C deformation signals (Panthi et al., 2023), of which sugars like glucose are rich. Indeed, when the subset of Raman shifts was used for validating PLS model, Q^2 value increased by 4 (Q^2=0.47) and 27 (Q^2=0.50) percentage points in cross validation and external validation, respectively. A model is generally considered good when Q^2>0.5 (Golbraikh & Tropsha, 2002). As for galactose, less than acceptable performances were obtained both in calibration and cross-validation. This is probably due to the high overlap of its spectral bands with other key components (Chen et al., 2019). When PLS model was tasked to predict lactose concentration, high performances were achieved when using the entire set of Raman shifts, reaching Q^2>0.7. Finally, excellent prediction was obtained from PLS regression when estimating lactic acid (Q^2>0.85) and pH (Q^2>0.9) when trained using all the Raman shifts.

Table 5.1 Calibration, Cross-Validation and test results in estimating glucose, galactose, lactose, lactic acid concentration and pH from *Streptococcus Thermophilus* acidification using PLS regression with Raman spectra as predictor variable.

Response variable	Selected Variables	# Latent variables	R^2	Q^2_{CV}	Q^2_{EV}
Glucose	All	3	0.62	0.43	0.23
	177-878 cm^{-1}	3	0.61	0.47	0.50
Galactose	All	3	0.42	0.12	0.14
	1248-1283 cm^{-1}; 2074-2081 cm^{-1}; 2520-2574 cm^{-1}	2	0.37	0.31	0.44
Lactose	All	8	0.87	0.73	0.77
	157-891 cm^{-1}; 1009-1449 cm^{-1}; 1576-1692 cm^{-1}; 2898-3023 cm^{-1};	4	0.81	0.70	0.75
Lactic acid	All	11	0.93	0.85	0.91
	157-891 cm^{-1}; 1009-1449 cm^{-1}; 1576-1692 cm^{-1}; 2898-3023 cm^{-1};	11	0.91	0.90	0.81
pH	All	8	0.97	0.97	0.90
	157-902 cm^{-1}; 1009-1471 cm^{-1}; 1570-1700 cm^{-1}; 2879-3033 cm^{-1};	5	0.95	0.95	0.87

The graphical outcomes of analytes concentration prediction in external validation is reported in figure 5.5. The optimal combination of selected Raman shifts and number of latent variables were used to build the models. In all cases except for galactose, the predicted values the models capture the main variability of the experimental response. The dispersion observed around the regression line differs among analytes, with responses for lactose and lactic acid showing tight clustering and hence better predictions. The lower predictive performance observed for glucose and galactose can be explained by their similar Raman. In contrast, lactose and lactic acid exhibit more distinctive Raman features, providing the model with stronger and more specific spectral information. Moreover, the lower concentrations of such sugars in milk, along with the complexity of other protein structure whose signal overlaps with that of glucose and galactose have likely contributed to reduce the signal-to-noise ratio.

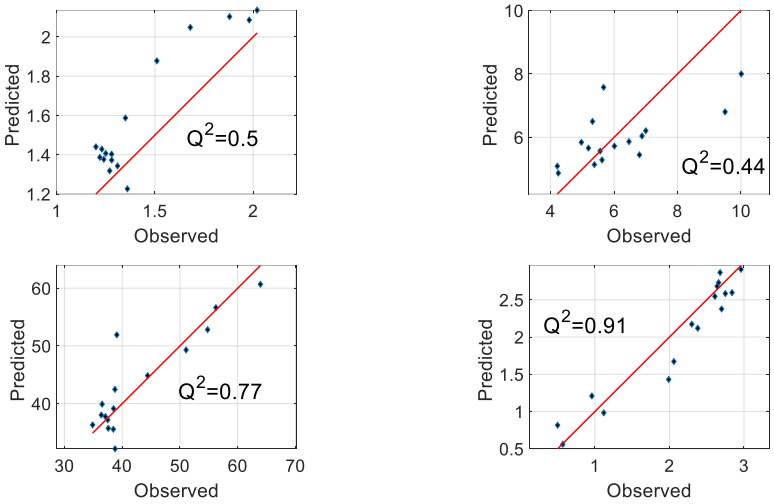


Figure 5.5 Parity plot from the application of PLS regression in external validation for glucose (a), galactose (b), lactose (c) and lactic acid (d) concentration.

Similarly, predicted vs observed plot were constructed for pH prediction in external validation, for which results are reported in figure 5.6. The high Q^2_{EV} value (0.9) aligns with that of lactic acid concentration prediction, as a result of the strong correlation between such two variables (pH decreasing is mainly induced by lactic acid production and to minor extent, with other organic acids). Such an estimation can be useful for technological purposes, such as predicting the acid cutting time, which is generally defined as the time required to reach the isoelectric point (pH=4.6-4.8) (Castillo, Lucey, & Payne, 2006).

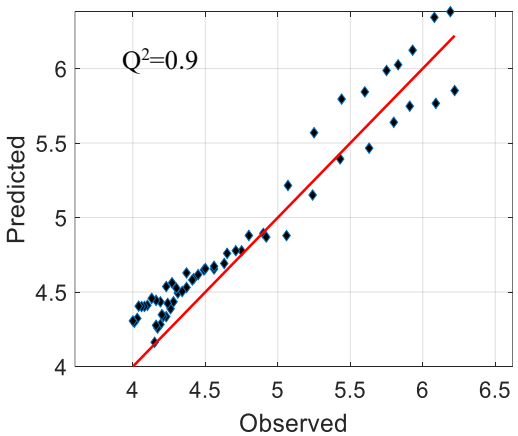


Figure 5.6 Parity plot from the application of PLS regression in external validation for pH.

The PLS models were then tested using external validation data sets for the computation of glucose, galactose, lactose and lactic acid. The temporal profile of analytes concentration measured by LC-MS/MS was compared with the corresponding estimated values (Figure 5.7). Validation results show that from Raman spectra, it is possible to satisfactorily predict the concentrations of glucose, lactose and lactic acid. For the lactose case, worse prediction was achieved in the first two hours of fermentation for T=40°C with a starter culture concentration of 0.175 g L^{-1} (Figure 5.7b). Some other noteworthy behavior can be highlighted from the dynamic profile of analytes' concentrations. At the beginning of the fermentation, significant concentrations of both glucose and galactose were detected. The initial presence of these monosaccharides can be attributed to residual amounts naturally occurring in the milk powder. As the process evolves, lactose consumption progresses until a critical pH range between 4.4 and 4.6 is reached. At this point, lactose degradation markedly slows down or ceases, a behavior that can be associated with the interruption of biomass production once the environment approaches a pH of 4.2-4.5 (Leroy & De Vuyst, 2004). Beyond this stage, only a limited utilization of glucose is observed, accompanied by continued lactic acid formation. A slight increase in galactose concentration was also recorded in some trials, which may be explained by the partial bacterial excretion of this sugar after lactose degradation, a phenomenon previously reported in certain Streptococcus thermophilus strains (Derkx et al., 2014). However, this trend was not consistently observed across all calibration assays, leaving its interpretation unclear. Further investigations are therefore needed to better elucidate this behavior, particularly when translating knowledge from well-mixed systems such as lactic acid bacteria cultivation (Taniguchi et al., 2004) to coagulating environments like curd formation.

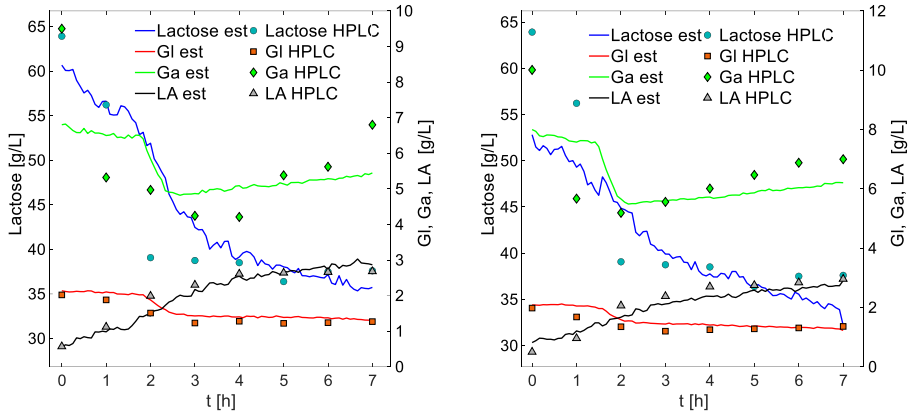


Figure 5.7 Comparison between LC-MS/MS measurements and PLS estimation using Raman sepctra for Lactose, Glucose (Gl), Galactose (Ga) and Lactic Acid (LA). Results refer to the test experiment conducted at T=40°C 2ith a starter culture concentration of 0.125 g L^{-1} (a) and T=40°C with a starter culture concentration of 0.175 g L^{-1} (b).

The same plot was derived for pH estimation, as reported in figure 5.8. The figure reports the comparison between estimated and experimentally measured pH values during the acidification process. In both panels, the predicted profiles closely follows the experimental data, albeit with some deviations at the early and late stages of acidification. The worse predictive performance observed for the test experiment conducted at a temperature of 40 °C and a starter concentration of 0.125 g L-1 attributed to the unexpectedly steeper pH decrease recorded under these conditions. Indeed, this experiment exhibited the lowest final pH (pH(t= 7 h) = 4), which was not represented in the calibration dataset and therefore not captured by the PLS model during training phase.

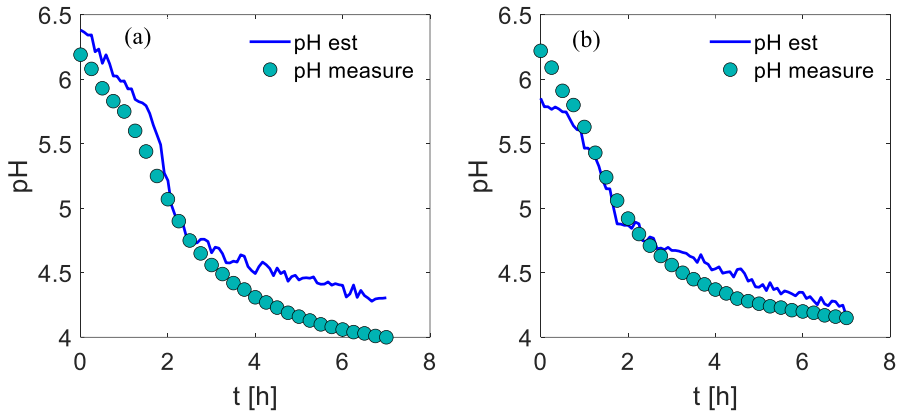


Figure 5.8 Comparison between measured pH and PLS estimation using the multivariate signal provided by Raman spectra. Results refer to the test experiment conducted at T=40°C with a starter culture concentration of 0.125 g L^{-1} (a) and T=40°C with a starter culture concentration of 0.175 g L^{-1} (b).

Generally, the prediction accuracy achieved for individual saccharides is lower compared to that obtained for lactic acid and pH. In particular, the test set results show that the models tend to underestimate the maximum lactose concentrations. This systematic deviation may be explained by the presence of spatial gradients in the reactor, which are likely to arise in the absence of mechanical mixing, which is often guaranteed in fermented products. Under such conditions, localized differences in substrate availability and metabolic activity can occur, leading to heterogeneities in the actual concentration profiles. As a result, discrepancies between the experimental values and the model predictions become more evident. Nonetheless, the proposed methodology offers exploratory information into galactose, while delivering acceptable predictions for glucose and lactose. Finally, particularly strong predictive performances were achieved for lactic acid and pH, which cover the most important role within the technological application.

5.4 Conclusions

The implementation of an inline/online monitoring system of key metabolites for milk coagulation offers multiple benefits within the dairy environment, including enhanced quality tracking, increased safety, and reduced operating costs. As outlined in previous chapters, this task is challenging. Raman spectroscopy was coupled with multivariate modeling to provide a real time determination of the concentrations of glucose, galactose, lactose, lactic acid, and pH during combined acid-rennet coagulation of milk by *Streptococcus Termophilus*. The applied method is experimentally straightforward as it requires no sample preparation, making it a promising candidate for industrial applications. The system investigated in this study could be developed for rapid analysis of the coagulating system composition (sugars and lactic acid) as well as technological parameters (pH). The use of multiple PLS regression models allowed to achieve satisfactorily responses prediction. From cross-validation and external validation procedures, acceptable models' performance were obtained for glucose, good predictive ability was obtained for lactose while high Q^2 were observed for lactic acid and pH. The selection of Raman shift windows as predictor variables for PLS models allowed to improve the goodness-of-prediction of monosaccharides concentration thanks to the region-specific signal response for such compounds. However, overlapping peaks of their spectra reduced model accuracy. On the other hand, the whole feature set was required to efficiently calibrate the PLS models when estimating lactose concentration, lactic acid concentration and pH. Although the method allows to monitor the concentration of key metabolites, its ability to fully capture the metabolic dynamics of the system remains a challenge due to the system's complexity. Further studies are required to extend the current knowledge of lactic acid fermentation to more complex problems such as curd formation, and implement it within the industrial frameworks.

Overall, the work presented in this chapter serves as a connecting element between the implementation of PAT tools, discussed in chapter 3 and 4 with the forthcoming metabolomic investigation, bridging the issue of dairy process monitoring with the exploration of the biochemical picture of cheesemaking. Indeed, the four metabolites described in this chapter showed significant variations in concentration across different thermal treatments and ripening times (Chapter 7), thereby demonstrating the interplay between the different process steps and the need of establishing an interconnected PAT pipeline throughout the cheesemaking process.

5.5 References

Abdelgawad, A. R., Guamis, B., & Castillo, M. (2014). Using a fiber optic sensor for cutting time prediction in cheese manufacture from a mixture of cow, sheep and goat milk. Journal of Food Engineering, 125(1), 157–168. https://doi.org/10.1016/j.jfoodeng.2013.10.001

Bouteille, R., Gaudet, M., Lecanu, B., & This, H. (2013). Monitoring lactic acid production during milk fermentation by in situ quantitative proton nuclear magnetic resonance spectroscopy. Journal of Dairy Science, 96(4), 2071–2080. https://doi.org/10.3168/jds.2012-6092

Castillo, M., Lucey, J. A., & Payne, F. A. (2006). The effect of temperature and inoculum concentration on rheological and light scatter properties of milk coagulated by a combination of bacterial fermentation and chymosin. Cottage cheese-type gels. International Dairy Journal, 16(2), 131–146. https://doi.org/10.1016/j.idairyj.2005.02.005

Castillo, M., Lucey, J. A., Wang, T., & Payne, F. A. (2006). Effect of temperature and inoculum concentration on gel microstructure, permeability and syneresis kinetics. Cottage cheese-type gels. International Dairy Journal, 16(2), 153–163. https://doi.org/10.1016/j.idairyj.2005.02.004

Chandrasekaran, K., Chandrasekaran, K., Jayaraman, G., & Bhatt, N. (2025). Model maintenance, monitoring, and control framework using In-Situ NIR spectroscopy and offline process data for Lactococcus lactis fermentation. Chemical Engineering Journal, 519. https://doi.org/10.1016/j.cej.2025.165116

Chen, D., Li, Y., Tan, Z., Huang, Z. xuan, Zong, J., & Li, Q. feng. (2019). On-line monitoring of key nutrients in yoghurt samples using digitally labelled Raman spectroscopy. International Dairy Journal, 96, 132–137. https://doi.org/10.1016/j.idairyj.2019.04.006

Czaja, T. P., Vickovic, D., Pedersen, S. J., Hougaard, A. B., & Ahrné, L. (2023). Spectroscopic characterisation of acidified milk powders. International Dairy Journal, 142. https://doi.org/10.1016/j.idairyj.2023.105664

De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. International Journal of Pharmaceutics, 417(1–2), 32–47. https://doi.org/10.1016/j.ijpharm.2010.12.012

Derkx, P. M. F., Janzen, T., Sørensen, K. I., Christensen, J. E., Stuer-Lauridsen, B., & Johansen, E. (2014). The art of strain improvement of industrial lactic acid bacteria without the use of recombinant DNA technology. Microbial Cell Factories, 13(1). https://doi.org/10.1186/1475-2859-13-S1-S5

Eilers, P. H. C. (2003). A perfect smoother. Analytical Chemistry, 75(14), 3631–3636. https://doi.org/10.1021/ac034173t

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! Journal of Molecular Graphics and Modelling, 20(4), 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1

Hayes, E., Greene, D., O'Donnell, C., O'Shea, N., & Fenelon, M. A. (2023). Spectroscopic technologies and data fusion: Applications for the dairy industry. Frontiers in Nutrition, 9. https://doi.org/10.3389/fnut.2022.1074688

Leroy, F., & De Vuyst, L. (2004). Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends in Food Science and Technology, 15(2), 67–78. https://doi.org/10.1016/j.tifs.2003.09.004

Panthi, R. R., Shibu, S. N., Ochalski, T. J., & O'Mahony, J. A. (2023). Raman spectra of micellar casein powders prepared with wet blending of glycomacropeptide and micellar casein concentrate. International Journal of Dairy Technology, 76(2), 429–435. https://doi.org/10.1111/1471-0307.12920

Rhiel, M., Ducommun, P., Bolzonella, I., Marison, I., & Von Stockar, U. (2002). Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. Biotechnology and Bioengineering, 77(2), 174–185. https://doi.org/10.1002/bit.10134

Rodriguez, R., Vargas, S., Estevez, M., Quintanilla, F., Trejo-Lopez, A., & Hernández-Martínez, A. R. (2013). Use of Raman spectroscopy to determine the kinetics of chemical transformation in yogurt production. Vibrational Spectroscopy, 68, 133–140. https://doi.org/10.1016/j.vibspec.2013.06.004

Sibono, L., Manis, C., Zucca, F., Atzori, L., Errico, M., Tronci, S., Casula, M., Dedola, A., Pes, M., Caboni, P., & Grosso, M. (2024). Metabolomic profiling of Fiore Sardo cheese: Investigation of the influence of thermal treatment and ripening time using univariate and multivariate classification techniques. Food Chemistry, 456. https://doi.org/10.1016/j.foodchem.2024.139930

Sibono, L., Tronci, S., Hajrizaj, R., Christensen, K. V., Errico, M., & Grosso, M. (2023). Optimization and kinetic analysis of untreated brewers' spent grain saccharification process via enzymatic hydrolysis. Biochemical Engineering Journal, 198. https://doi.org/10.1016/j.bej.2023.109044

Sivakesava, S., Irudayaraj, J., & Ali, D. (2001). Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and FT-Raman spectroscopic techniques. Process Biochemistry, 37(4), 371–378. https://doi.org/10.1016/S0032-9592(01)00223-0

Taniguchi, M., Tokunaga, T., Horiuchi, K., Hoshino, K., Sakai, K., & Tanaka, T. (2004). Production of L-lactic acid from a mixture of xylose and glucose by co-cultivation of lactic acid bacteria. Applied Microbiology and Biotechnology, 66(2), 160–165. https://doi.org/10.1007/s00253-004-1671-x

Tarapata, J., Lobacz, A., & Zulewska, J. (2021). Physicochemical properties of skim milk gels obtained by combined bacterial fermentation and renneting: Effect of incubation temperature at constant inoculum level. International Dairy Journal, 123. https://doi.org/10.1016/j.idairyj.2021.105167

Torrestiana, B. S., Brito de la Fuente, E., Lacroix, C., & Choplin, L. (1994). Modelling the acidifying activity profile of Lactobacillus bulgaricus cultures. Applied Microbiology and Biotechnology, 41(2), 192–196. https://doi.org/10.1007/BF00186958

Zhang, Z. Y., Sha, M., & Wang, H. Y. (2017). Laser perturbation two-dimensional correlation Raman spectroscopy for quality control of bovine colostrum products. Journal of Raman Spectroscopy, 48(8), 1111–1115. https://doi.org/10.1002/jrs.5179

Chapter 6: Missing data imputation and classification algorithms

In the present chapter, the problem of sample classification when missing data appear within the data matrix is undertaken. Particularly, different algorithms are proposed and validated in order to identify the best solution in terms of classification and prediction performances. In this section, the investigation of data retrieved from two different data sources allows to emphasize the benefits provided by multiple platforms when handling missing data. To this concern, Pecorino Romano cheese is presented as a case study. This chapter refers to a previously published work by the author (Sibono et al., 2023).

6.1 The problem of dairy food nutritional improvement and missing data imputation: the effect of cheese seasonality on cheese fatty acid and mineral profile as a case study

In the context of increasingly complex industrial processes characterized by several highly correlated variables, common univariate analysis approaches often fail to describe a system's characteristics comprehensively. The introduction of multivariate statistical analysis has received significant attention in several fields in the process industry, such as quality control, fault detection, process monitoring, process control, and predictive modelling (MacGregor and Kourti, 1995; Miletic et al., 2004). Within the framework of the food industry, increasing consumer awareness of food quality and safety has led to a growing demand for the development of healthier products (Mishra et al., 2022; Renes et al., 2019), thus driving the food sector toward the implementation of improved technological and manufacturing techniques that rely on multivariate information (Khattab et al., 2019). Specifically, dairy manufacturers are focused on monitoring and controlling the levels of lipids, proteins, minerals, and vitamins in their products, as these compounds strongly impact human health (Matera et al., 2018). Different studies have reported that sheep's milk cheese could have a beneficial effect on human health due to the presence of bioactive compounds such as conjugated linoleic acids (CLA, e.g., *C18:2 cis 9*, *trans 11*), vaccenic acid (*C18:1 trans 11*), and omega-3 family compounds, which play important roles as nutraceutical substances (Addis et al., 2015; Nudda et al., 2021). The concentration of these nutritional compounds in cheese products is known to vary significantly throughout different cheese production seasons owing to many factors, such as changes in animal feed fatty acid composition and lactation stages (Addis et al., 2015). Therefore, developing an effective classification tool for cheese samples to study the effect of a specific stimulus or treatment, such as seasonality, on the properties of the cheese is imperative for ensuring the control of product characteristics. To date, metabolomics has been a powerful discipline with respect to analyzing food products, especially dairy products, as it enables the employment of a variety of analytical techniques together with statistical tools, thus facilitating the assessment of product quality and safety for several food matrices (Caboni et al., 2019; Karlsson et al., 2022; Suh, 2022). However, some limitations may impair the quality of data retrieved from food matrices. Indeed, the failure of the measuring/sampling method or the unreliable peak value given by the spectrometer may origin of missing values within the data matrix (Hrydziuszko and Viant, 2012; Ispirova et al., 2020). Missing data are defined as missing entries into the data matrix (Hrydziuszko and Viant, 2012), which prevent one to use conventional

data analytics methodologies. There are three possible solution to handle missing data (Ispirova et al., 2020):

1) Remove the variable for which data is missing.
2) Replacing missing with plausible estimates such as mean or median obtained from the other samples within the same class for which the variable value is known.
3) Estimating missing data with imputation algorithms.

Discarding observations and/or variables with incomplete data would result in a loss of valuable information and hinder the prediction of characteristics that might otherwise be successfully inferred in an analysis (van Ginkel et al., 2020). Therefore, it is imperative to properly infer missing values when handling metabolomic data.

In the context of food industry, the challenge of managing missing values remains an issue of great concern across various food matrices, such as pasta and tea, where imputation algorithms are employed to deal with such an issue (Delaporte et al., 2019; Mercier et al., 2017). To this concern, Probabilistic Principal Component Analysis (PPCA) offers the possibility to reconstruct a data matrix wherever missing data are present while retaining a model parameter solution equivalent to that obtained from the conventional PCA (Nyamundanda et al., 2010). In addition, PPCA allows for the achievement of a more robust classification model thanks to its probabilistic approach, which enables a better assessment of the uncertainty in the resulting model response (Scano et al., 2021). Currently, the study of the effects of any cheese production stimuli (e.g. cheese seasonality) on the nutritional properties of the product while dealing with missing entries within data matrix appears to be lacking. The present study aims to expand the current advancements in food-omics by addressing the presence of completely at random missing values via investigating the influence of seasonality on the fatty acid and mineral profiles of Pecorino Romano cheese samples. PPCA coupled with Linear Discriminant Analysis (LDA) was first employed to classify cheese samples produced during winter (January), mid-spring (April), and early summer (June) and identify the variables responsible for such classifications. Then, the model was employed in reconstructing the missing information. The results obtained from PPCA were compared with those obtained from more consolidated approaches, namely, Partial Least Squares Discriminant Analysis (PLS-DA) in the case of a classification problem and Partial Least Squares Regression (PLSR) in the case of the inference of unknown data values. The robustness of the models was evaluated using a hybrid cross-external validation approach for classification and Montecarlo cross-validation for missing data inference.

6.2 Materials and methods

For this specific study, the analytical procedures presented in the materials and methods were carried out by external collaborators, whose contributions are acknowledged at the end of the dissertation. The author, however, focused on the metabolomic data analysis and the interpretation of the results.

6.2.1 Dataset structure

The dataset examined in this study was obtained from Pecorino Romano PDO cheese samples produced in 12 dairies situated in Sardinia (Italy) over three distinct months, namely, January, April, and June. The ripening time was 8 months. Once the percentage of Fatty Acid Methyl Esters (% FAME) and mineral concentration were experimentally determined, a 45 × 81 matrix was obtained. Specifically, 73 variables were related to % FAME, and the remaining 8 corresponded to minerals (Ca, Mg, Na, K, P, S, Zn, and Fe). A total of 45 samples were analyzed, of which 33 offered complete % FAME and mineral concentration information, while the remaining 12 lacked this information. In more detail, regarding January and April, there were 24 samples (12 per production month) for which both the % FAME and mineral concentration were known. Hereafter, these samples will be referred to as “complete samples”. In addition to these complete samples, three samples per month lacked % FAME information, and these samples constituted the missing data series of the experimental data matrix. Hereafter, these samples will be referred to as “partial samples”, as only their mineral content is known. Similarly, there are nine complete samples for June along with six additional partial samples. Thus, dataset is balanced for all three seasons. Further information regarding the dataset partitioning are reported in table 6.1.

6.2.2 Analytical methods

6.2.2.1 Cheese Composition and Nitrogen Fractions

Cheese samples were analyzed with respect to pH using pH meter (420 A, Orion, Boston, MA, USA), dry matter (DM) (ISO5534:2004, 2004), fat using the Soxhlet method (Soxhlet, 1879), total nitrogen (TN) (ISO8968-12014, 2014), protein (TN*6.38), soluble nitrogen (SN) at pH 4.6, soluble nitrogen in 12% trichloroacetic acid (SN-TCA), soluble nitrogen in 10% phosphotungstic acid (SN-PTA) (GRIPON

et al., 1975), NaCl (ISO5943:2006IDF88:2006, 2006), and ash. To determine the ash content, cheese samples (5 g) were pre-dried at 102 °C for 24 h and then calcined at 550 °C in a muffle furnace (Gelman Instrument, Opera, Italy). The operating conditions were set to reach 550 °C in 8 h and then maintain 550 °C for another 8 h (ISO5545IDF90:2008, 2008). In this work, the effect of the cheese production season on the macro-components was studied first.

6.2.2.2 Fatty Acid Methyl Ester Analysis

The fatty acid profiles of the cheese samples were determined according to the analytical procedure reported in (Addis et al., 2015). In brief, FAMEs were obtained through the basic trans-methylation of the cheese fat extracted according to the methodology reported by Jiang et al. (Jiang et al., 1996) and analyzed using a gas chromatograph coupled with a flame ionization detector (GC-FID). Each fatty acid methyl ester (FAME) was identified based on retention time and compared with a standard mixture of 37 pure components.

6.2.2.3 Elemental Analysis

Briefly, elemental analysis was conducted using an inductively coupled plasma optical emission spectrometry ICP-OES spectrometer (OPTIMA 7300 DV, Perkin Elmer, Waltham, MA, USA) a GemTip Cross-Flow II nebulizer (Perkin Elmer) equipped with an autosampler (SC-2 DX, Elemental Scientific Inc., Omaha, NE, USA). The detailed analytical procedure can be found in the study conducted by Lai et al. (Lai et al., 2023).

6.2.2.4 Chemometric Techniques

The first analysis to be carried out was the effect of the cheese production season on macro-components. The Tukey test was used to evaluate the statistical significance between the mean values of two production months for each macro-component. The selected threshold p-value for the test was $p < 0.05$. Fatty acid profiles and mineral content were analyzed from qualitative and quantitative perspectives. Concerning the classification task, in this study, an unsupervised method is compared with a supervised approach. PPCA is a statistical method that enables dimensionality reduction in high-dimensional datasets while accounting for noise and missing data (Tipping and Bishop, 1999). After performing PPCA, LDA was applied to the first two principal components to classify the data into three groups (one for each production season). LDA is used to find a linear combination of the principal components (PCs), thus constructing the boundaries where the Mahalanobis distances to the centroids of each class are equal (Brereton, 2009). Therefore, the linear boundaries maximize the separation between the groups. Once the sample scores were

computed and LDA was performed, the presented model was employed to classify cheese samples. Partial Least Squares Discriminant Analysis (PLS-DA) was then employed in order to carry out a comparison between the two classifiers' performances. After performing the classification, the estimation of % FAME from the mineral profile was performed using PPCA to assess its capability to manage missing data. PLS regression was also employed for comparative analysis in this investigation. It is important to specify that the PLS algorithm is unsuitable for handling missing data, meaning that the only variables completely characterized for each sample (i.e., those regarding mineral composition) were used as predictor variables for PLS-DA and PLS regression. A unity variance preprocessing step was conducted before executing the abovementioned algorithms.

6.2.2.5 Models' Validation

In order to properly validate the PPCA and PLS-DA models in classification, a Monte Carlo cross-validation (Xu and Liang, 2001) combined with external validation was employed, and can be described through the following 4 -step procedure:

Step 1. Creation of the calibration set. The set is created by randomly selecting approximately 85% of the available complete samples. Further partial samples are added into the calibration set to constitute approximately 15 to 35% (Depending on the production month, as June samples were richer in missing data than those of April and January) of the total calibration set.

Step 2. Validation set construction. The remaining complete samples from step 1 are included for model cross-validation. Additional partial samples are included in the validation set, comprising approximately 30% to 65% (Again, the percentage of partial samples depends on the production month) of the total validation set. These samples are exclusively used for external validation, meaning that they are not replaced in the calibration set.

Step 3. Calibrate the models. The entire calibration dataset is utilized to tune the PPCA and PLS-DA models.

Step 4. Calculate the percentage of correctly classified samples (%CC). The PPCA and PLS-DA models are tested through the validation samples, and their performance is assessed through the accuracy in classification.

The presented methodology is applied to select the optimal number of LVs for PPCA and PLS-DA. For each PCs number (or LVs number), the abovementioned procedure is executed 20 times, and

the models' performances are assessed by computing the average %CC obtained for validation samples across all iterations. The optimal number of PCs and LVs is determined as the one at which %CC was maximum. It is important to highlight that in this procedure, the additional validation partial samples are not replaced across iterations and are never used for model calibration, thereby resembling an external validation. Indeed, the models inferred the class of these external partial samples in each iteration, while the calibration set comprises distinct sample groups each time it is randomly constructed. As a result, a different score estimation for external validation samples is expected to be obtained from each iteration as well, allowing the quantification of parameter estimation sensitivity when a fraction of samples in the calibration set is replaced. Table 6.1 explains how the samples are distributed among the calibration and validation sets.

Table 6.1 Number of samples in the calibration and validation sets for each month of cheese production.

	January		April		June	
	Number of complete samples	Number of partial samples	Number of complete samples	Number of partial samples	Number of complete samples	Number of partial samples
Calibration set	10	2	10	2	8	4
validation set	2[a]	1[b]	2[a]	1[b]	1[a]	2[b]
Total	12	3	12	3	9	6

[a] Samples employed to construct the internal cross-validation set

[b] Samples employed to construct external validation set

Concerning the assessment of PPCA and PLS in the quantitative estimation of % FAME, a Monte Carlo cross-validation method was employed to evaluate these models' performances. Specifically, 20% of the complete samples were randomly selected for each production month and used for the internal validation step, while the remaining ones were retained for calibration (Xu and Liang, 2001). Once calibrated, the PPCA and PLS models were tested via their application to the validation samples by predicting the % FAME using the samples' mineral compositions. In both models, the

reliability of the % FAME prediction was assessed using the coefficient of determination in cross validation (Q^2) and the Root Mean Square of Errors in Cross-Validation (RMSECV). The entire procedure was carried out for a total of 20 iterations, and the overall Q^2 and RMSECV were calculated by averaging the values obtained in each iteration.

First, macro component data were analyzed using the statistical software Minitab 21.1 (Minitab Inc., State College, PA, USA, 2022). Then, % FAME and mineral concentration (%*w*/*w*) analyses were carried out in the MATLAB® environment. PPCA was carried out using the built-in functions provided by the Statistics toolbox, whereas validation was performed using an author-built routine. Conversely, LDA and PLS-DA were performed using the Classification Toolbox (Ballabio and Consonni, 2013).

6.3 Results and Discussion

6.3.1 Pecorino Romano PDO Cheese Macro Composition Analysis

Table 6.2 shows the physicochemical composition and proteolysis indices (SN/TN, SN-TCA/TN, and SN-PTA/TN) for Pecorino Romano PDO cheese produced in January, April, and June. Based on the Tukey test, it was determined that the month of production statistically affected the fat and protein content of the cheese. In this study, it was observed that the fat content in cheese tends to decrease from winter to spring; then, it exhibits a significant increase in cheeses during the early summer. The protein content in cheese tends to exhibit a complementary behavior with respect to fat content, increasing in the cheese produced from January to April and then decreasing in the cheese produced in early summer. As reported by other authors, this behavior is mainly caused by the natural variation in the fat/protein ratio in milk induced by the advancing production season (Guinee et al., 2007). The increase in the fat/protein ratio in sheep's milk and cheese as the season progresses is a typical phenomenon occurring in sheep's milk produced in Sardinia, and it is mainly linked to the stage of lactation, diet, and the rearing method practiced on the island. In particular, in the final stage of lactation, there is a progressive decrease in the volume of milk produced with a concentration of certain macro-components (particularly fat). Sheep's milk, produced in the summer period, is characterized by a high fat content, which is generally not compensated by an equal increase in protein content (Addis et al., 2015).

Table 6.2 Macro composition and proteolysis indices for Cheeses produced in three different seasons (Mean ± Standard Deviation).

Parameter	January	April	June
pH	$5.06^{a}\pm0.12$	$5.08^{a}\pm0.08$	$5.02^{a}\pm0.14$
SS (w/w %)	$68.20^{a}\pm1.59$	$68.76^{a}\pm1.26$	$68.62^{a}\pm1.20$
Fat (w/w %)	$33.90^{a}\pm1.50$	$32.75^{b}\pm1.01$	$34.93^{a}\pm1.50$
Protein (w/w %)	$25.19^{b}\pm0.96$	$26.40^{a}\pm0.82$	$24.42^{b}\pm0.97$
SN/TN	$14.02^{a}\pm2.23$	$13.51^{a}\pm2.09$	$13.81^{a}\pm1.86$
SN-TCA/ TN	$11.62^{a}\pm2.13$	$11.16^{a}\pm1.91$	$10.78^{a}\pm2.03$
SN-PTA/TN	$8.73^{a}\pm2.38$	$8.43^{a}\pm2.32$	$8.81^{a}\pm1.57$
NaCl (w/w %)	$4.44^{a}\pm0.98$	$4.51^{a}\pm0.73$	$4.68^{a}\pm0.98$
Ash (w/w %)	$7.20^{a}\pm0.93$	$7.41^{a}\pm0.71$	$7.21^{a}\pm0.95$

[a-b] Statistically different means in the same raw were annotated with different superscripts. SS: soluble solids; SN: soluble nitrogen; TCA: trichloroacetic acid; PTA: phosphotungstic acid

6.3.2 Multivariate Statistical Analysis on % FAME and Mineral Profiles

Many studies have investigated the influence of production season and reported its strong influence on the milk fatty acid profile and chemical composition of sheep milk products (Altomonte et al., 2019; Serrapica et al., 2020; Todaro et al., 2014). Indeed, it is well known that production season strongly affects the fatty acid composition of milk and, then of cheese, as a consequence of variations in animal feed, pasture availability, and the fatty acid composition of grass lipids (Addis et al., 2015; Verma et al., 2020). Similarly, seasonal changes in animal diet have been found to affect the mineral composition profile of cheese (Bontempo et al., 2019). Therefore, variations in % FAME and mineral concentration throughout the production season are expected to provide valuable information with which to develop a classifier that can be used to distinguish Pecorino Romano PDO cheese based on its manufacturing period. PPCA coupled with LDA was employed for such a task since PPCA is reliable in terms of reducing the complexity of a multivariate observed space and is efficient in reconstructing missing information. In contrast, LDA enables the discrimination of cheese samples based on the coordinates of the first two PPCA

scores in a two-dimensional latent space. The dataset investigated in this chapter is affected by the occurrence of missing completely at random data, because the unavailable information does not depend neither on the values of observable variables, nor on the value of the missing information itself (Sportisse et al., 2020). Figure 6.1 reports the score plot of the first two principal components and the results of LDA classification for the calibration, cross-validation, and external validation samples obtained in an iteration. Similar results were obtained in the other iterations. The first two PCs, which explained 47% and 13% of the total variance in the calibration set, were used to identify the main differences between samples.

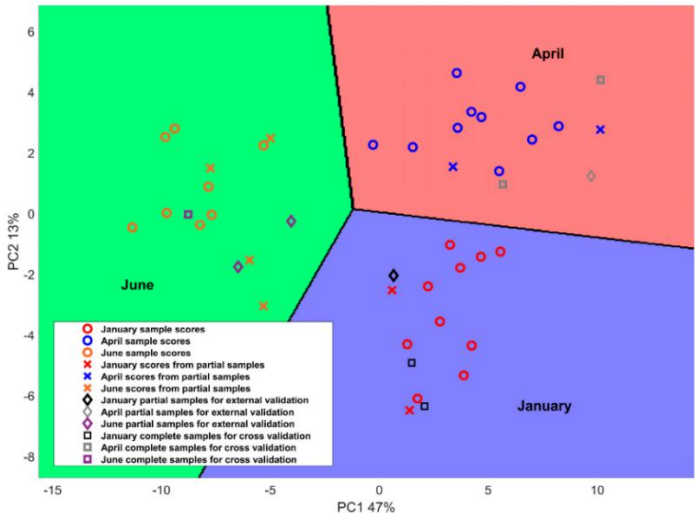


Figure 6.1 Score plot obtained from PPCA, and the relative boundaries obtained from LDA delimiting the regions to which the three different classes belong: Pecorino Romano produced in January, April, and June.

Notably, cheeses produced during the winter and spring seasons show positive scores regarding PC1, while negative scores were reported for early summer. Concerning PC2, cheeses produced in April exhibited positive scores, while negative scores were observed for January. Regarding June, both positive and negative PC2 score values are reported in the plot. Similar results were obtained by Nudda et al., who reported a multivariate statistical analysis of the FA content in cheese samples produced in different seasons (Nudda et al., 2021). The same authors merged the early spring with the winter season and the late spring with the summer season, thus obtaining a binary analysis

from the cheese seasonality point of view. The optimal number of PCs and LVs may vary depending on the overall dataset and the samples selected to calibrate the model. For this reason, cross-validation combined with external validation was employed to ensure the generalizability and reliability of the results. The results are reported in Figure 6.2.

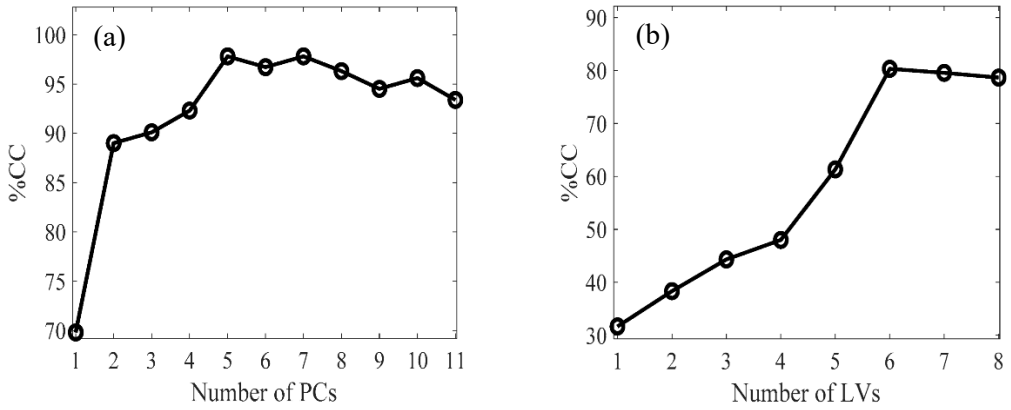


Figure 6.2 Validation results. The performance regarding PPCA is shown for each PC (a), while the results of PLS-DA are presented with respect to the number of LVs (b). The accuracy of classification is expressed as the percentage of correctly classified samples.

Reproduced from: Sibono, L.; Grosso, M.; Tronci, S.; Errico, M.; Addis, M.; Vacca, M.; Manis, C.; Caboni, P., Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction, Metabolites, 2023, 13, 877.

It is evident that the application of the PPCA + LDA method to % FAME and mineral information yielded good classification results over different production times Figure 6.2a. Indeed, the maximum %CC, equal to 98%, was obtained from PPCA + LDA for five principal components, explaining 77% of the total variance. The model demonstrates satisfactory accuracy and robustness when a moderate number of PCs are used. Regarding the PLS-DA, the highest accuracy was achieved when using six LVs, resulting in a %CC of 80%. Accordingly, it appears that introducing partial information regarding % FAME into a PPCA model significantly improves the performance of the DA classifier. Conversely, using only mineral information coupled with a PLS-DA model leads to lower, yet still acceptable, performance accuracy in classification (Figure 6.2b).

6.3.3 PPCA Loading analysis

In order to assess the impact of cheese production seasons on the content of individual fatty acids (FAs) and minerals, an analysis of loadings derived from PPCA and the identification of the most important variables with respect to constructing the first two PCs are reported in this section. The results of this analysis can provide valuable insights into which variables have the greatest impact on the observed variability in data. Particular attention has been paid to specific fatty acids, namely, rumenic acid, vaccenic acid, and omega-3 compounds, owing to their beneficial effects on human health. Figure 6.3 shows the biplot obtained using the PPCA for the calibration dataset.

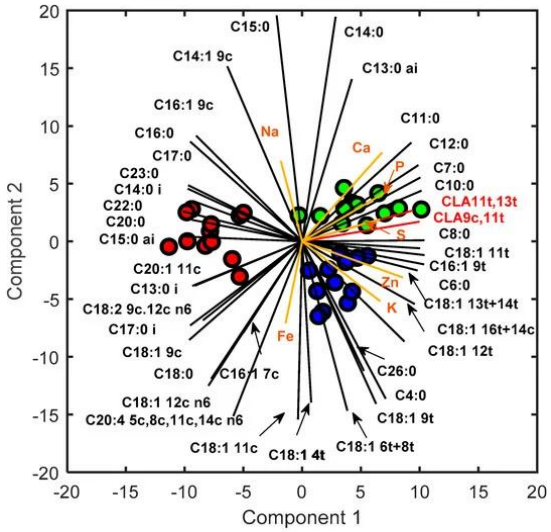


Figure 6.3 PPCA biplot: Scores related to Pecorino Romano samples produced in January, April, and June are reported in blue, green, and red, respectively. Loadings related to the metabolites are reported in black, whereas those related to minerals are in orange.

Reproduced from: Sibono, L.; Grosso, M.; Tronci, S.; Errico, M.; Addis, M.; Vacca, M.; Manis, C.; Caboni, P., Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction, Metabolites, 2023, 13, 877.

Regarding the % FAME analysis, both PC1 and PC2 displayed positive loadings for rumenic acid, vaccenic acid, and omega3 compounds, such as alpha-linolenic acid (C18:3 9c,12c,15c n3) and eicosapentaenoic acid (C20:5 5c,8c,11c,14c,17c n3), indicating a positive correlation with April. Furthermore, high negative PC1 loadings were observed with regard to oleic acid, resulting in the determination of its positive correlation with June. Notably, these results are in line with those found in a previous studies, even with respect to examining different food matrices like cow milk

(Addis et al., 2015; Lopez et al., 2014). These findings can be explained via the combined effect of pasture quality and lactation stage (Caboni et al., 2019). In particular, the abundance of CLA and omega-3 compounds in cheese produced in the spring season is strongly associated with the wide availability, in the grazed pasture, of their precursor (C18:3 9c,12c,15c n3) during the grazing season (Addis et al., 2015). In contrast, the increase in oleic acid during the late lactation stages may be due to the energy requirement induced by the worsening of pasture quality (Nudda et al., 2021).

Saturated fatty acids play a crucial role in defining the nutritional characteristics of food. Short-chain saturated FAs are strongly correlated with January; medium-chain saturated FAs are positively correlated with April. On the other hand, June is characterized by an abundance of long-chain saturated FAs. These loading trends highlight a clear dependence of the length of the saturated FA chain on the production season, which is probably due to the abovementioned reasons. Regarding the mineral loading analysis, S, P, and Ca showed positive loadings for both PC1 and PC2, indicating a positive correlation between these minerals and April as the month of cheese production, likely owing to the abundance of minerals during spring when the pasture quality is high. The higher presence of protein (caseins) in the cheese produced in April (Table 6.2) compared to the other periods could explain the high presence in the spring cheese of Ca and P, as they are key elements in the formation of the cheese protein matrix. K and Zn display positive PC1 and negative PC2 loadings, suggesting a positive correlation with January. For Fe, a weak dependence on the winter season was observed, while Mg did not strongly contribute to constructing the first two principal components as its loading values were low. Interestingly, Na loadings were negative in the first PC and positive in the second one, indicating that Na content is prevalent in June. A simple explanation for this result is that cheeses produced in June show a tendentially, though not significantly, higher salt content than the other cheeses, as previously reported by other authors (Addis et al., 2015). Regarding the mineral profile behavior, similar results were obtained in other studies (Altomonte et al., 2019; Dar et al., 2014; Li et al., 2022).

6.3.4 Missing data reconstruction

The PPCA model's effectiveness in reconstructing missing % FAME data was evaluated by comparing the estimated values with their corresponding experimental values (With the samples for which %FAME was known) from the cross-validation samples. During each iteration, % FAME

information was removed from the validation samples, estimated by the model, and residuals were computed. Figure 6.4 compares the RMSECV for two models: PLS and PPCA. The x-axis shows the number of PCs/LVs involved in each model, while the y-axis shows the RMSECV values. According to the graph, PPCA performed better than PLS when a smaller number of components is considered. On the other hand, as the number of model parameters increases, the performance of PLS appears to improve, while PPCA continues to perform well. With six components, the performance of PLS regression is slightly superior to that of PPCA, resulting in a lower RMSE. This slight difference is maintained at higher dimensions.

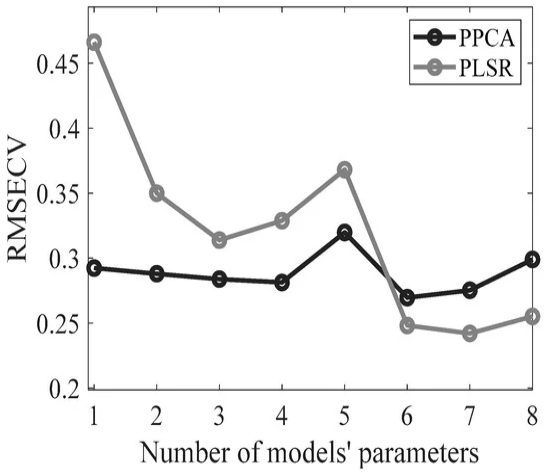


Figure 6.4 Root mean square error of cross-validation for PPCA and PLS models while varying the number of PCs and LVs, respectively.

Reproduced from: Sibono, L.; Grosso, M.; Tronci, S.; Errico, M.; Addis, M.; Vacca, M.; Manis, C.; Caboni, P., Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction, Metabolites, 2023, 13, 877.

Cross-validation analysis showed that the best data fit for the PPCA model was obtained with six PCs, resulting in a Q^2 value of 99.43%. On the other hand, the optimal PLS model was set with seven LVs, for which a Q^2 value of 99.53% was obtained. Thus, the performance of the models is comparable. It is worth noting that the presented PPCA model, like any model suitable for handling missing data, can be used as a classification and inferential tool (García-Laencina et al., 2010). Indeed, the same model can simultaneously provide informative score values that enable the classification of cheese samples and the reconstruction of the whole data matrix, thereby facilitating the prediction of missing values. Specifically, when the optimal number of PCs obtained

in classification (i.e., 5 PCs) is employed to reconstruct data using PPCA, the result is an Q^2 value of 99.20% and an RMSE of 0.32 (Figure 6.4) meaning that good performance is maintained regarding missing data inference. In contrast, the PLS approach requires two different models, namely, PLS-DA and PLSR, to accomplish classification and inference tasks, respectively. In general, PPCA preserves high prediction performance, even when implemented with a parsimonious number of parameters, whereas PLSR achieves superior performances only when model complexity is relatively high due to the involvement of several latent variables. Furthermore, when implementing a PPCA model, the computational cost may drastically increase as the number of PCs, even when the dataset is relatively small. Hence, the ability of the PPCA model to retain good performance with a limited number of PCs may suggest such a probabilistic model as a potential candidate for addressing this kind of problem. However, it important to acknowledge that PPCA is performed under the assumption of linear relationships between observed and latent variables, which might render it inappropriate when the data-generating mechanism involves nonlinear dependencies, which can be common in biological and biochemical processes. Nonetheless, such a linear assumption resulted not to impair the quality of the classification techniques.

6.4 Conclusions

The application of a probabilistic model such as PPCA facilitates the monitoring of food quality when there are missing values in experimental observations. As reported herein, a comparative study was employed in order to assess the benefits offered by partial knowledge of the investigated variables when using a probabilistic model compared to a well-established multivariate model such as PLS. The case study reported in this work concerns the examination of the effect of seasonality on the fatty acid and mineral profiles for Pecorino Romano cheese, aiming to identify the key variables capable of discriminating cheese samples based on their production month. For classification purposes, such models were coupled with discriminant analysis algorithms, and Montecarlo cross-validation combined with external validation was employed to assess the classifiers' reliability. The results showed that both models could effectively perform classification tasks, although PPCA showed particularly remarkable performance. Then, the ability of PPCA to handle missing data was exploited to reconstruct missing % FAME information, and PPCA was then compared with PLS regression, for which % FAME values were fitted by using only mineral compositions as predictor variables. The results showed that both models offer high predictive performance in the estimation of % FAME. Overall, it can be concluded that PPCA may

be a valuable modelling strategy, as it simultaneously provides a dimensionality reduction strategy couplable with classification tools while computing unknown information related to the presence of missing values, which is particularly common in the food industry.

6.5 References

Addis, M., Fiori, M., Riu, G., Pes, M., Salvatore, E., Pirisi, A., 2015. Physico-chemical characteristics and acidic profile of PDO Pecorino Romano cheese: Seasonal variation. Small Rumin. Res. 126, 73–79. https://doi.org/10.1016/j.smallrumres.2015.03.014

Altomonte, I., Conte, G., Serra, A., Mele, M., Cannizzo, L., Salari, F., Martini, M., 2019. Nutritional characteristics and volatile components of sheep milk products during two grazing seasons. Small Rumin. Res. 180, 41–49. https://doi.org/10.1016/j.smallrumres.2019.10.003

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods 5, 3790–3798. https://doi.org/10.1039/c3ay40582f

Bontempo, L., Barbero, A., Bertoldi, D., Camin, F., Larcher, R., Perini, M., Sepulcri, A., Zicarelli, L., Piasentier, E., 2019. Isotopic and elemental profiles of Mediterranean buffalo milk and cheese and authentication of Mozzarella di Bufala Campana PDO: An initial exploratory study. Food Chem. 285, 316–323. https://doi.org/10.1016/j.foodchem.2019.01.160

Brereton, R.G., 2009. Chemometrics for Pattern Recognition, Chemometrics for Pattern Recognition. Wiley.

Caboni, P., Murgia, A., Porcu, A., Manis, C., Ibba, I., Contu, M., Scano, P., 2019. A metabolomics comparison between sheep's and goat's milk. Food Res. Int. 119, 869–875. https://doi.org/10.1016/j.foodres.2018.10.071

Dar, A.A., Jadhav, R.K., Dimri, D., Khan, A.A., Khan, H.M., Sharm, M.C., 2014. Effects of physiological status and seasonal variation on plasma mineral profile of sheep in Kashmir valley. Sci. Res. Essays 9, 69–76. https://doi.org/10.5897/sre2013.5694

Delaporte, G., Cladière, M., Camel, V., 2019. Missing value imputation and data cleaning in untargeted food chemical safety assessment by LC-HRMS. Chemom. Intell. Lab. Syst. 188, 54–62. https://doi.org/10.1016/j.chemolab.2019.03.005

García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: A review. Neural Comput. Appl. 19, 263–282. https://doi.org/10.1007/s00521-009-0295-6

GRIPON, J.C., DESMAZEAUD, M.J., LE BARS, D., BERGERE, J.L., 1975. Etude du rôle des micro-organismes et des enzymes au cours de la maturation des fromages. II. - Influence de la présure commerciale. Lait 55, 502–516. https://doi.org/10.1051/lait:197554828

Guinee, T.P., Mulholland, E.O., Kelly, J., Callaghan, D.J.O., 2007. Effect of protein-to-fat ratio of milk on the composition, manufacturing efficiency, and yield of cheddar cheese. J. Dairy Sci. 90, 110–123. https://doi.org/10.3168/jds.S0022-0302(07)72613-9

Hrydziuszko, O., Viant, M.R., 2012. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. Metabolomics 8, 161–174. https://doi.org/10.1007/s11306-011-0366-4

ISO5534:2004, 2004. Cheese and Processed Cheese—Determination of the Total Solids Content. ISO Geneva, Switz.

ISO5545IDF90:2008, 2008. Rennet Caseins and Caseinates, Determination of Ash. ISO Geneva, Switz.

ISO5943:2006IDF88:2006, 2006. Cheese and Processed Cheese Products—Determination of Chloride Content—Potentiometric Titration Method. ISO Geneva, Switz.

ISO8968-12014, 2014. IDF Milk and Milk Products—Determination of Nitrogen Content—Part 1: Kjeldahl Principle and Crude Protein Calculation. ISOGeneva, Switz.

Ispirova, G., Eftimov, T., Seljak, B.K., 2020. Evaluating missing value imputation methods for food composition databases. Food Chem. Toxicol. 141. https://doi.org/10.1016/j.fct.2020.111368

Jiang, J., Bjoerck, L., Fondén, R., Emanuelson, M., 1996. Occurrence of Conjugated Cis-9,Trans-11-Octadecadienoic Acid in Bovine Milk: Effects of Feed and Dietary Regimen. J. Dairy Sci. 79, 438–445. https://doi.org/10.3168/jds.S0022-0302(96)76383-X

Karlsson, T., Winkvist, A., Rådjursöga, M., Ellegård, L., Pedersen, A., Lindqvist, H.M., 2022. Identification of Single and Combined Serum Metabolites Associated with Food Intake. Metabolites 12. https://doi.org/10.3390/metabo12100908

Khattab, A.R., Guirguis, H.A., Tawfik, S.M., Farag, M.A., 2019. Cheese ripening: A review on modern technologies towards flavor enhancement, process acceleration and improved quality assessment. Trends Food Sci. Technol. 88, 343–360. https://doi.org/10.1016/j.tifs.2019.03.009

Lai, G., Caboni, P., Piras, C., Pes, M., Sitzia, M., Addis, M., Pirisi, A., Scano, P., 2023. Development and Chemico-Physical Characterization of Ovine Milk-Based Ingredients for Infant Formulae. Appl. Sci. 13. https://doi.org/10.3390/app13010653

Li, S., Delger, M., Dave, A., Singh, H., Ye, A., 2022. Seasonal Variations in the Composition and Physicochemical Characteristics of Sheep and Goat Milks. Foods 11. https://doi.org/10.3390/foods11121737

Lopez, C., Briard-Bion, V., Ménard, O., 2014. Polar lipids, sphingomyelin and long-chain unsaturated fatty acids from the milk fat globule membrane are increased in milks produced by cows fed fresh pasture based diet during spring. Food Res. Int. 58, 59–68. https://doi.org/10.1016/j.foodres.2014.01.049

MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. Control Eng. Pract. 3, 403–414. https://doi.org/10.1016/0967-0661(95)00014-L

Matera, J., Luna, A.S., Batista, D.B., Pimentel, T.C., Moraes, J., Kamimura, B.A., Ferreira, M.V.S., Silva, H.L.A., Mathias, S.P., Esmerino, E.A., Freitas, M.Q., Raices, R.S.L., Quitério, S.L., Sant'Ana, A.S., Silva, M.C., Cruz, A.G., 2018. Brazilian cheeses: A survey covering physicochemical characteristics, mineral content, fatty acid profile and volatile compounds. Food Res. Int. 108, 18–26. https://doi.org/10.1016/j.foodres.2018.03.014

Mercier, S., Mondor, M., Marcos, B., Moresoli, C., Villeneuve, S., 2017. Estimation of missing values in a food property database by matrix completion using PCA-based approaches. Chemom. Intell. Lab. Syst. 166, 37–48. https://doi.org/10.1016/j.chemolab.2017.04.008

Miletic, I., Quinn, S., Dudzic, M., Vaculik, V., Champagne, M., 2004. An industrial perspective on implementing on-line applications of multivariate statistics. J. Process Control 14, 821–836. https://doi.org/10.1016/j.jprocont.2004.02.001

Mishra, B., Mishra, A.K., Kumar, S., Mandal, S.K., Lakshmayya, N.S.V., Kumar, V., Baek, K.H., Mohanta, Y.K., 2022. Antifungal metabolites as food bio-preservative: Innovation, outlook, and challenges. Metabolites 12. https://doi.org/10.3390/metabo12010012

Nudda, A., Correddu, F., Cesarani, A., Pulina, G., Battacone, G., 2021. Functional Odd- and Branched-Chain Fatty Acid in Sheep and Goat Milk and Cheeses. Dairy 2, 79–89. https://doi.org/10.3390/dairy2010008

Nyamundanda, G., Brennan, L., Gormley, I.C., 2010. Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 11. https://doi.org/10.1186/1471-2105-11-571

Renes, E., Gómez-Cortés, P., de la Fuente, M.A., Linares, D.M., Tornadijo, M.E., Fresno, J.M., 2019. CLA-producing adjunct cultures improve the nutritional value of sheep cheese fat. Food Res. Int. 116, 819–826. https://doi.org/10.1016/j.foodres.2018.09.016

Scano, E.A., Grosso, M., Pistis, A., Carboni, G., Cocco, D., 2021. An in-depth analysis of biogas production from locally agro-industrial by-products and residues. An Italian case. Renew. Energy 179, 308–318. https://doi.org/10.1016/j.renene.2021.07.050

Serrapica, F., Masucci, F., Di Francia, A., Napolitano, F., Braghieri, A., Esposito, G., Romano, R., 2020. Seasonal variation of chemical composition, fatty acid profile, and sensory properties of a mountain pecorino cheese. Foods 9. https://doi.org/10.3390/foods9081091

Sibono, L., Grosso, M., Tronci, S., Errico, M., Addis, M., Vacca, M., Manis, C., Caboni, P., 2023. Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction. Metabolites 13. https://doi.org/10.3390/metabo13070877

Soxhlet, F., 1879. Die gewichtsanalytische Bestimmung des Milchfettes. Dinglers Polytech. J 232, 461–465.

Sportisse, A., Boyer, C. & Josse, J. Imputation and low-rank estimation with Missing Not At Random data. Stat Comput 30, 1629–1643 (2020). https://doi.org/10.1007/s11222-020-09963-5

Suh, J.H., 2022. Critical review: Metabolomics in dairy science – Evaluation of milk and milk product quality. Food Res. Int. 154. https://doi.org/10.1016/j.foodres.2022.110984

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B Stat. Methodol. 61, 611–622. https://doi.org/10.1111/1467-9868.00196

Todaro, M., Bonanno, A., Scatassa, M.L., 2014. The quality of Valle del Belice sheep's milk and cheese produced in the hot summer season in Sicily. Dairy Sci. Technol. 94, 225–239. https://doi.org/10.1007/s13594-013-0155-0

van Ginkel, J.R., Linting, M., Rippe, R.C.A., van der Voort, A., 2020. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. J. Pers. Assess. 102, 297–308. https://doi.org/10.1080/00223891.2018.1530680

Verma, A., Meitei, N.S., Gajbhiye, P.U., Raftery, M.J., Ambatipudi, K., 2020. Comparative analysis of milk triglycerides profile between jaffarabadi buffalo and holstein friesian cow. Metabolites 10, 1–27. https://doi.org/10.3390/metabo10120507

Xu, Q.S., Liang, Y.Z., 2001. Monte Carlo cross validation. Chemom. Intell. Lab. Syst. 56, 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Chapter 7: Untargeted metabolomics for dairy food biomarkers identification

Within the cheese making process, thermal treatment and ripened cheese stand as a pretreatment and final product manufacturing stages, respectively. Such processes are strongly intertwined in defining the quality of cheese. In this section, different algorithms are explored to assess the impact of two key process factors (i.e. thermal treatment and ripening time) on the metabolic composition of cheese. Both univariate and multivariate approaches are applied to provide a broader perspective on food data analysis. This article refers to a previously published work by the author (Sibono et al., 2024).

7.1 Overview of metabolomic-based approach in the dairy framework

Due to its complexity, cheese characteristics are influenced by a multitude of biochemical processes. The understanding of such effects is mandatory within the industrial framework, as they embrace food quality, safety and sensory properties of the food matrix. This means that the development and application of strategies that can accurately predict cheese characteristics would be highly beneficial to the dairy industry (Afshari et al., 2021). In this context, Metabolomics-based methodologies for cheese analysis have expanded significantly over the last decades (García-Pérez et al., 2024). For example, untargeted metabolomics based on UHPLC-Orbitrap-HRMS was used to discriminate the metabolomics profiles of Parmigiano Reggiano PDO cheeses obtained from different feeding strategies and geographical regions (Becchi et al., 2023). In another study, GC/TOF-MS-based metabolomics was used for sensometric applications (Ochi et al., 2012). Concerning the cheese-safety, Gas Chromatography-Mass Spectrometry (GC-MS) was also used to investigate the spoilage potential of different bacterial species (Jaffrès et al., 2011). Ergo, metabolomics has the potential to facilitate the identification and quantification of biomarkers associated with cheese typification, enabling improved quality control and traceability within the dairy industry.

7.2 Overview of the effect of milk thermal treatment and cheese ripening time on final product quality

The Food and Agriculture Organization of the United Nations (FAO) estimated that 6 billion people worldwide consume milk and milk products (FAO, 2023; OECD/FAO, 2021). Milk and milk products are an important source of protein, calcium, and other essential nutrients (Silva et al., 2017). They are also a major source of calories and energy for many people. There is an undeniable relationship between cheese and territory in terms of culture and recognized typical products. Such a relationship is covered by PDO label. Therefore, in order to produce high-quality food products that meet the legal requirements and consumer demands, it is imperative to implement advanced technological and monitoring methods within the cheese-making industry (Khattab et al., 2019) while preserving regional traditions associated with the product and its production technologies. The analysis of cheese flavor and safety includes hundreds of different chemical compounds produced by food microbiota metabolisms such as biogenic amines (BA), fatty acids, amino acids (AA) and sugars, to name a few (Hassan et al., 2012; Khattab et al., 2019; Sibono et al., 2023). Among the several variables that influence the concentration of such metabolites, cheese ripening time and milk heat treatment are considered to be the primary factors in the development of cheese characteristics (Atasoy & Türkoğlu, 2008; Combarros-Fuertes et al., 2016; Santiago-López et al., 2018). It has been broadly reported that ripening time has a strong effect on cheese microbial metabolic reactions such as amino acid decarboxylation (Benkerroum, 2016; Caboni et al., 2019a; Kandasamy et al., 2021) due to the advancement of the proteolysis effect (Santiago-López et al., 2018), and fat degradation phenomena associated with lipolytic enzymes (Collins et al., 2003). Consequently, cheese ripening is responsible for determining the extent of accumulation of amino acids, some of which serve as precursors for the synthesis of BA, such as tyramine, histamine, putrescine, and cadaverine (Kandasamy et al., 2021; Moniente et al., 2023). Excessive intake of these compounds is thought detrimental to human health. (Combarros-Fuertes et al., 2016; del Rio et al., 2024; Ferrante & Mercogliano, 2023). Concerning milk heat treatment, some studies analyzed the effect of the thermization process on cheese properties, identifying the key features that discriminate heat-treated milk cheese from raw milk cheese (Anedda et al., 2021; Caboni et al., 2019b; Dedola et al., 2020; Natrella et al. 2023). In particular, Caboni, et al. (2019b) studied the polar low-molecular-weight metabolites profile to differentiate raw milk cheese from cheese whose milk was previously treated at 68 °C for 30 s. However, the analysis of the joint effect given by milk thermization and cheese ripening time is still to be explored. Moreover, milk thermization

can be carried out within a relatively broad operative window that ranges from the mildest process of 57 °C for 15 seconds to the most severe thermal treatment of 68 °C for 30 seconds, meaning that further analysis within the thermal treatment stage have to be conducted. In this framework, the application of GC-MS to metabolomics has proven to be highly effective in analyzing dairy food quality, allowing for a widespread identification of biomarkers associated with a specific product attribute (Putri et al., 2022). Moreover, it is recognized that statistical analysis of metabolomic data is very challenging due to the nature of the data and the complexity of metabolomics as a research field.

7.3 Investigation of the influence of thermal treatment and ripening time on cheese metabolic profile: Fiore Sardo cheese as a case study.

In this work, the effect of heat treatment and ripening time on the metabolite profile of Fiore Sardo cheese was studied by means of GC-MS analysis coupled with untargeted metabolomics approaches. The ultimate goal was to develop classification tools capable of discriminating between cheeses subjected to different ripening times (105 and 180 days) and different thermization processes (non-thermized milk, milk thermized under mild conditions, and milk thermized under severe conditions). It should be noted that PDO Fiore Sardo is produced from raw milk, since thermal treatment is not allowed under the traditional production rules. This investigation was approached from a comparative perspective, as it evaluates the performance of univariate classifiers compared to the widely employed Partial Least Squares Discriminant Analysis (PLS-DA). The methodology introduced here offer a contribute for identifying cheese products that comply with PDO specifications.

7.3.1 Materials and Methods

For this specific study, the analytical procedures presented in the materials and methods were carried out by external collaborators, whose contributions are acknowledged at the end of the manuscript. The author focused on the metabolomic data analysis and the interpretation of the results. The experimental protocol is presented anyway.

7.3.1.1 Chemicals

Analytical LC grade (99.9% of purity) methanol, chloroform, and hexane as well as pyridine, methoxamine hydrochloride, potassium chloride, N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) were purchased from Sigma Aldrich (Milan, Italy). Bi-distilled water was obtained with a MilliQ purification system (Millipore, Milan, Italy). Analytical standards were used to prepare a QA samples. Succinic acid, 2,2,3,3-D4-succinic acid, malic acid, citric acid, aspartic acid, pyruvic acid, ketoglutaric acid, oxalacetic acid, fumaric acid, lactic acid, butyric acid, beta-hydroxybutyric acid, glycolic acid, myoinositol, scylloinositol, ribitol, alanine, glycine, glutamine, leucine, isoleucine, serine, homoserine, ornithine, phenylalanine threonine, proline, valine, glucose, fructose, xylitol, ribitol and phosphate were purchased from Sigma Aldrich (Milan, Italy).

7.3.1.2 Cheese Production and Sampling

Cheese samples of this study were produced from bulk milk from mature Sarda sheep collected from February 2019 to June 2019, leading to a total of 45 cheese-making trials from the experimental farm of Agris Sardegna and transformed into the Fiore Sardo type cheese. For each cheese-making trial, 130 kg of bulk raw sheep's milk were divided into 3 aliquots: 31 kg raw milk (RM), 32 kg low thermized milk (LTM) and 32 kg of high thermized milk (HTM). The eventual heat treatment occurred in a batch process using a tubular infrared exchanger (model Stoutz-Actinator, ACTINI Group, Evian-les-Bains, France), where the milk was heated to 57 °C for LTM and 68 °C for HTM for 30 seconds, respectively, and then rapidly cooled to 38 °C. The processing diagram is reported in Figure 7.1. The process was repeated under the same conditions, to obtain 3 replicates for each milk thermal treatment, within three separate days each month, in the Agris Sardegna pilot plant. In each cheese-making session day, three vats were used in batch: one with RM, one with LTM and one with the HTM alternately. The cheesemaker was the same for all cheese making trials, in order to reduce the experimental variability. After each process, approximately 1.5-1.7 kg were produced. At the end of the experimental campaign, 31 raw milk cheeses (RC), 32 cheeses at 57 °C (HCA) and 32 cheeses at 68 °C (HCB) were obtained, half of which were analyzed after 105

days of ripening (16 RC cheeses, 16 HCA cheeses and 16 HCB cheeses), while the other half after 180 days (15 RC cheeses, 16 HCA cheeses and 16 HCB cheeses).

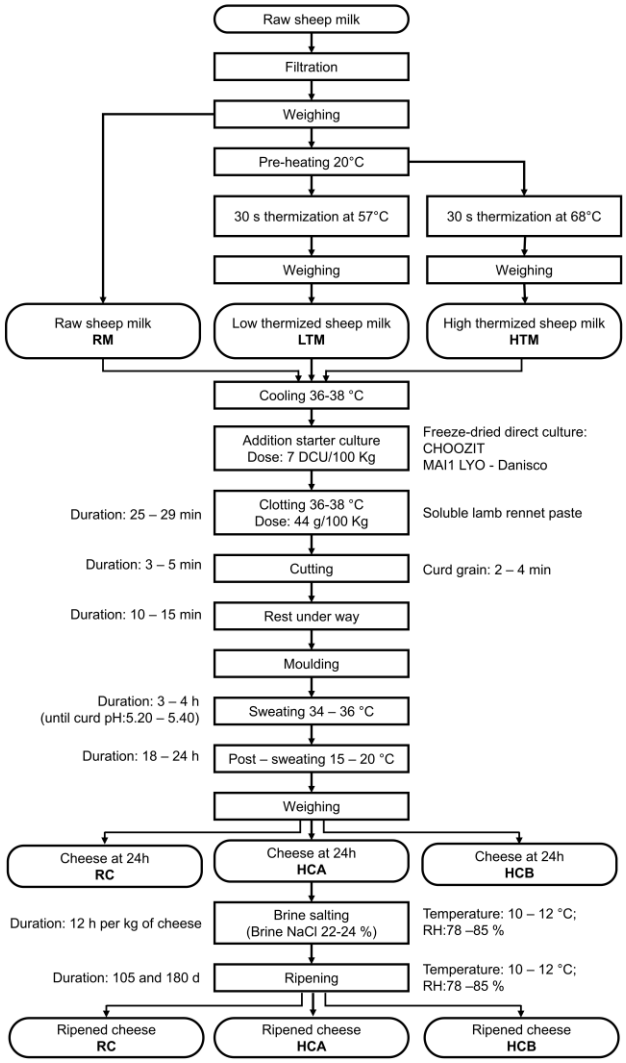


Figure 7.1 Fiore Sardo cheese production flowchart. RC = cheese from raw milk; HCA = cheese from thermized milk at 57 °C; HCB = cheese from thermized milk at 68 °C.

Reproduced from: Leonardo Sibono, Cristina Manis, Francesca Zucca, Luigi Atzori, Massimiliano Errico, Stefania Tronci, Mattia Casula, Alessio Dedola, Massimo Pes, Pierluigi Caboni, Massimiliano Grosso, Metabolomic profiling of Fiore Sardo cheese: Investigation of the influence of thermal treatment and ripening time using univariate and multivariate classification techniques, Food Chemistry, Volume 456, 2024, 139930.

7.3.1.3 Sample preparation and GC-MS Analysis

Fine shredded cheese samples were stored in sterile plastic Falcon tubes at −20 °C before analysis. After thawing on ice, 50 mg of cheese were extracted with 250 µL of methanol, and 125 µL of chloroform (Caboni et al., 2016). Samples were sonicated at 4 °C for 15 min at 25 kHz, through three cycles of 5 min each using ExstractorOne (GM solution, Cagliari, Italy). Subsequently, 380 µL of chloroform and 90 µL of aqueous 0.2 M KCl containing 5 mg/L of 2,2,3,3-D4-succinic acid as the internal standard, were added and finally vortexed for 10 s. The solution obtained was centrifuged at 17700 rcf for 10 min, and 0.200 mL of the aqueous layer was extracted from each sample and dried under a gentle nitrogen stream. The dried samples were derivatized first using 50 µL of methoxyamine hydrochloride dissolved in pyridine at 10 mg/mL, homogenized for 20 s, and kept at room temperature for 17 h. Then, 100 µL of N-methyl-N-(trimethylsilyl)trifluoroacetamide was added to the mixture and vortexed. Following a 1-hour incubation period, 600 µL of hexane was added, and the samples were subjected to GC-MS analysis.

One microliter of derivatized samples was injected splitless in a 6850-gas chromatograph coupled with a 5973 mass spectrometer (Agilent Technologies Inc., Santa Clara, CA). The injector temperature was set to 200 °C. The gas flow rate through the column was 1mL/min. The fused silica capillary column was a DB5-MS 0.25 µm column (30 m × 0.25 mm i.d.; J&W Scientific Inc., Folsom, CA). The initial temperature was 50 °C for 3 minutes in isothermal mode, which was then increased to 250 at 3 °C/min and maintained at 250 °C for 25 min. The transfer line and the ion source temperatures were 280 and 180 °C, respectively. Ions were generated at 70 eV with electron ionization and were recorded at 1.6 scans/s in the mass range m/z 50 to 550. GC-MS data analysis was conducted by integrating each peak of the resolved chromatogram. The identification of metabolites was performed using the NIST08 standard mass spectra library (NIST, 2005), and, when available, compared to analytical standards. Confirmation of metabolites was performed by comparison of their relative retention times and mass fragmentation as well as retention indices as calculated according to Kováts for alkanes C9−C36 (Kováts, 1958). In the untargeted metabolomics pipeline, authentic standards submitted to the same silylation procedure of cheese samples were used. Furthermore, QA and QC samples were employed to improve the integrity and robustness of our datasets while obviating the need for repeated analysis. Spectral peaks associated with potentially formed adducts (M+TMS, M+2TMS, M+3TMS, M+4TMS, M+5TMS were examined for each metabolite using Chromeleon software 7.2.8 (Themofisher, Massachusetts, USA). When multiple adducts were detected, the most reproducible (technical replicate CV) MS1 adduct was

selected for annotation and peak area integration. Metabolite reproducibility was then checked by evaluating consistency across replicate experiments; metabolites which are not consistent across replicates were labelled as irreproducible.

7.3.1.4 Dataset description

The metabolic dataset consists of 95 cheese samples and 745 GC-MS features. Cheese samples were classified based on different heat treatments and different ripening times. Among these samples, 31 were obtained from Raw Cheese (RC) milk samples (i.e. not thermally treated), 32 were obtained from 57 °C heat-treated milk (HCA), and 32 were subjected to a 68 °C thermal treatment (HCB).

7.3.1.5 Data pre-analysis

PCA was carried out on unity variance preprocessed data to reduce the complexity of the high-dimensional dataset, assess the trends among samples and eventually identify outliers (Varmuza & Filzmoser, 2016). The captured variance V_C, explained by a specific PC, was computed using components' eigenvalues expressed as percentages relative to the total variance (Brereton, 2009). Before the classification step, a two-way ANOVA test was used on raw data to identify the metabolites contributing to sample discrimination based on heat treatment and ripening time. The following independent variables were considered in the analysis: ripening time, denoted by letter A, and milk thermal treatment, indicated by letter B, with a number of levels equal to *a* and *b*, respectively. The GC-MS experimental observations obtained for each combination of these variables were denoted as x_{ijk}, where "*i*" represents the i-th level of the A factor, "j" represents the j-th level of the B factor, and "*k*" represents the k-th replication of the experiment. The effect model can be thus expressed as (Montgomery, 2013):

$$x_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1,2 \dots a \\ j = 1,2 \dots b \\ k = 1,2 \dots n \end{cases} \tag{7.1}$$

In Eq. 7.1, μ is the overall mean, τ_i is the effect of the i-th level of the independent variable A, γ_j is the effect of the j-th level of the independent variable B, $(\tau\gamma)_{ij}$ is the interaction effect of the i-th level of A and the j-th level of B, ε_{ijk} is the residual for the ijk observation. To be rigorous, the two-

way ANOVA tests the null hypothesis, which concerns the equality to zero of treatment effects with respect to the factors investigated. The level of significance α was set equal to 0.01. The procedure was performed for each metabolite, meaning that the effect of ripening time, thermal treatment and the interaction of both factors was tested on all the 745 spectral peaks collected. The ANOVA test revealed a set of metabolites found to be statistically significant in distinguishing between the ripening time and the thermal treatment of cheese samples. It is essential to note that, in situations involving multiple comparisons, the risk of committing Type I errors drastically increases (Lee & Lee, 2018). A more stringent approach was thus used to address this concern and control the inflation of false positives. More in detail, to identify the metabolites that are mostly responsible for variations in ripening, Benjamini-Hochberg method, also known as the False Discovery Rate (FDR; (Benjamini & Hochberg, 1995)) correction was employed for a FDR value equal to 0.01. This technique provided the opportunity to undertake a further elimination process with the objective of identifying the most significant metabolites that distinguish cheese samples based on their ripening times and thermal treatments.

7.3.1.6 Classification tools

For the sake of comparative analysis, data were analyzed by means of a univariate and a multivariate approach, and the performances were assessed in terms of accuracy and False Positive Rate (FPR). Firstly, a code employing univariate statistics was developed as a two-class classifier that identifies cheese samples' maturation time and thermal treatment. Specifically, the implemented code uses a t-test-based classification method to classify samples with a pairwise approach (Liao & Akritas, 2007). Hence, samples subjected to the same thermal treatment were binarily classified depending on the two ripening times. In contrast, for samples aged for the same ripening time, all possible pairs of thermal treatments among the three investigated were analogously tested. A detailed explanation of such a technique is reported Chapter 2, section 1, subsection 3. Besides the development of a univariate classifier, PLS-DA is the multivariate counterpart proposed for comparative purposes (Wold et al., 2001). A pairwise classification approach is also adopted in PLS-DA for an unbiased comparison between the two models. Regarding the thermal treatment, each pair of classes is tested separately, meaning samples subjected to pairwise discrimination are labelled 1 for one class and 2 for the other. Regarding ripening time, the value 1 was assigned to cheese ripened for 105 days, whereas a value of 2 was assigned to cheese ripened for 180 days. The Bayesian assignation criterion collocated samples in a specific class (Ballabio & Consonni, 2013). Models' goodness of classification was assessed through

3-fold cross-validation and averaging performance indices (i.e. Accuracy and FPR). In order to reduce the variance of the estimated performance measure, cross-validation was repeated 100 times with different 3-fold subsets at each time (Berrar, 2018).

The optimal number of PLS-DA latent variables was defined as the number that satisfies the following criteria:

1) The error rate in cross-validation was minimized (i.e. the accuracy was maximized);

2) The proportion of samples not assigned by the Bayesian criteria was less than 20% to ensure an unbiased assessment of the model's accuracy.

Univariate analysis and PCA were performed through Matlab®2023a (Mathworks Inc. Natick, MA, USA). The univariate procedure was carried out by means of codes developed by the authors, and "Milano Chemometrics" toolbox (version 6.0) was involved in performing PLS-DA (Ballabio & Consonni, 2013).

7.3.2 Results and Discussion

7.3.2.1 GC-MS data pre-analysis

Principal Component Analysis was initially performed to explore dissimilarities and similarities among 95 cheese samples. A 99.99% confidence ellipsoid was employed in the 3D score plot (Figure 7.2), which illustrates the first three principal components, in order to identify any potential outliers within the data. After outliers removal, the resulting dataset was a 94 x 745 matrix.

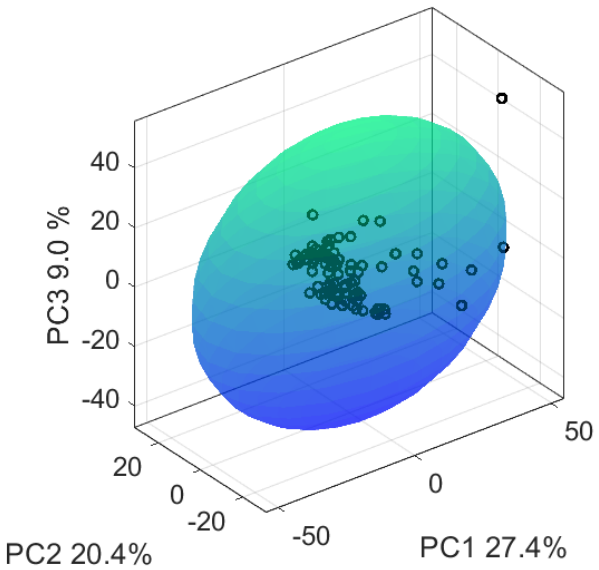


Figure 7.2 3D-scoreplot for the whole experimental data.

A two-way ANOVA test was performed on 745 GC-MS features to assess the impact of a specific factor, comprising the interaction term, on the individual metabolite. The hypothesis tests showed that 42 of these features underwent a statistically significant variation due to the effect of at least one of the two investigated factors. Specifically, 28 features were found to be influenced by the cheese ripening time. At the same time, 16 features were affected by the heat treatment. Two metabolites were simultaneously affected by both factors. The contribution of the interaction

term was remarkably lower since only 1 metabolite manifested a significant effect of the interaction term, which showed p-values smaller than 0.01. A Venn diagram is provided in Figure 7.3, displaying all the metabolites that have shown a statistically significant effect ($p<0.01$) from both factors, including the interaction effect. Such figure is supplemented with the metabolites identification number, listed in table 7.1. Metabolite 301 (arachidonoyl amide) resulted as the most important metabolite from the ripening time analysis, although it was also found to change significantly over the interaction term. Analogously, metabolite 1 (tyramine) resulted as the one that mostly changed over the thermal treatments, although the effect of ripening time was also found to be statistically significant.

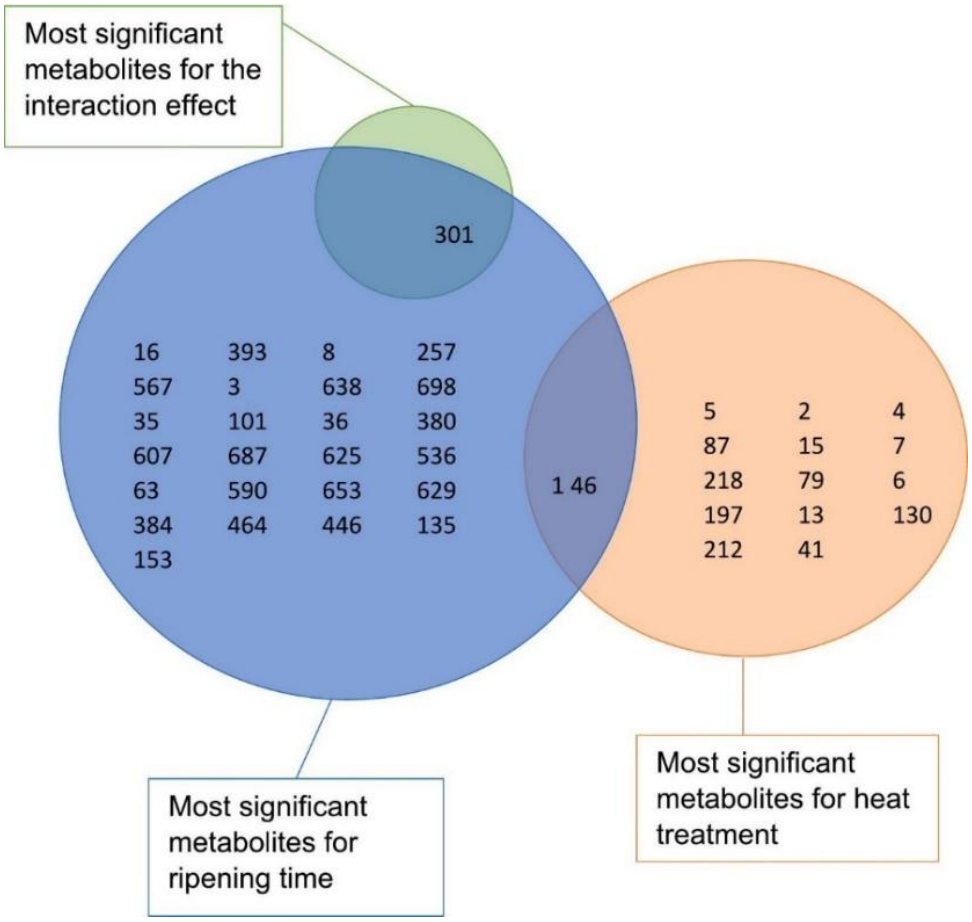


Figure 7.3 Venn diagram displaying the metabolites for which the individual factors and interaction effect resulted to be significant ($p<0.01$) from the 2-way ANOVA test. The metabolite number is shown in ascendent order in the p-value.

Here follows the metabolites identification number which can aid for the reading of Venn diagram provided above.

Table 7.1 List of the metabolites detected with the two-way ANOVA test with a significance level a=0.01.

Label	Factors	Metabolite	Label	Factors	Metabolite
1	A and B	Tyramine	197	B	**Unknown**
2	B	Cadaverine	212	B	Lactose 8TMS
3	A	4-aminobutanoic acid	218	B	Lactose 8TMS
4	B	Phenylalanine	257	A	Proline
5	B	Malic acid	301	A and AB	Arachidonoyl amide
6	B	**Unknown**	380	A	Myo-Inositol 6TMS
		Methyl glycocholate			
7	B	3TMS	384	A	Valine
8	A	Citric acid	393	A	Phenylpropanoic acid
13	B	**Unknown**	446	A	Myo-Inositol
15	B	Leucine	464	A	Citric acid 2TMS
16	A	Glucose	536	A	Galactose
35	A	Erythrose	567	A	Lactic acid
36	A	**Unknown**	590	A	**Unknown**
41	B	Glutamine	607	A	**Unknown**
46	A and B	Urea	625	A	**Unknown**
		4-aminobutanoic acid 3			**Unknown**
63	A	TMS	629	A	
79	B	Urea 2TMS	638	A	Succinic acid
87	B	Lactose	653	A	Maltose 8TMS
101	A	D-Fructose	687	A	**Unknown**
130	B	Glycerol 3TMS	698	A	**Unknown**
135	A	Turanose			
153	A	**Unknown**			

** Factor A is the ripening time; Factor B is the thermal treatment; the term AB represents the interaction between the two factors.*

A Benjamini-Hochberg method was used to select the significant metabolites while simultaneously maintaining small type I errors. The false discovery rate was set equal to 0.01. Table 7.2 presents the metabolites and their p-values derived from the two-way ANOVA whose null hypothesis was rejected using the Benjamini-Hochberg method. Concerning maturation time, 8 metabolites were not discarded by the FDR correction, while only one metabolite was retained when the correction method assessed the effect of heat treatment. Finally, the interaction term was discarded for all metabolites.

Interestingly, it was obtained that 4-aminobutanoic remarkably changed over different ripening times. It is reported that cheese starters produce 4-aminobutanoic acid during ripening (Nomura et al., 1998). Particularly, Nomura et al. reported that a cell extract of a Lactic Acid Bacterium (LAB) from the *Lactococcus lactis* strain, showed a significant glutamate decarboxylase activity, which catalyzes the decarboxylation of glutamate into 4-aminobutanoic acid. Therefore, it can be inferred that during the 75-day ripening period between cheeses subjected to 105 days of maturation and those ripened for 180 days, the decarboxylation reactions may have been a key factor in the production of 4-aminobutanoic acid. Also, the level of tyramine was identified to change significantly with temperature, indicating that thermal milk treatment significantly influences the level of this metabolite. More in detail, the analysis revealed that the tyramine level in raw milk cheeses was moderate and importantly increased in cheeses made from milk heated at 57 °C (HCA). Conversely, a marked decrease was observed in the tyramine content of HCB samples. These findings are consistent with the research conducted by Kebary et al., who found a similar trend when comparing raw milk cheese to cheese made from milk heated to 70 °C, 75 °C, and 80 °C (Kebary et al., 1999). The increased tyramine content in the lower-temperature heated milk cheese may be due to the favorable conditions created by such heat treatments for the growth of proteolytic bacteria and/or the activation of proteolytic enzymes (Kebary et al., 1999). Conversely, higher temperatures may have reduced proteolytic bacteria concentration and denatured decarboxylase enzymes, resulting in lower tyramine levels. Similarly, according to the ANOVA test, cadaverine, the second most significant metabolite in terms of thermal treatment effect, decreased abruptly with HCB. However, there was no significant difference between raw milk cheese and cheese made from 57 °C heated milk for such a metabolite. Noteworthy is that arachidonoyl amide exhibited the lowest p-value for ripening time analysis. Recent works reported that several fermented food matrices like cheese are rich in endocannabinoid metabolites (e.g., arachidonoyl amide) due to the action of microbial activity (Yılmaz et al., 2023). Other compounds

like long chain-carboxylic acids and proline are known to vary over different ripening times (Conte et al., 2021; Becchi et al., 2023; Tekin & Güler, 2019). While the formers are known to show complex behaviour along different ripening times (Caboni et al., 2019b), other studies reported the upregulation of proline in cheeses subjected to prolonged ripening times (Becchi et al., 2023; Singh et al., 2023).

Table 7.2 Results of the Benjamini-Hochberg procedure, showing metabolite codes and corresponding p-values.

Ripening time		Thermal treatment	
Metabolite	p-value	Metabolite	p-value
(4-aminobutanoic acid)	0.0014	Tyramine	8.734e-05
Citric acid	0.0005		
Glucose	0.0003		
Urea	0.0011		
Proline	0.0009		
Arachidonoyl amide	1.581e-05		
Phenylpropanoic acid	0.0005		
Lactic acid	0.0014		

The application of both univariate and multivariate classification models (reported in the following subsection) to classify cheese samples and identify the metabolites accountable for such classification allowed to validate the significance of the compounds shown in Table 7.2. The employment of ANOVA coupled with the False Discovery Rate (FDR) correction served as an initial screening process for identifying a subset of metabolites considered relevant to this study.

7.3.2.2 Classification of cheese samples

Following the detection of relevant metabolites, two different classification techniques were applied to discriminate cheese samples according to the ripening time and thermal treatment. As long as the univariate classifier tests the difference between two means, a pairwise approach was adopted during analysis, meaning that comparisons were made between ripening time classes

within the same thermal treatment level and vice versa. This process was performed for each metabolite, and the ones that exhibited the highest accuracy are documented in Table 7.3, which presents the classification metrics obtained through 3-fold cross-validation. Specifically, the metabolites that yielded the best performance metrics when subjected to t-test classification, as well as groups of metabolites whose accuracy values surpassed a threshold ranging from 70% to 80%, are reported. Notably, the employment of the t-test classifier demonstrated good efficacy in classifying samples based on ripening time, as accuracy was greater than 75 % in all cases. On the other hand, results show that classification performance was less favourable in terms of thermal treatment, particularly in comparison between RC and HCA, indicating that the average peak values between these classes were not considerably different given the same ripening time. Indeed, no relevant metabolites (except for maltose) resulted in good candidates for discerning between these thermal treatments, regardless of the ripening time. A milder treatment temperature likely led to fewer noticeable differences in metabolites content with respect to the HCB case. It is worth noting that the tyramine based t-test classifier was found to be the most effective for discerning HCB from the other thermal treatments. Concerning the ripening time analysis, results show that acceptable performances (Maroco et al., 2011) were obtained for arachidonoyl amide and phenylpropanoic acid, for which the effect of different ripening times was statistically significant from the two-way ANOVA test and whose null hypothesis was also rejected following FDR correction. These outcomes validate the involvement of such metabolites in cheese ripening process. Furthermore, the 4-aminobutanoic acid also showed a notable result in discriminating HCB180 from HCB105, thus confirming the importance of this compound during cheese ripening. These results confirm the findings identified with the ANOVA tests and FDR corrections carried out by means of the Benjamini-Hochberg method.

Overall, the findings reported in this analysis indicate that the metabolites considered as discriminant for ripening time were mainly fatty acids, interestingly supplemented by one endocannabinoid, while biogenic amines such as tyramine, accompanied by two soluble sugars (i.e., maltose and lactose) resulted in predominantly discriminating samples on a thermization process basis.

Table 7.3 Univariate classifier performance metrics.

Classes	Metabolites		Univariate classifier cross validation indices		
	Most performant metabolite	Performant metabolites	Accuracy given by performant metabolites	Best Accuracy[a]	Best FPR
RC180 vs RC105	Arachidonoyl amide	Arachidonoyl amide	>75%	76%	14%
HCA180 vs HCA105	Phenylpropanoic acid	Phenylpropanoic acid	>80%	81%	19%
HCB180 vs HCB105	4-aminobutanoic acid	4-aminobutanoic acid, Glutamic acid, Erythrose, Hydroxycaproic acid, Norleucine, Hydroxyisocaproic acid, Norvaline, Hydroxyisocaproic acid 2, U1	>75%	80%	26%
RC180 vs HCA180	U2	U2	>70%	73%	25%
RC180 vs HCB180	Tyramine	Tyramine, Lactic acid, U3, Glutamine, Urea, U5, Lactose	>70%	73%	32%
HCA180 vs HCB180	Tyramine	Tyramine	>75%	77%	32%

RC105 vs HCA105	Maltose	Maltose	>70%	72%	25%
RC105 vs HCB105	Glycilglutamic acid	Malic acid, Glycilglutamic acid	>75%	80%	27%
HCA105 vs HCB105	U6	U6,U7	>75%	75%	19%

[a] Accuracy correspondent to the metabolite whose univariate classification resulted in the best performance. U = not annotated metabolite. Frequently, mass spectrometry-based metabolic profiles encompass numerous peaks that remain unidentified, making it challenging to correlate them with the metabolic physiology of the system under investigation (Kanani et al., 2008).

The performance metrics were also evaluated using PLS-DA for sample classification. In this case, the impact of ripening time and thermal treatment on the metabolite content was not analyzed on a single-metabolite basis, but embraced the multivariate dataset. The results of this analysis are presented in Table 7.4, which reports the classification metrics obtained through a 3-fold cross-validation coupled with a PLS-DA model.

Table 7.4 PLS-DA performance metrics in cross validation.

Classes	Multivariate classifier Cross validation indices		Optimal number of latent variables
	FPR	Accuracy	
RC180 vs RC105	25%	73%	3
HCA180 vs HCA105	11%	88%	4
HCB180 vs HCB105	18%	83%	8
RC180 vs HCA180	25%	59%	7
RC180 vs HCB180	36%	62%	7
HCA180 vs HCB180	35%	66%	3
RC105 vs HCA105	31%	69%	2
RC105 vs HCB105	25%	74%	10
HCA105 vs HCB105	29%	72%	1

A comparison between tables 7.3 and 7.4 illustrates that the PLS-DA classifier generally performs less accurately than the univariate classifier when assigned to discriminate cheese samples based on heat treatment. On the other hand, higher performances were recorded for ripening time, especially for samples taken from heat-treated milk (i.e., HCA and HCB). Overall, the univariate classifier performed better at classifying cheese samples. This result may be due to the presence of several uninformative metabolites, or the scarce correlation among the metabolites content (Saccenti et al., 2014). To support the occurrence of these phenomena, the correlation between all pairs of variables was examined using a heatmap (Figure 7.4a), complemented with the distribution of correlation coefficients (Figure 7.4b). From this analysis, it was obtained that only 17.6% of the examined correlation coefficients showed absolute values greater than 0.5, suggesting that the overall behavior of the variables exhibits low correlation. This outcome suggests the potential appearance of uninformative variables which have led to data overfitting, meaning that the use of a univariate classifier may serve as a good candidate for feature selection

purposes. Similar classification accuracies to those reported in this study were reported by other authors (Azcarate et al., 2015; Teixido-Orries et al., 2023).

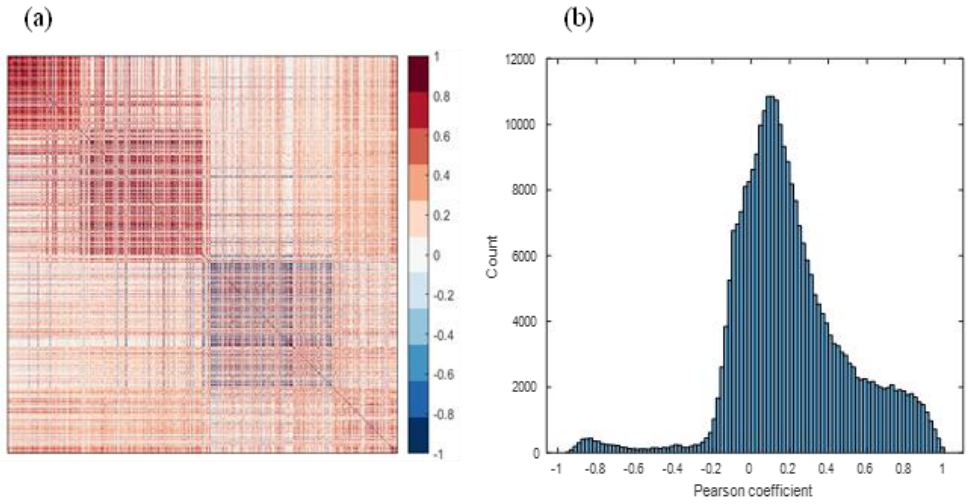


Figure 7.4 Heatmap of Pearson correlation coefficients (a) along with their distributions for the GC-MS data of each couple of variables (b).

By way of example, Figure 7.5 displays the algorithm iterations performed to determine the optimal number of latent variables for PLS-DA in the discrimination between HCA180 and HCA105 samples. As depicted, the selection of 10 latent variables resulted in the minimum error rate in cross-validation. However, it was noted that about 50% of the samples were classified as unassigned. In order to balance the trade-off between error rate minimization and the number of unassigned samples, four latent variables were selected as part of the model. For the samples analyzed in this study, utilizing the collective information from the entire metabolic profile may not be sufficient in accurately classifying samples from differently heated milk with the same ripening time. However, PLS-DA classification performances were adequate when the classifier was asked to distinguish samples based on ripening time with the same thermal treatment. These results suggest that ripening time exerts a more evident influence on metabolites profile under the same thermal conditions, thus offering more details into the dynamics of the ripening process.

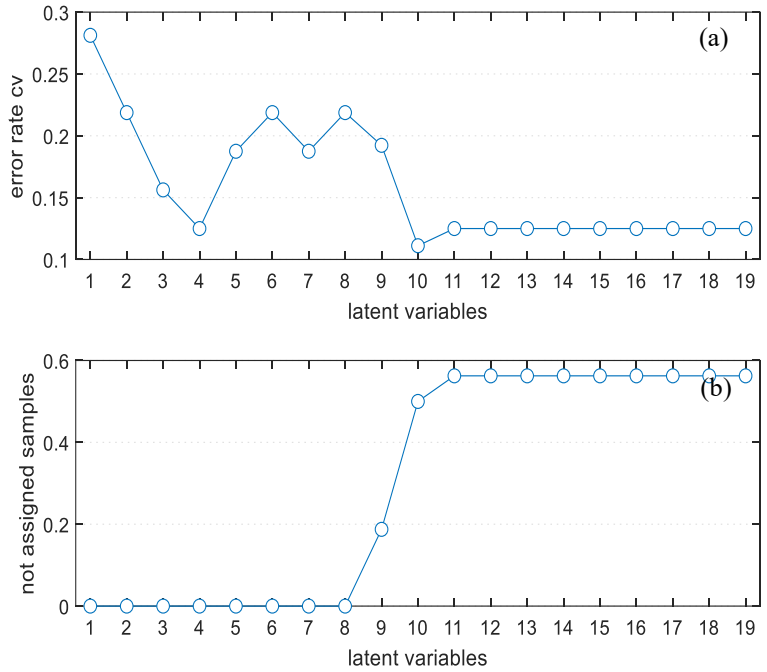


Figure 7.5 Error rate (a) and ratio of not assigned samples (b) in cross validation as a function of the latent variables number included in the PLS-DA model using a Bayesian assignation criterion.

7.4 Conclusions

This study aimed to identify which metabolites were mainly responsible for the discrimination of cheese samples based on ripening time and thermal treatment. Two-way ANOVA was employed to test the hypothesis, showing that 42 metabolites were statistically significant in discriminating at least one of the two factors, comprising the interaction effect. The Benjamini-Hochberg method was applied to assess, by means of a more conservative approach, the significance of content variations of metabolites over different ripening times and thermal treatments. As a result of the analysis, 9 metabolites exhibited significant changes over the course of ripening, while only one metabolite underwent a significant change due to thermal treatment. Proline and 4-aminobutanoic acid were found to be upregulated in late ripening cheese, as confirmed by the statistical significance downstream FDR correction. The concentration of compounds belonging to biogenic amines, saccharides and carboxylic acids classes resulted to significantly change when milk was treated at 68 °C, compared to raw milk or milk treated at 57 °C. A univariate t-test classifier was then utilized to classify the test samples and showed improved accuracy compared to PLS-DA in terms of thermal treatment classification. Conversely, multivariate classifier performances were higher when it was tasked with ripening time-based classification. The results of this work support the use of a univariate approach in classifying the cheese samples under investigation. For this case study, the analysis of individual metabolites may provide precise information and be an effective tool for sample classification. In contrast, joint information from the entire metabolic profile did not yield optimal results in terms of an accurate overall classification. This outcome may be due to the presence of uninformative predictor variables (i.e. metabolite content) for which the multivariate model did not fit experimental data as accurately as the univariate approach. For such an issue, it is important to consider the predictor variables correlation and the analysis's purpose when deciding whether to use a univariate or multivariate model.

In conclusion, the application of an untargeted metabolomic proved to be a promising tool for a comprehensive understanding of the influence of thermal treatment and ripening time on the metabolic profile of Fiore Sardo cheese. Identifying the key metabolites responsible for such effects permitted the development of a classification tool capable of identifying cheese loaves subjected to different treatments.

7.5 References

Afshari, R., Pillidge, C. J., Dias, D. A., Osborn, A. M., & Gill, H. (2021). Biomarkers associated with cheese quality uncovered by integrative multi-omic analysis. Food Control, 123. https://doi.org/10.1016/j.foodcont.2020.107752

Anedda, R., Melis, R., & Curti, E. (2021). Quality Control in Fiore Sardo PDO Cheese: Detection of Heat Treatment Application and Production Chain by MRI Relaxometry and Image Analysis. Dairy, 2(2), 270–287. https://doi.org/10.3390/dairy2020023

Atasoy, A. F., & Türkoğlu, H. (2008). Changes of composition and free fatty acid contents of Urfa cheeses (a white-brined Turkish cheese) during ripening: Effects of heat treatments and starter cultures. Food Chemistry, 110(3), 598–604. https://doi.org/10.1016/j.foodchem.2008.02.046

Azcarate, S. M., De Araújo Gomes, A., Alcaraz, M. R., Ugulino De Araújo, M. C., Camiña, J. M., & Goicoechea, H. C. (2015). Modeling excitation-emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety. Food Chemistry, 184, 214–219. https://doi.org/10.1016/j.foodchem.2015.03.081

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical Methods, 5(16), 3790–3798. https://doi.org/10.1039/c3ay40582f

Becchi, P. P., Rocchetti, G., Vezzulli, F., Lambri, M., & Lucini, L. (2023). The integrated metabolomics and sensory analyses unravel the peculiarities of mountain grassland-based cheese production: The case of Parmigiano Reggiano PDO. Food Chemistry, 428. https://doi.org/10.1016/j.foodchem.2023.136803

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benkerroum, N. (2016). Biogenic Amines in Dairy Products: Origin, Incidence, and Control Means. Comprehensive Reviews in Food Science and Food Safety, 15(4), 801–826. https://doi.org/10.1111/1541-4337.12212

Berrar, D. (2018). Cross-validation. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1–3, 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Brereton, R. G. (2009). Chemometrics for Pattern Recognition. In Chemometrics for Pattern Recognition. https://doi.org/10.1002/9780470746462

Caboni, P., Maxia, D., Scano, P., Addis, M., Dedola, A., Pes, M., Murgia, A., Casula, M., Profumo, A., & Pirisi, A. (2019a). A gas chromatography-mass spectrometry untargeted metabolomics approach to discriminate Fiore Sardo cheese produced from raw or thermized ovine milk. Journal of Dairy Science, 102(6), 5005–5018. https://doi.org/10.3168/jds.2018-15885

Caboni, P., Murgia, A., Porcu, A., Demuru, M., Pulina, G., & Nudda, A. (2016). Gas chromatography-mass spectrometry metabolomics of goat milk with different polymorphism at the αS1-casein genotype locus. Journal of Dairy Science, 99(8), 6046–6051. https://doi.org/10.3168/jds.2015-10537

Caboni, P., Murgia, A., Porcu, A., Manis, C., Ibba, I., Contu, M., & Scano, P. (2019b). A metabolomics comparison between sheep's and goat's milk. Food Research International, 119, 869–875. https://doi.org/10.1016/j.foodres.2018.10.071

Collins, Y. F., McSweeney, P. L. H., & Wilkinson, M. G. (2003). Lipolysis and free fatty acid catabolism in cheese: A review of current knowledge. International Dairy Journal, 13(11), 841–866. https://doi.org/10.1016/S0958-6946(03)00109-2

Combarros-Fuertes, P., Fernández, D., Arenas, R., Diezhandino, I., Tornadijo, M. E., & Fresno, J. M. (2016). Biogenic amines in Zamorano cheese: Factors involved in their accumulation. Journal of the Science of Food and Agriculture, 96(1), 295–305. https://doi.org/10.1002/jsfa.7093

Conte, P., Cinquanta, L., Lo Meo, P., Mazza, F., Micalizzi, A., & Corona, O. (2021). Fast field cycling NMR relaxometry as a tool to monitor Parmigiano Reggiano cheese ripening. Food Research International, 139. https://doi.org/10.1016/j.foodres.2020.109845

Dedola, A. S., Piras, L., Addis, M., Pirisi, A., Piredda, G., Mara, A., & Sanna, G. (2020). New analytical tools for unmasking frauds in raw milk-based dairy products: Assessment, validation and application to fiore sardo PDO cheese of a RP-HPLC method for the evaluation of the α-l-fucosidase activity. Separations, 7(3), 1–11. https://doi.org/10.3390/separations7030040

del Rio, B., Fernandez, M., Redruello, B., Ladero, V., & Alvarez, M. A. (2024). New insights into the toxicological effects of dietary biogenic amines. Food Chemistry, 435. https://doi.org/10.1016/j.foodchem.2023.137558

FAO. (2023). Gateway to dairy production and products.

Ferrante, M. C., & Mercogliano, R. (2023). Focus on histamine production during cheese manufacture and processing: A review. Food Chemistry, 419. https://doi.org/10.1016/j.foodchem.2023.136046

García-Pérez, P., Becchi, P. P., Zhang, L., Rocchetti, G., & Lucini, L. (2024). Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity. Trends in Food Science and Technology, 147. https://doi.org/10.1016/j.tifs.2024.104481

Hassan, F. A. M., Abd El-Gawad, M. A. M., & Enab, A. K. (2012). Flavour compounds in cheese (review). International Journal of Academic Research, 4(5), 169–181. https://doi.org/10.7813/2075-4124.2012/4-5/a.20

Jaffrès, E., Lalanne, V., Macé, S., Cornet, J., Cardinal, M., Sérot, T., Dousset, X., & Joffraud, J. J. (2011). Sensory characteristics of spoilage and volatile compounds associated with bacteria isolated from cooked and peeled tropical shrimps using SPME-GC-MS analysis. International Journal of Food Microbiology, 147(3), 195–202. https://doi.org/10.1016/j.ijfoodmicro.2011.04.008

Kanani, H., Chrysanthopoulos, P. K., & Klapa, M. I. (2008). Standardizing GC-MS metabolomics. Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 871(2), 191–201. https://doi.org/10.1016/j.jchromb.2008.04.049

Kandasamy, S., Yoo, J., Yun, J., Kang, H. B., Seol, K. H., & Ham, J. S. (2021). Quantitative analysis of biogenic amines in different cheese varieties obtained from the korean domestic and retail markets. Metabolites, 11(1), 1–16. https://doi.org/10.3390/metabo11010031

Kebary, K. M. K., El-Sonbaty, A. H., & Badawi, R. M. (1999). Effects of heating milk and accelerating ripening of low fat Ras cheese on biogenic amines and free amino acids development. Food Chemistry, 64(1), 67–75. https://doi.org/10.1016/S0308-8146(98)00085-5

Khattab, A. R., Guirguis, H. A., Tawfik, S. M., & Farag, M. A. (2019). Cheese ripening: A review on modern technologies towards flavor enhancement, process acceleration and improved quality assessment. Trends in Food Science and Technology, 88, 343–360. https://doi.org/10.1016/j.tifs.2019.03.009

Kováts, E. (1958). Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helvetica Chimica Acta, 41(7), 1915–1932. https://doi.org/10.1002/hlca.19580410703

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? Korean Journal of Anesthesiology, 71(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

Liao, S. M., & Akritas, M. (2007). Test-based classification: A linkage between classification and statistical testing. Statistics and Probability Letters, 77(12), 1269–1281. https://doi.org/10.1016/j.spl.2007.03.012

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & De Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes, 4. https://doi.org/10.1186/1756-0500-4-299

Moniente, M., Botello-Morte, L., García-Gonzalo, D., Virto, R., Pagán, R., Ferreira, V., & Ontañón, I. (2023). Combination of SPE and fluorescent detection of AQC-derivatives for the determination at sub-mg/L levels of biogenic amines in dairy products. Food Research International, 165. https://doi.org/10.1016/j.foodres.2022.112448

Montgomery, D. (2013). Design and analysis of experiments. John Wiley \& Sons, Inc.

Natrella, G., Gambacorta, G., Squeo, G., & Faccia, M. (2023). Impact of Milk Thermization on the Quality Characteristics of P.D.O. “Canestrato Pugliese” Ovine Hard Cheese. Foods, 12(5). https://doi.org/10.3390/foods12051080

NIST. (2005). Mass spectral library (NIST/EPA/NIH).

Nomura, M., Kimoto, H., Someya, Y., Furukawa, S., & Suzuki, I. (1998). Production of γ-Aminobutyric Acid by Cheese Starters during Cheese Ripening. Journal of Dairy Science, 81(6), 1486–1491. https://doi.org/10.3168/jds.S0022-0302(98)75714-5

Ochi, H., Naito, H., Iwatsuki, K., Bamba, T., & Fukusaki, E. (2012). Metabolomics-based component profiling of hard and semi-hard natural cheeses with gas chromatography/time-of-flight-mass spectrometry, and its application to sensory predictive modeling. Journal of Bioscience and Bioengineering, 113(6), 751–758. https://doi.org/10.1016/j.jbiosc.2012.02.006

OECD/FAO. (2021). OECD-FAO Agricultural Outlook 2021–2030. OECD-FAO Agricultural Outlook 2021–2030.

Putri, S. P., Ikram, M. M. M., Sato, A., Dahlan, H. A., Rahmawati, D., Ohto, Y., & Fukusaki, E. (2022). Application of gas chromatography-mass spectrometry-based metabolomics in food science and technology. Journal of Bioscience and Bioengineering, 133(5), 425–435. https://doi.org/10.1016/j.jbiosc.2022.01.011

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. B. (2014). Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics, 10(3), 361–374. https://doi.org/10.1007/s11306-013-0598-6

Santiago-López, L., Aguilar-Toalá, J. E., Hernández-Mendoza, A., Vallejo-Cordoba, B., Liceaga, A. M., & González-Córdova, A. F. (2018). Invited review: Bioactive compounds produced during cheese ripening and health effects associated with aged cheese consumption. Journal of Dairy Science, 101(5), 3742–3757. https://doi.org/10.3168/jds.2017-13465

Sibono, L., Grosso, M., Tronci, S., Errico, M., Addis, M., Vacca, M., Manis, C., & Caboni, P. (2023). Investigation of Seasonal Variation in Fatty Acid and Mineral Concentrations of Pecorino Romano PDO Cheese: Imputation of Missing Values for Enhanced Classification and Metabolic Profile Reconstruction. Metabolites, 13(7). https://doi.org/10.3390/metabo13070877

Sibono, L., Manis, C., Zucca, F., Atzori, L., Errico, M., Tronci, S., Casula, M., Dedola, A., Pes, M., Caboni, P., & Grosso, M. (2024). Metabolomic profiling of Fiore Sardo cheese: Investigation of the influence of thermal treatment and ripening time using univariate and multivariate classification techniques. Food Chemistry, 456. https://doi.org/10.1016/j.foodchem.2024.139930

Silva, H. L. A., Balthazar, C. F., Esmerino, E. A., Vieira, A. H., Cappato, L. P., Neto, R. P. C., Verruck, S., Cavalcanti, R. N., Portela, J. B., Andrade, M. M., Moraes, J., Franco, R. M., Tavares, M. I. B., Prudencio, E. S., Freitas, M. Q., Nascimento, J. S., Silva, M. C., Raices, R. S. L., & Cruz, A. G. (2017). Effect of sodium reduction and flavor enhancer addition on probiotic Prato cheese processing. Food Research International, 99, 247–255. https://doi.org/10.1016/j.foodres.2017.05.018

Singh, T. P., Arora, S., Borad, S. G., Bam, J., Paul, V., Thomas, R., & Sarkar, M. (2023). Fatty acid and amino acid profiling, antioxidant activity and other quality characteristics of vacuum packed cheddar style-yak milk cheese during ripening. Food Bioscience, 51. https://doi.org/10.1016/j.fbio.2022.102213

Teixido-Orries, I., Molino, F., Femenias, A., Ramos, A. J., & Marín, S. (2023). Quantification and classification of deoxynivalenol-contaminated oat samples by near-infrared hyperspectral imaging. Food Chemistry, 417. https://doi.org/10.1016/j.foodchem.2023.135924

Tekin, A., & Güler, Z. (2019). Glycolysis, lipolysis and proteolysis in raw sheep milk Tulum cheese during production and ripening: Effect of ripening materials. Food Chemistry, 286, 160–169. https://doi.org/10.1016/j.foodchem.2019.01.190

Varmuza, K., & Filzmoser, P. (2016). Introduction to Multivariate Statistical Analysis in Chemometrics. In Introduction to Multivariate Statistical Analysis in Chemometrics. https://doi.org/10.1201/9781420059496

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yılmaz, C., Kocadağlı, T., & Gökmen, V. (2023). Determination of endocannabinoids in fermented foods of animal and plant origin by liquid chromatography tandem mass spectrometry. Food Chemistry, 427. https://doi.org/10.1016/j.foodchem.2023.136766

Chapter 8: Feature selection algorithms to enhance classification models and biomarker identification in dairy product profiling.

With this chapter, the discussion of the -omics subject matter in this dissertation comes to an end. In this part of the thesis, the coagulation process (discussed in the process monitoring chapters) and cheese ripening (examined through the -omics chapters) are brought together and analyzed from a lipidomic perspective. The lipidomic profile of cheese from two different rennet sources across 18 months of ripening is investigated in this section. The problem of high dimensionality, which is commonly encountered in lipidomic datasets, is here addressed through the application of feature selection algorithms.

8.1 The role of different rennet sources within the dairy framework

Historically, the most used rennet enzyme is calf rennet, derived from the extract of the suckling-calf abomasa (Andrén, 2021). However, the cheesemaking industry is interested in exploring alternatives to animal-derived milk coagulants owing to their limited availability, escalating price and concerns related to dietary preferences, including vegetarianism, and religious beliefs (Shah et al., 2014). Conversely, the use of plant-based clotting enzymes is suitable for the production of cheeses which adheres to emerging markets. To this concern, in the last decade, extracts of *Cynara Cardunculus* plant have achieved an increasing interest in several industrial frameworks, such as cheesemaking, due to its wide availability and multifunctional properties (Gominho et al., 2011). *Cynara Cardunculus* is a herbaceous perennial diploid plant from which artichoke and cardoon are cultivated (Conceição et al., 2018; Timón et al., 2019).

To date, several works studied the different effects induced by vegetable rennet (such as *Cynara Cardunculus*) and animal rennet on the final product's quality, and elucidated their intricate behavior within different cheese ripening durations (Pacifico et al., 2024; Prieto et al., 2004; Rampanti et al., 2023; Sanjuán et al., 2002; Şengül et al., 2014). The cited studies contributed to defining the physico-chemical, proteolytic and proteomic characteristics of vegetable rennet-based cheese. This focus is attributed to the pivotal role of casein proteins, the main metabolites engaged in renneting processes. Indeed, although rennet enzymes are predominantly acknowledged for their proteolytic functions in degrading milk caseins during cheesemaking, few studies have demonstrated the possible lipolytic activity of rennet. As an example, one study examined the fluctuations in the concentrations of short, medium, and long-chain fatty acids across three different lamb rennets employed in the production of Pecorino Romano Cheese (Addis et al., 2005). Similarly, Santillo et al. examined the probiotic lipolytic activity contained in the same rennet paste for quantifying fatty acids increasing during ripening, within the same type cheese (Santillo et al., 2009). However, only animal-based rennet has been subjected to lipidomic analysis. Consequently, there is a significant gap in a broader characterization of the lipidic profiles of cheeses produced using vegetable rennet, with particular regard to higher molecular weight compounds such as sphingolipids, phospholipids and triacylglycerols. This lack extends to understanding the differences between vegetable-rennet cheeses and animal-rennet ones. Moreover, it is widely recognized that extracts of *Cynara Cardunculus* are affected by high variability of the biochemical profile (Conceição et al., 2018; Gomes et al., 2019), which can

significantly slow down its introduction in the dairy industry. A broader knowledge of vegetable rennet-based cheese metabolomics can provide a powerful benchmark for enhancing and standardizing the final product's quality. Such metabolomic extension opens up the possibility to apply chemometric analysis tools for developing more reliable classification models for detecting different renneting technologies that can be employed in the industrial dairy panorama.

8.2 Untargeted Lipidomics by LC-QTOF-MS for Biomarker Selection and Discrimination of plant- and animal-derived Rennet in Cooked-Curd ovine cheese at Different Ripening Stages

The present study proposes a lipidomic study with the aim of extending the current set of biomarkers for the recognition of different types of rennet during cheese ripening. Specifically, the content of fatty acids, triglycerides, sphingolipids and phospholipids obtained through LCQTOF-MS measurements was coupled with a chemometric approach to study the lipidic profile of cooked-curd ovine cheeses coagulated through a coagulating enzyme extracted from the *Cynara Cardunculus* plant and the one obtained from calf rennet-based cheeses for different ripening times. The application of a robust classification model for discriminating samples belonging to different classes coupled with univariate techniques facilitated the identification of lipidic features that exhibited significant variations across the investigated treatments.

8.2.1 Materials and methods

For the study presented in this chapter, the analytical procedures presented in the materials and methods were carried out by external collaborators, whose contributions are acknowledged at the end of the manuscript. The author focused on the metabolomic data analysis and the interpretation of the results. The experimental protocol is presented anyway.

8.2.1.1 Dataset description

The metabolic dataset includes 18 cheese samples and 1,018 LCQTOF-MS features, combining 738 from the positive ion mode and 280 from the negative ion mode. The data matrix included two classes based on the rennet type used in cheese production: animal rennet and vegetal-rennet coagulated cheeses, each comprising nine samples. Within each rennet type category, the samples

are further subdivided based on ripening time, with three samples belonging to each of the following ripening periods: 24 hours, 12 months, and 18 months.

8.2.1.2 Cheese production

The cheese samples were obtained from three experimental cheesemaking trials carried out on three different days in a dairy farm located in Sardinia (Italy) according to the cheesemaking technology for hard, cooked-curd sheep's milk cheese. Every cheesemaking was performed using bulk sheep's milk, which was first divided into two aliquots, thermized (68 °C x 120 s) and then cooled to 37–38 °C. Subsequently, a thermophilic starter culture (CHOOZIT® SU CASU, Danisco, Copenaghen, Denmark) and a secondary support culture (CHOOZIT® FLAV 54, Danisco, Copenaghen, Denmark) were added to the milk. One aliquot of milk was coagulated using a commercial liquid calf rennet (Caglificio Manca, Thiesi, Italy; chymosin 75%, pepsin 25%; 150 IMCU (International Milk Clotting Unit) mL-1) at a dosage of 35 mL 100 L–1 milk while the other one was coagulated with the commercial liquid vegetable rennet Gallium® containing plant enzymes extracted from *Cynara cardunculus* (Laboratorio Prodor, Bobbio, Italy; 55 IMCU mL-1) at a dosage of 100 mL 100 L–1 milk. The dosage of each rennet was selected so as to standardize the clotting time to approximately 10 minutes. After an additional 15-minute period necessary to allow curd firming, the curd was subsequently subjected to further mechanical and physical processing according to the diagram reported in figure 8.1. The cheeses wheels obtained for each cheesemaking process were sampled one day from production and after a ripening period of 12 months and 18 months under controlled conditions (10 - 12 °C and 78–85% relative humidity); in particular, two opposite wedges weighing a total of approximately 1.5 kg (10% of the entire wheel) were taken from each wheel of cheese weighing 15-16 kg. The wedges were then ground, homogenised and stored at -20 °C until further analysis. Further details about the production process are reported in a previous work (Pes et al., 2014).

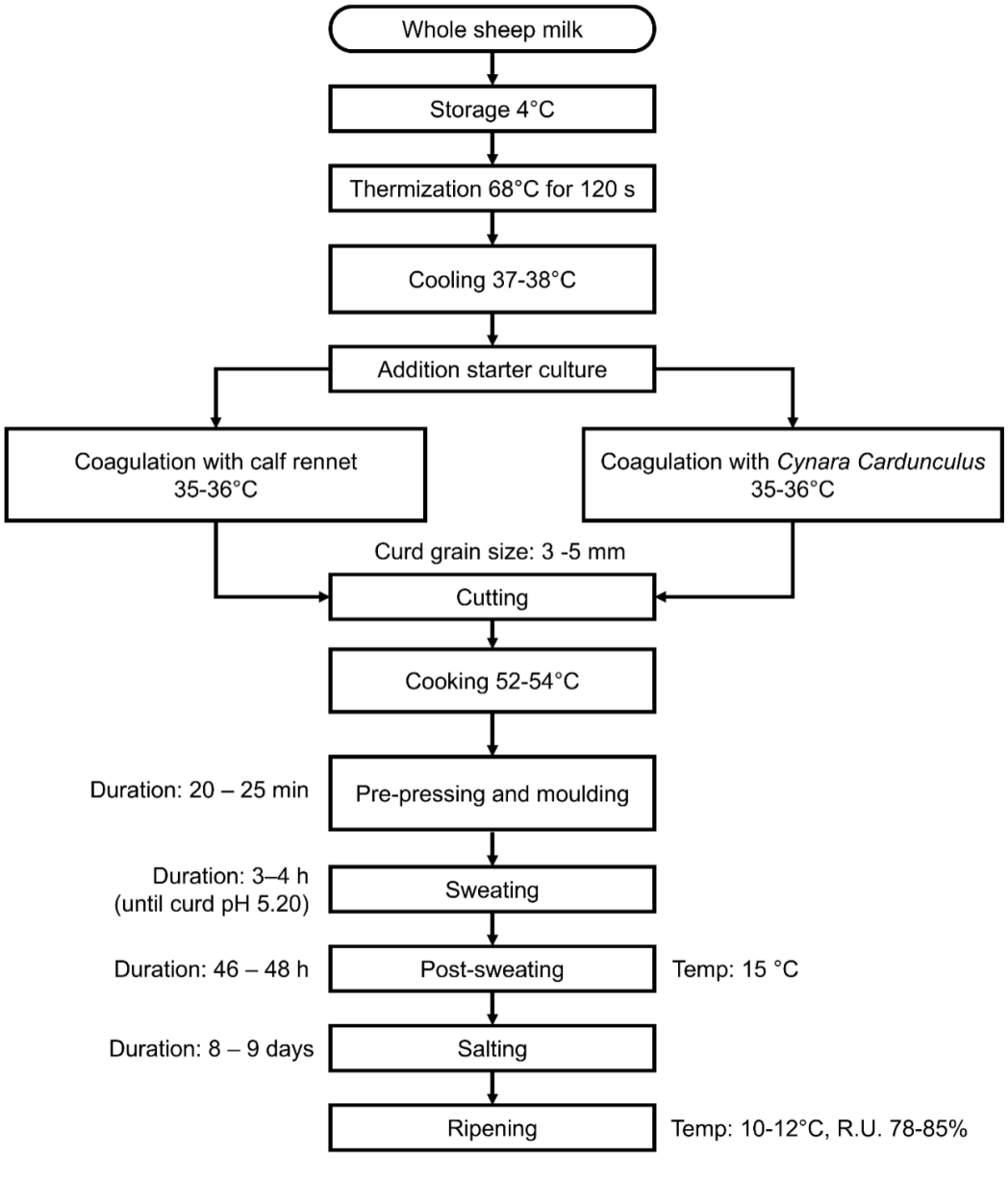


Figure 8.1 Flowchart of the Cheesemaking process.

Reproduced from: Leonardo Sibono, Massimiliano Grosso, Massimiliano Errico, Stefania Tronci, Giacomo Lai, Massimo Pes, Margherita Addis, Pierluigi Caboni, Cristina Manis, Untargeted lipidomics by LC-QTOF-MS for biomarker selection and discrimination of plant- and animal-derived rennet in cooked-curd ovine cheese at different ripening stages, Food Research International, 2025, 117987, ISSN 0963-9969, https://doi.org/10.1016/j.foodres.2025.117987.

8.2.1.3 Sample preparation and lipidomic analysis.

Lipid fractions from animal-rennet and vegetal-rennet coagulated cheese samples were extracted following a modified Folch method (FOLCH et al., 1957). Briefly, 10 mg of each sample were mixed with 250 µL of methanol and 125 µL of chloroform, followed by vortexing at 15-min intervals for 1 h. An additional 350 µL of chloroform and 150 µL of methanol were then added, and the mixture was incubated for a further hour. After centrifugation at 17,700 × g for 10 min, the resulting organic (400 µL) layer was carefully collected into separate glass vials and evaporated under a gentle nitrogen stream.

The dried lipid extract was reconstituted in 20 µL of a methanol:chloroform solution (1:1, *v/v*) and diluted with 800 µL of isopropanol:acetonitrile:water (2:1:1, *v/v*) for analysis by UHPLC-QTOF-MS in positive and negative ionization mode. The analysis of the lipid extract was carried out using a 6560 LC-QTOF-MS system coupled to an Agilent 1290 Infinity II UHPLC platform. For positive ionization mode, 2.0 µL of each sample was injected onto a Kinetex C18 column (1.6 µm, 100 mm × 2.1 mm; Phenomenex, Bologna, Italy), while 10 µL were injected for negative ionization mode. Chromatographic separation was performed at 50 °C, with a constant flow rate of 0.5 mL/min.

The mobile phase composition differs by ionization mode. In positive mode, eluent A consisted of 10 mM ammonium formate in 60% Milli-Q water/40% acetonitrile, while eluent B was 90% isopropanol/10% acetonitrile (9:1) containing 10 mM ammonium formate. For negative ionization mode, 10 mM ammonium acetate replaced ammonium formate in both phases. The elution gradient began with 60% A, decreased to 50% at 2.1 min, then to 30% by 10 min, and further reduced to 1% A, maintained for 1.9 min, before re-equilibrating to initial conditions within 1 min.

Mass spectrometric detection employed an Agilent Jet Stream ESI source, operating in both ionization modes. Source conditions were as follows: gas temperature at 250 °C, drying gas flow at 5 L/min, nebulizer pressure at 20 psig, sheath gas temperature at 275 °C, and sheath gas flow at 12 L/min. The capillary voltage was set to +4000 V (positive mode) and −3000 V (negative mode); nozzle voltage was 500 V, fragmentor voltage 400 V, skimmer 65 V, and octopole RF 750 V. The mass range acquired was m/z 50–1700, with a collision energy of 20 eV in positive mode and 25 eV in negative mode, allowing for up to 3 precursors per MS/MS cycle.

Instrument calibration was performed prior to analysis using a standard Agilent tuning mix (m/z 50–1700). A reference mass solution was continuously infused throughout the run for real-time mass axis correction. Data acquisition was performed using the Agilent MassHunter LC/MS Acquisition software, and the results were processed with the Agilent Lipid Annotator software

(Agilent Technologies, Santa Clara, USA) as part of a comprehensive lipidomics workflow.
Data acquired with LCQTOF-MS in both, positive and negative ionization mode, were pre-processed with the software MassHunter Workstation suite (Agilent Technologies, Santa Clara, USA). This software (Mass Profinder 10.0) allowed to perform time alignment and deconvolution of signals removing the background noise and unrelated ions obtained from blank samples through the tool called Recursive Feature Extraction (RFE), yielding a matrix containing all the features present across all samples. Overall, a data matrix of 738 features for the positive ion mode and 280 features for the negative ion mode was obtained. These filtered matrices were combined and then subjected to statistical analysis. Discriminant metabolites were annotated using iterative MS/MS acquisition, which provided characteristic fragmentation spectra at a collision energy (CE) of 20 eV.

8.2.1.4 Statistical analysis: Forward PLS-DA feature selection and sample classification

A Two-way ANOVA test was conducted on raw LCQTOF-MS data to assess whether the investigated factors and their interaction have a significant effect on each variable. After conducting the univariate test, two sets of p-values (one set for each main factor) were arranged in ascending order. Each p-value denotes the significance of the factor's effect on every individual feature. Univariate tests such as two-way ANOVA, though useful for preliminary feature screening, may yield high false negative rates after multiple testing correction. To account for the multivariate structure of the data, an iterative forward variable selection approach was implemented, initialized with the top-ranked ANOVA features and followed by PLS-DA. Briefly, the algorithm is initialized by collecting a subset of variables from the original matrix, comprising the first two variables with the lowest p-values determined by two-way ANOVA. PLS-DA was executed on unity variance preprocessed data subset. This iterative procedure involved a progressive addition of each variable and validation of the subset-based model through a Montecarlo cross-validation with 1000 simulations (Xu & Liang, 2001). For each subset, the optimal number of LV was determined based on achieving the highest accuracy. After examining the entire feature space, the model with the highest accuracy was chosen as the best-performing one. In each Monte Carlo simulation, 20% of the dataset for validation and the remaining for calibration. This procedure refers to a forward selection interval-PLS-DA algorithm, wherein variables are sequentially added to an initially low-feature subset candidate until the assessment of the overall feature space (Nørgaard et al., 2000). The variables subset which corresponded to the best discrimination accuracy was selected as the biomarkers set for the specific factor (i.e. rennet type and ripening time). PLS-DA was performed

using the Classification Toolbox 6.0 from Milano Chemometrics (Ballabio & Consonni, 2013). Figure 8.2 shows a scheme of the implemented algorithm.

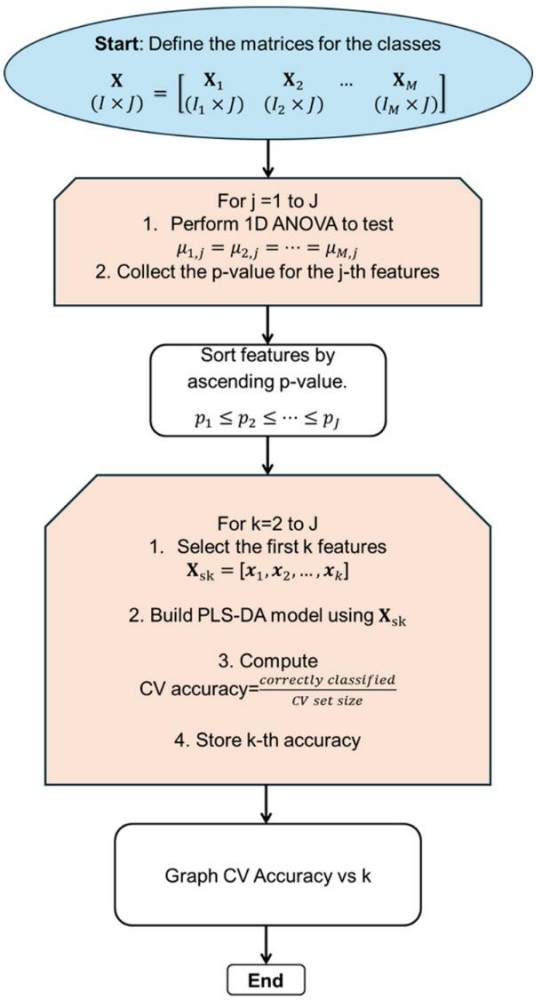


Figure 8.2 Schematic representation of forward-selection algorithm for biomarkers determination. In the presented algorithm, I_1, I_2 and I_M respectively denote the number of samples corresponding to class 1,2 and M for the data matrix **X** with I samples and J features, while $\mu_{1,j}$, $\mu_{2,j}$ and $\mu_{M,j}$ are the relevant class mean values associated with the feature j. p-values rank is then used to sort feature vectors used to generate the subset Data Matrix $\mathbf{X}_{sk}$.

Reproduced from: Leonardo Sibono, Massimiliano Grosso, Massimiliano Errico, Stefania Tronci, Giacomo Lai, Massimo Pes, Margherita Addis, Pierluigi Caboni, Cristina Manis, Untargeted lipidomics by LC-QTOF-MS for biomarker selection and discrimination of plant- and animal-derived rennet in cooked-curd ovine cheese at different ripening stages, Food Research International, 2025, 117987, ISSN 0963-9969, https://doi.org/10.1016/j.foodres.2025.117987.

8.2.2 Results and discussion

After conducting ANOVA on LCQTOF-MS data, a set of 1018 p-values corresponding to each metabolite was sorted in ascending order. Due to the high dimension of the data matrix, part of such variables is expected to be redundant or uninformative for the assessment of the effect of the stimuli under investigation. The employment of a local modeling procedure involving a variable window selection can significantly improve both the model's predictive ability and computational time (Nørgaard et al., 2000). Indeed, the overall computational complexity of PLS algorithm is proportional to the square of the number of variables (Qin et al., 2022).

In the following discussion, a lipidomic investigation of the impact of different rennet sources (vegetable vs animal) on the content of a broad set of lipids in cheese is presented for the first time. The subsequent analysis of ripening time provides an additional means of validation of the applied chemometric method and the overall structure of the collected data, while providing a deeper understanding of both triglycerides and phospholipids lipolysis mechanisms.

8.2.2.1 Rennet type analysis

The discrimination potential of LCQTOF-MS data was assessed through forward stepwise interval PLS-DA, which selects the subset of metabolites associated with the maximum accuracy of classification. For each subset, the optimal number of LVs was selected as the one that exhibited the best accuracy. Figure 8.3 depicts the classification accuracy in cross validation for 1018 variables subsets and their associated optimal number of LVs. For the case of vegetable vs animal rennet, a 40 metabolites subset was selected as the optimal window, giving rise to classification accuracy in cross-validation of 88%. Conversely, only 56% of accuracy was achieved when the entire dataset was considered for the analysis. It is worth noting that the employment of two LVs guaranteed the achievement of the optimal performance when the first 40 variables were included in the window, meaning that the essential discriminatory information was captured in a parsimonious bidimensional latent space. On the other hand, when data dimensionality was high (i.e., number of features > 450), the optimal number of latent variables (LVs) increased sharply. This was likely due to the increased system complexity, which in turn reduced the model's ability to reliably discriminate between samples.

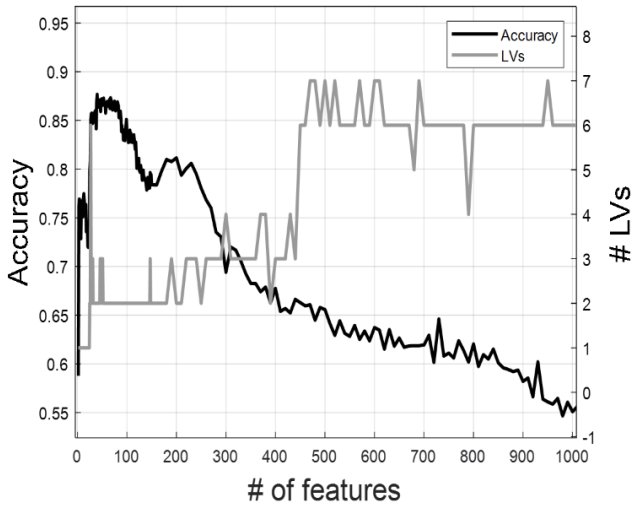


Figure 8.3 Classification accuracy in cross-validation as a function of the number of features considered in the PLS-DA classifier for the rennet type (black line), and the corresponding optimal number of LVs obtained from each subset (grey line).

Table 8.1 indicates the metabolites included in the optimal subset and their associated p-value from two-way ANOVA test. The rennet analysis identified a broad range of lipidic families. In the context of this study, significant differences were observed between the two rennet types in the content of phosphatidylserines (PS), triacylglycerols (TG), fatty acids (FA), phosphatidylcholines (PC), ceramides (Cer), sphingomyelins (SM), phosphatidylethanolamine (PE) and monoacylglycerols (MG). Such compounds were identified in other cheese lipidomic studies with diverse biomarker selection purposes using an unspecified rennet type (Haddad et al., 2022; Tomaiuolo et al., 2023). It is worth noting that among the presented molecules, several triacylglycerol-derived compounds such as MGs and FAs resulted to significantly vary between cardoon rennet-based cheeses and the calf rennet-based ones. This outcome can be due to the relatively high abundancy of triacylglycerols in every milk substrates, including sheep's milk, as they account for 95% of total fat (Zhao et al., 2022), supporting their main involvement in lipolytic phenomena. The remaining fraction (5%) is mainly constituted of phospholipids. The five major phospholipids in dairy products are PC, PE, SM, PI and PS and are located in the milk fat globule membrane (Verardo et al., 2013).

Their considerable contribution to the discrimination of cheese samples obtained from different rennet can be attributed to a possible high content of phospholipases enzymes in the calf rennet. These enzymes are known to facilitate the degradation of the fat globule membrane, which promotes the releasing of its constituents in the dairy matrix (Patton & Keenan, 1975).

Table 8.1 Two-way ANOVA p-values for the subset of metabolites demonstrating optimal performance of the PLS-DA classifier for rennet type.

Metabolite	p-value	Metabolite	p-value
U1	0.0062	PE (18:1_18:1)	0.0269
PS (18:1_18:1)	0.0077	Cer (d18:1_23:0)	0.0272
TG (51:5)	0.0085	PE (P-18:1_18:1)	0.0274
FA (21:3)0.0162	0.0162	dihydroxy FA (20:0)	0.0301
FA (22:5)	0.0166	U4	0.0303
14:0 Cholesteryl ester	0.0166	MG(21:0)	0.0304
SM (23:0;O5)	0.0167	MG(16:0)	0.0315
PC (14:0)	0.0180	U5	0.03263
PS (18:0_22:0)	0.0198	Octadecanedioic acid	0.0328
U3	0.0198	PI (18:0_18:2)	0.0328
Cer (d18:0_23:0)	0.0203	Cholesterol glucuronide	0.0330
SM (d33:1)	0.0227	PS (18:1_20:0)	0.0335
FA (18:2)	0.0232	FA (19:1)	0.0343
FA (24:2)	0.0252	FA (13:0)	0.0346
PC (P-16:0_18:4)	0.0254	FA (22:1)	0.0382
FA (21:1)	0.0258	FAHFA(24:5-17:0)	0.0382
SM (d36:1)	0.0261	N-(4-benzenesulfonamide)arachidonoyl amide	0.0396

PE (18:0_18:1)	0.0261	U6	0.0400
FA (24:1)	0.0266	U7	0.0404
SM (d37:1)	0.0267	Cer (m18:0_24:0)	0.0405

U: Unknown; PS: PhosphatidylSerines; TG: TriacylGlycerols; FA: Fatty Acids; PC: PhosphatidylCholines; Cer: Ceramides, SM: SphingoMyelins; PhosphatidylEthanolamine; MG: MonoacylGlycerols.

To this regard, differences between calf and cardoon rennet lipidomics are depicted in Figure 8.4, which shows the biplot obtained from the PLS-DA model including the optimal lipids subset. As can be seen, the LV space displays a clear separation between the scores belonging to the different groups. From a visual inspection of loading values, it can be concluded that only 3 metabolites were clearly upregulated in cardoon-based cheese samples, while the other 37 resulted to show either a strong or a weak upregulation in the calf rennet-based samples. This pattern is consistent with the mean values of LCQTOF-MS peaks observed in each class comparison across all metabolites.

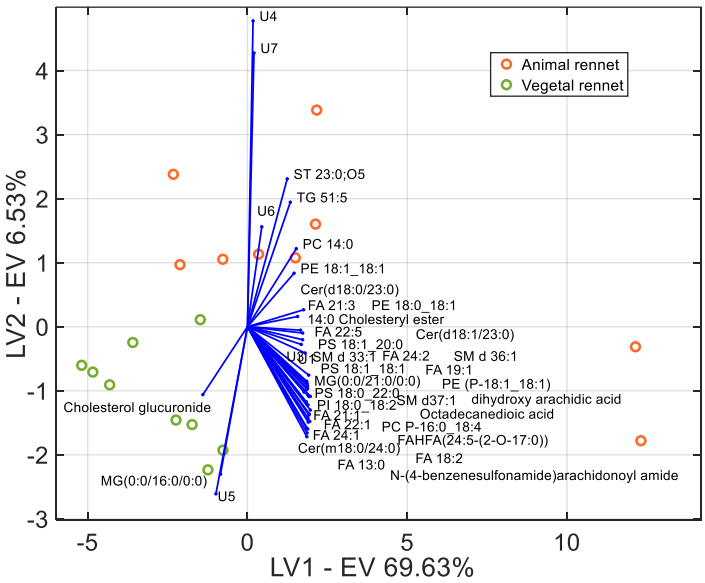


Figure 8.4 PLS biplot showing scores for different rennet types and the lipids loading values obtained from the optimal features subset of rennet analysis. For a clear visual analysis, loading values were multiplied 10 times.

Reproduced from: Leonardo Sibono, Massimiliano Grosso, Massimiliano Errico, Stefania Tronci, Giacomo Lai, Massimo Pes, Margherita Addis, Pierluigi Caboni, Cristina Manis, Untargeted lipidomics by LC-QTOF-MS for biomarker selection and discrimination of plant- and animal-derived rennet in cooked-curd ovine cheese at different ripening stages, Food Research International, 2025, 117987, ISSN 0963-9969, https://doi.org/10.1016/j.foodres.2025.117987.

It has been demonstrated that several animal rennet contain pregastric lipolytic enzymes and esterases that trigger the formation of FA, PC, PE, SN and Cer (Moschopoulou, 2011). This aspect can justify the main upregulation of lipids in the animal rennet-derived cheese. Interestingly, similar loading values were identified between Cer and SM compounds. This is probably due to the fact that ceramides are generally considered structurally similar to sphingomyelins (Z. Liu et al., 2020). Noteworthy, the only discernible pattern among cholesterols was remarked with cholesterol glucuronide, where a positive correlation was observed between its lipid content and cheese derived from cardoon rennet. This result aligns with the lipophilic composition of the cardoon plant (i.e. *Cynara cardunculus* L.) documented by Ramos et al., whose study indicated that sterols account for 1 to 11 % of the total detected compounds in all morphological parts of the plant (Ramos et al., 2013). The same authors detected a noticeable amount of MG(16:0/0:0/0:0), coherently with the upregulation of MG(0:0/16:0/0:0) in cheese samples obtained from cardoon renneting. Therefore, the upregulation of such compounds can be related to their high content in the rennet. In order to further evaluate the prediction quality of the model, the cross-validation task was repeated on a 10000 iterations Montecarlo simulation using the optimal subset of features, from which it resulted that the accuracy value is approximately 84%, which is 4 percentage points less than the one obtained from the algorithm reported in figure 8.3, but yet considered as consistent.

8.2.2.2 Ripening time analysis

Ripening stage duration is a dominant factor for FAs profile of cheese due to lipolysis of TGs (Haddad et al., 2022). The present chemometric analysis of the effect of ripening time on the lipidic profile of cheese allows to acquire a deeper understanding of cheese lipolytic processes.

Figure 8.5 depicts the accuracy of ripening time classification in cross validation for 1017 subsets (1018 variables) and their associated optimal number of LVs concerning the analysis of cheese ripening time. A 15-feature subset was selected as it achieved an optimal CV classification accuracy of 91%. On the other hand, only 74% accuracy was achieved when the entire dataset was considered for the analysis. In this case, the optimal performance was achieved from a three LVs-model when the first 15 lipids were included in the window, meaning that the essential discriminatory information was explained in a tridimensional latent space. Also in this scenario, when dealing with high data dimensionality (#features>350), the optimal number of LVs showed a sudden increase along with a remarkable decreasing of model's accuracy.

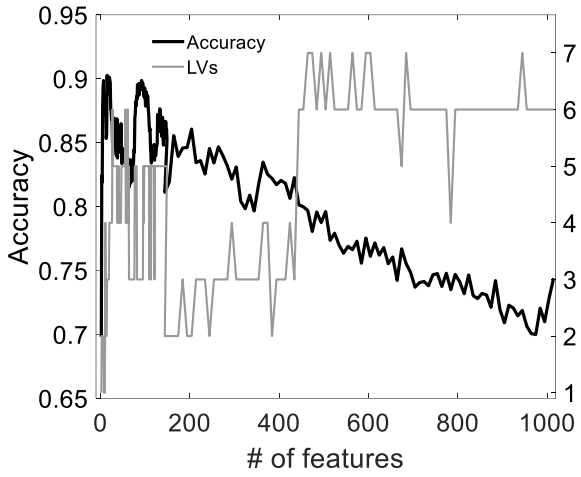


Figure 8.5 Classification accuracy in cross-validation as a function of the number of features considered in the PLS-DA classifier for the ripening time (black line), and the corresponding optimal number of LVs obtained from each subset (grey line).

In the examination of cheese ripening time, as delineated in table 8.2, the first lipidic group found to exhibit significant variation across different ripening durations is that of FAs. Notably, the associated p-values for ripening time are considerably lower compared to those linked with the type of rennet used. In fact, cheese ripening is one of the most influential factor for determining the metabolic composition (Scano et al., 2019).

Table 8.2 Two-way ANOVA p-values for the subset of metabolites demonstrating optimal performance of the PLS-DA classifier for ripening time.

Metabolite	p-value	Metabolite	p-value	Metabolite	p-value
LysoPE 18:1	2.2610e-07	FA 11:0	0.0002	hydroxy-octadecadienoic acid	0.0003
Cer(d18:1/26:0)	5.7332e-07	hydroxy palmitic acid	0.0002	DG 15:0_22:2	0.0004
FA 22:2	0.0001	PS 18:1_22:0	0.0002	FA 18:2	0.0004
U2	0.0001	FA 18:1	0.0002	FA 15:1	0.0004
TG 18:0_18:4_18:0)	0.0001	FA 12:0	0.0003	U8	0.0004

It is important to specify that none of the compounds reported in Table 8.2 resulted to significantly vary between the different rennet types, meaning that the following rationale for ripening times is statistically consistent for both calf and cardoon rennet-based cheeses. Figure 8.6 reports the PLS-DA biplot obtained from the optimal subset for ripening time cheese classification. A clear separation of samples ripened for different times can be observed from the scores' positions. The 24-hour ripened cheese samples exhibit prominently negative score values along the LV1 axis, contrasting with the predominantly positive or marginally negative scores observed for the 12 and 18-month clusters. Furthermore, LV2 delineates a notable difference between the 12-month and 18-month clusters, with the former displaying high values compared to the latter, indicative of discernible compositional differences between these ripening durations. The analysis of loading values provides a positive correlation of medium to long chain FAs content with cheese samples obtained from 18 months of ripening. Such an outcome can be attributed to the occurrence of multiple phenomena:

1) Following an initial phase of the lipolysis process, during which the outer and sterically less hindered position of the triglyceride (the sn-3) is preferentially hydrolyzed, the lipolytic enzymes are able to hydrolyze, in a later ripening time, even the less accessible

sites corresponding to the stereochemical number positions 1 and 2, where medium and long-chain fatty acids are preferentially esterified (Caboni et al., 2019). Indeed, the loadings position of TG (18:0_18:4_18:0), placed in the middle of 12 and 18 months-ripened cheese clusters, as well as the subsequent upregulation of C18 FAs in 18 months-ripened cheese is probably related to a delay of ester bonds cleavage due to steric hindrance.

2) A higher content of LysoPE 18:1 is detected in 12 months-ripened cheese. Probably, the hydrolysis of this phospholipid through the action of lysophospholipase enzymes contributed to the remarkable increasing of FA 18:1 concentration in 18 months-old cheeses (García-Cano et al., 2020).

Noteworthy, two hydroxy-fatty acids (OH-FAs) resulted to be upregulated in 18 months-ripened samples. Traces of OH-FAs were detected in other works, which recognized this group of lipids as an important marker for characterizing foodstuffs (Jenske & Vetter, 2009). The different regulation of hydroxy-palmitic acid and hydroxy-octadecanoic acid in cheese with different ripening time, as reported in this study, is probably related to the hydroxylation of fatty acids resulting from their microbial catabolism (Collins et al., 2003).

Ultimately, due to the characteristic times of lipolysis reactions, it can be deduced that none of the here studied lipids are apparent to be upregulated in 24-hour cheeses. This can be ascribed to the short residence time of lipolysis enzymes in the cheese substrate.

A complementary assessment of the model's predictive performance (using the optimal subset of features) was carried out through a Monte Carlo cross-validation scheme with 10000 resampling iterations. The resulting accuracy was 90%, which is consistent with that found from that obtained by iterating the PLS-DA model 1000 times (91%).

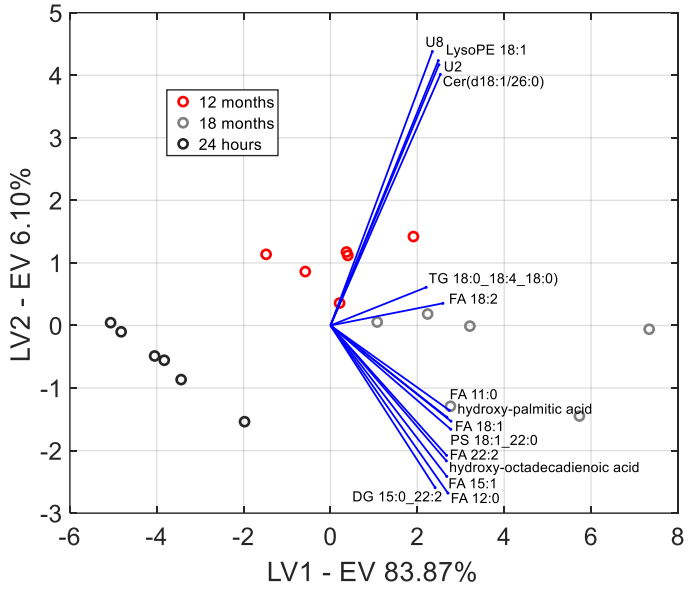


Figure 8.6 PLS biplot showing scores for different ripening times and the lipids loading values obtained from the optimal features subset of ripening time analysis.

In addition to the primary feature selection strategy adopted in this study, an ensemble-based decision tree approach was also considered. Specifically, a TreeBagger model (MATLAB's implementation of bootstrap-aggregated decision trees) was employed to obtain an alternative ranking of variable importance (Hajian et al., 2022). However, a comparison among different feature selection algorithms is out of the scope of this section. Therefore, additional details about the methodology are provided in Appendix D.

8.3 Conclusions

Cynara cardunculus from which rennet is extracted has great potential for different applications within pharmaceuticals, energy as well as in the food industry, thanks to its multifunctional biochemical profile. This study presented for the first time a lipidomic investigation of cheese deriving from vegetable and animal coagulant enzymes, proving the cardoon-derived cheese to be significantly different in its fat's composition to the traditional calf rennet. The utilization of chemometrics coupled with LCQTOF-MS data provides a powerful tool to detect the variability of specific biomarkers within complex biomatrices such as cheese.

Several multivariate phenomena, such as lipolytic processes, may exhibit diverse extent of variations under different factors, leading to disparities in performance when employing a classifier model like PLS-DA for the investigation of more than one stimulus. Specifically, it was observed that the model tasked to predict ripening times generally outperformed that of rennet type classification. This discrepancy underscores the pronounced role of ripening time in lipidic processes, contrasting with the primarily proteolytic aspects associated with rennet type. However, a proper selection of metabolites subset allowed the achievement of high performances of the classification model for both factors. Regarding rennet type investigation, a broad picture of biomarkers was identified, encompassing various lipid classes such as triglycerides and various phospholipids. In particular, several phosphatidylserines, triacylglycerols, phosphatidylcholines, ceramides, sphingomyelins, phosphatidylethanolamine, monoacylglycerols and free fatty acids resulted to be upregulated in animal rennet-based cheeses. This diversity reflects the significant dissimilarities in the lytic behavior of two coagulants from different sources (i.e. vegetable and animal). Conversely, biomarkers associated with ripening time were predominantly found within fatty acid molecules. This observation can be attributed to the high lipolysis extent of triacylglycerols, the main class of lipids involved during cheese ripening (Haddad et al., 2022).

These findings emphasize the importance of metabolite selection in achieving robust predictive outcomes for diverse factors influencing cheese quality. Moving forward, such insights could support cheesemaking research in establishing efficient formulations of plant rennet aimed at reducing enzyme's profile variability and optimizing cheese production processes for product quality enhancement.

8.4 References

Addis, M., Piredda, G., Pes, M., Di Salvo, R., Scintu, M. F., & Pirisi, A. (2005). Effect of the use of three different lamb paste rennets on lipolysis of the PDO Pecorino Romano Cheese. International Dairy Journal, 15(6–9), 563–569. https://doi.org/10.1016/j.idairyj.2004.07.018

Andrén, A. (2021). Milk-Clotting Enzymes. Food Engineering Series, 349–362. https://doi.org/10.1007/978-3-030-55482-8_14

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical Methods, 5(16), 3790–3798. https://doi.org/10.1039/c3ay40582f

Caboni, P., Maxia, D., Scano, P., Addis, M., Dedola, A., Pes, M., Murgia, A., Casula, M., Profumo, A., & Pirisi, A. (2019). A gas chromatography-mass spectrometry untargeted metabolomics approach to discriminate Fiore Sardo cheese produced from raw or thermized ovine milk. Journal of Dairy Science, 102(6), 5005–5018. https://doi.org/10.3168/jds.2018-15885

Collins, Y. F., McSweeney, P. L. H., & Wilkinson, M. G. (2003). Lipolysis and free fatty acid catabolism in cheese: A review of current knowledge. International Dairy Journal, 13(11), 841–866. https://doi.org/10.1016/S0958-6946(03)00109-2

Conceição, C., Martins, P., Alvarenga, N., Dias, J., Lamy, E., Garrido, L., Gomes, S., Freitas, S., Belo, A., Brás, T., Paulino, A., & Duarte, M. F. (2018). Cynara cardunculus: Use in Cheesemaking and Pharmaceutical Applications. Technological Approaches for Novel Applications in Dairy Processing. https://doi.org/10.5772/intechopen.76530

FOLCH, J., LEES, M., & SLOANE STANLEY, G. H. (1957). A simple method for the isolation and purification of total lipides from animal tissues. The Journal of Biological Chemistry, 226(1), 497–509. https://doi.org/10.1016/s0021-9258(18)64849-5

García-Cano, I., Rocha-Mendoza, D., Kosmerl, E., & Jiménez-Flores, R. (2020). Purification and characterization of a phospholipid-hydrolyzing phosphoesterase produced by Pediococcus acidilactici isolated from Gouda cheese. Journal of Dairy Science, 103(5), 3912–3923. https://doi.org/10.3168/jds.2019-17965

Gomes, S., Belo, A. T., Alvarenga, N., Dias, J., Lage, P., Pinheiro, C., Pinto-Cruz, C., Brás, T., Duarte, M. F., & Martins, A. P. L. (2019). Characterization of Cynara cardunculus L. flower from Alentejo as a coagulant agent for cheesemaking. International Dairy Journal, 91, 178–184. https://doi.org/10.1016/j.idairyj.2018.09.010

Gominho, J., Lourenço, A., Palma, P., Lourenço, M. E., Curt, M. D., Fernández, J., & Pereira, H. (2011). Large scale cultivation of Cynara cardunculus L. for biomass production-A case study. Industrial Crops and Products, 33(1), 1–6. https://doi.org/10.1016/j.indcrop.2010.09.011

Haddad, L., Francis, J., Rizk, T., Akoka, S., Remaud, G. S., & Bejjani, J. (2022). Cheese characterization and authentication through lipid biomarkers obtained by high-resolution 1H NMR profiling. Food Chemistry, 383. https://doi.org/10.1016/j.foodchem.2022.132434

Hajian, G., Behinaein, B., Etemad, A., & Morin, E. (2022). Bagged tree ensemble modelling with feature selection for isometric EMG-based force estimation. Biomedical Signal Processing and Control, 78, 104012. https://doi.org/10.1016/j.bspc.2022.104012

Jenske, R., & Vetter, W. (2009). Concentrations of medium-chain 2- and 3-hydroxy fatty acids in foodstuffs. Food Chemistry, 114(3), 1122–1129. https://doi.org/10.1016/j.foodchem.2008.10.067

Liu, Z., Li, C., Pryce, J., & Rochfort, S. (2020). Comprehensive Characterization of Bovine Milk Lipids: Phospholipids, Sphingolipids, Glycolipids, and Ceramides. Journal of Agricultural and Food Chemistry, 68(24), 6726–6738. https://doi.org/10.1021/acs.jafc.0c01604

Moschopoulou, E. (2011). Characteristics of rennet and other enzymes from small ruminants used in cheese production. Small Ruminant Research, 101(1–3), 188–195. https://doi.org/10.1016/j.smallrumres.2011.09.039

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy, 54(3), 413–419. https://doi.org/10.1366/0003702001949500

Pacifico, S., Caputo, E., Piccolella, S., & Mandrich, L. (2024). Exploring New Fruit- and Vegetable-Derived Rennet for Cheese Making. Applied Science, 14(6), 2257. https://doi.org/https://doi.org/10.3390/app14062257

Patton, S., & Keenan, T. W. (1975). The Milk fat globule membrane. BBA - Reviews on Biomembranes, 415(3), 273–309. https://doi.org/10.1016/0304-4157(75)90011-8

Pes, M., Salvatore, E., Addis, M., Furesi, S., Di Salvo, R., Fiori, M., Piga, C., & Pirisi, A. (2014). Produzione di un formaggio a pasta cotta da latte di pecora. Scienza e Tecnica Lattiero-Casearia, 65(5–6), 143–157.

Prieto, B., Franco, I., Fresno, J. M., Prieto, J. G., Bernardo, A., & Carballo, J. (2004). Effect of ripening time and type of rennet (farmhouse rennet from kid or commercial calf) on proteolysis during the ripening of León cow milk cheese. Food Chemistry, 85(3), 389–398. https://doi.org/10.1016/j.foodchem.2003.07.016

Qin, Y., Lou, Z., Wang, Y., Lu, S., & Sun, P. (2022). An analytical partial least squares method for process monitoring. Control Engineering Practice, 124. https://doi.org/10.1016/j.conengprac.2022.105182

Ramos, P. A. B., Guerra, A. R., Guerreiro, O., Freire, C. S. R., Silva, A. M. S., Duarte, M. F., & Silvestre, A. J. D. (2013). Lipophilic extracts of cynara cardunculus L. var. altilis (DC): A source of valuable bioactive terpenic compounds. Journal of Agricultural and Food Chemistry, 61(35), 8420–8429. https://doi.org/10.1021/jf402253a

Rampanti, G., Raffo, A., Melini, V., Moneta, E., Nardo, N., Saggia Civitelli, E., Bande-De León, C., Tejada Portero, L., Ferrocino, I., Franciosa, I., Cardinali, F., Osimani, A., & Aquilanti, L. (2023). Chemical, microbiological, textural, and sensory characteristics of pilot-scale Caciofiore cheese curdled with commercial Cynara cardunculus rennet and crude extracts from spontaneous and cultivated Onopordum tauricum. Food Research International, 173. https://doi.org/10.1016/j.foodres.2023.113459

Sanjuán, E., Millán, R., Saavedra, P., Carmona, M. A., Gómez, R., & Fernández-Salguero, J. (2002). Influence of animal and vegetable rennet on the physicochemical characteristics of Los Pedroches cheese during ripening. Food Chemistry, 78(3), 281–289. https://doi.org/10.1016/S0308-8146(02)00098-5

Santillo, A., Albenzio, M., Quinto, M., Caroprese, M., Marino, R., & Sevi, A. (2009). Probiotic in lamb rennet paste enhances rennet lipolytic activity, and conjugated linoleic acid and linoleic acid

content in pecorino cheese. Journal of Dairy Science, 92(4), 1330–1337. https://doi.org/10.3168/jds.2008-1598

Scano, P., Cusano, E., Caboni, P., & Consonni, R. (2019). NMR metabolite profiles of dairy: A review. International Dairy Journal, 90, 56–67. https://doi.org/10.1016/j.idairyj.2018.11.004

Şengül, M., Erkaya, T., Dervişoğlu, M., Aydemir, O., & Gül, O. (2014). Compositional, biochemical and textural changes during ripening of Tulum cheese made with different coagulants. International Journal of Dairy Technology, 67(3), 373–383. https://doi.org/10.1111/1471-0307.12120

Shah, M. A., Mir, S. A., & Paray, M. A. (2014). Plant proteases as milk-clotting enzymes in cheesemaking: A review. Dairy Science and Technology, 94(1), 5–16. https://doi.org/10.1007/s13594-013-0144-3

Timón, M. L., Andrés, A. I., Otte, J., & Petrón, M. J. (2019). Antioxidant peptides (<3 kDa) identified on hard cow milk cheese with rennet from different origin. Food Research International, 120, 643–649. https://doi.org/10.1016/j.foodres.2018.11.019

Tomaiuolo, M., Nardelli, V., Mentana, A., Campaniello, M., Zianni, R., & Iammarino, M. (2023). Untargeted Lipidomics and Chemometric Tools for the Characterization and Discrimination of Irradiated Camembert Cheese Analyzed by UHPLC-Q-Orbitrap-MS. Foods, 12(11). https://doi.org/10.3390/foods12112198

Verardo, V., Gómez-Caravaca, A. M., Gori, A., Losi, G., & Caboni, M. F. (2013). Bioactive lipids in the butter production chain from parmigiano reggiano cheese area. Journal of the Science of Food and Agriculture, 93(14), 3625–3633. https://doi.org/10.1002/jsfa.6152

Xu, Q. S., & Liang, Y. Z. (2001). Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, 56(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Zhao, L., Zhang, J., Ge, W., & Wang, J. (2022). Comparative Lipidomics Analysis of Human and Ruminant Milk Reveals Variation in Composition and Structural Characteristics. Journal of Agricultural and Food Chemistry, 70(29), 8994–9006. https://doi.org/10.1021/acs.jafc.2c02122

Chapter 9: Conclusions

This chapter recapitulates the main aspects of this dissertation, sums up the conclusive remarks and highlights the future perspectives

Cheese is a multifunctional fermented food resulting from a sequence of processes that transform raw milk into a stable product. This transformation involves biochemical, physical, and microbiological changes that occur on each production step and contribute to the final quality attributes. Another issue involves the industrialization of dairy product manufacturing that reflects the PDO-protected recipe but involves the use of practices that are complex from a food safety perspective, such as the use of raw milk. This aspect represents one of the most controversial issues for institutions in the relationship between food safety and gastronomic tradition. By understanding and controlling the effects of intermediate manufacturing stages it could be possible to ensure product consistency and maintain a balance between food safety and the protection of dairy traditions.

In this thesis, the problem of quality monitoring of cheesemaking process has been addressed. The production of cheese was examined as a stepwise process, where each intermediate product was subjected to investigation. The achievement of a comprehensive product quality monitoring may be challenging if considering that cheese is featured by inherent variability associated to a strong mutability of feedstock properties and different production stimuli. Such sources of variability significantly affect the growth of microorganisms involved, the chemico-phisical properties of the intermediate items and consequently, the desired product quality. These aspects were addressed by following the sequential pathway of cheesemaking, starting from milk and continuing towards the final product.

The development of a milk coagulation monitoring system, through the application of Raman spectroscopy coupled with multivariate statistical analysis allowed to deal with the first issue of cheese production: real-time prediction of cutting time using an indirect and noninvasive measure of gelation evolution. To this concern, the spectra information was condensed through dimensionality reduction algorithms and exploited to predict the technological time of coagulation interruption (Cutting time), demonstrating that Raman spectroscopy can effectively provide comparable benefits to those obtained with conventional techniques already reported in the literature. The following step was to further exploit Raman spectroscopy power by providing a chemical knowledge of the gelation process, which is prevented if using other real-time monitoring tools such as NIR or fluorescence. After identifying the key vibrational modes of the process, the dynamic behavior of Principal Component scores obtained from Raman spectra and elastic modulus were embedded within a state estimator (Kalman filter) supported by a kinetic model to keep track of the state of the process during coagulation. Such a study allowed to achieve a time by

time knowledge of the chemical and physical properties of the system, which usually requires invasive sampling or atline measurements. The final step of the monitoring problem was to push it towards an industrial perspective, by developing a fault detection system. Different process fault were induced on coagulating milk in order to test whether the Raman sensor coupled with multivariate statistical process control is able to provide a spectral fingerprint of such failures and to timely trigger the alarm. To conclude the application of PAT tools on milk coagulation assays, Raman spectroscopy was applied to monitor the combined acid-enzymatic coagulation of milk, meaning that both bacteria and rennet were used to conduct gelation. In this preliminary study Raman signal was correlated to pH value as well as different analytes' concentration, whose actual value was measured through phmeter and LC-MS/MS, respectively. Specifically, PLS regression models were coupled with Raman spectroscopy to provide a noninvasive and fast quantitative estimation of pH, glucose concentration, galactose concentration, lactose concentration and lactic acid concentration. Overall, the results showed that good to high performances can be achieved when using Raman spectroscopy within the PAT paradigm of milk coagulation, thus extending the range of possible monitoring strategies.

The subsequent part of this thesis focused on the metabolomic analysis of cheese aimed to study the effect of multiple sources of variability: feedstock variability, induced by seasonal changes of milk, and technological variability, which included different thermal treatments, rennet types and ripening times. The metabolomic analysis identified a wide set of biomarkers that have the potential to contribute to assess the technological compliance to PDO label guidelines for the specific Sardinian cheese product. As for the case of feedstock properties, the research effort presented in this dissertation showed how the seasonal variation of grazing feed properties can unavoidably change the nutritional content of the final product. Particularly, for the first time it has been here reported that the fatty acid chain length gradually increases when milking season moves from January to June, proving that part of cheese nutritional profile can be precisely determined by knowing the production period. Such an investigation was put to test by the presence of missing data in the spectroscopical dataset, which were successfully handled through probabilistic modeling. Moreover, various groups of metabolites such as biogenic amines, sugars, organic acids and endocannabinoids, emerged as biomarkers associated with different milk thermal treatments, cheese ripening times, and their interaction, highlighting that the individual steps of cheesemaking are interdependent and must be considered collectively to fully understand their impact on the final product. Indeed, all the compounds investigated in chapter 5 (combined acid-enzymatic

coagulation of milk) resulted to be significantly affected by other processing steps, when metabolomic analysis was performed. Finally, a lipidomic study allowed to address the problem of rennet type classification, underlining a broad picture of lipidic biomarkers, encompassing various classes such as triglycerides, phospholipids, sphingolipids and free fatty acids, which resulted to be upregulated in animal rennet-based cheeses. This diversity reflected the significant dissimilarities in the lytic behavior of two coagulants from different sources (i.e. vegetable and animal). In this case as well, the factor of interest was investigated in conjunction with ripening time, further supporting the idea that process variables act synergistically along the production line. However, the classification task was initially undermined by the presence of numerous uninformative features. To tackle this issue, a variable selection algorithm combining principles from both univariate and multivariate statistics was applied, allowing the identification and removal of irrelevant variables. This approach enabled high predictive performance across various factors related to cheese production. The mathematical models employed in the thesis were experimentally validated by means of cross-validation and external validation techniques, supporting the robustness of the employed methodology.

The extraction of meaningful information from experimental datasets is not always immediate, due to noise, redundancy, and the multifactorial nature of the systems involved. It is therefore essential to carefully process data and to select, case by case, the most suitable analytical approach. The rationale presented herein can support future research in improving the analytical techniques and extend the operational variability. In particular, the following challenges encountered during the course of this work have to be tackled to enable its potential application within an industrial context:

1) Improvement of signal to noise ratio from Raman spectroscopy data: the analytical procedure for real time monitoring of milk coagulation deserve to be enhanced to further validate the robustness of the methodology. By improving the spectroscopical signal it is possible to achieve a better parameters estimation and to explore multiple factors ($CaCl_2$ concentration, protein and fat content, initial pH and so on).
2) Employment of Raman spectroscopy for curd composition: Raman spectroscopy has been widely used in literature for the detection of milk adulterant. Despite this work demonstrated that Raman sensors can provide valuable information regarding the

compounds concentration involved in milk renneting, further investigation may allow to extend the chemical characterization of the process after curd cutting.

3) Improve classification performance when dealing with process variables related to different production steps: the complex interconnection between thermal treatment and ripening can pose difficulties in achieving outstanding classification performances. Using multiple sources of data can provide high accuracies. The presented methodology could be further extended into data fusion analysis for the enhancement of the classification models and biomarker identification algorithms.
4) The extension of the presented methodologies to other dairy matrices holds the potential to further validate models presented herein. Despite the cheesemaking procedures well replicate the real-world scenarios, the conclusions drawn from the presented results are valid for the specific PDO products here investigated.

Overall, the results presented in this dissertation align with current demands in dairy research, particularly in the development of data-driven strategies for process monitoring, quality assessment, and product authentication. It is desirable that the research effort shown in doctoral thesis may serve as a starting point for addressing the ongoing challenges posed by the food industry.

APPENDIX A

Eq. 2.50 is obtained by building a portioned vector $\underline{z} = (\underline{t}, \underline{x})$. By applying the chain rule of probability, it is possible to express the joint probability $p(\underline{z}) = p(\underline{t}, \underline{x})$ as a function of the conditional probability $p(\underline{x}|\underline{t})$ and the marginal distribution $p(\underline{x})$:

$$p(\underline{t}, \underline{x}) = p(\underline{t})p(\underline{x}|\underline{t}) \tag{A.1}$$

By considering the natural logarithm of eq. A1 it is possible to obtain the following form:

$$ln\left(p(\underline{t}, \underline{x})\right) = ln\left(p(\underline{t})\right) + ln\left(p(\underline{x}|\underline{t})\right) \tag{A.2}$$

By substituting the distribution function indicated in eq. 2.49 to $\mathrm{p}(\underline{\mathrm{x}}|\underline{\mathrm{t}})$ and the assumption of $\underline{t} \sim N(\underline{0}, \underline{\mathrm{I}})$ to $\mathrm{p}(\underline{\mathrm{t}})$ eq. A.2 becomes:

$$ln\left(p(\underline{t}, \underline{x})\right) = -\frac{1}{2}\underline{t}^T\underline{t} - \frac{1}{2}\left(\underline{x} - \underline{\underline{W}}\,\underline{t} - \underline{\mu}\right)^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\left(\underline{x} - \underline{\underline{W}}\,\underline{t} - \underline{\mu}\right) + \text{const} \tag{A.3}$$

Where 'const' denotes terms independent from $\underline{\mathrm{t}}$ and $\underline{\mathrm{x}}$. Eq. A.3 is a quadratic function of $\underline{\mathrm{z}}$, which leads to the following expression:

$$ln\left(p(\underline{t}, \underline{x})\right) = -\frac{1}{2}\underline{t}^T\underline{t} - \frac{1}{2}\underline{x}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{x} + \frac{1}{2}\underline{x}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\underline{W}}\,\underline{t} + \frac{1}{2}\underline{x}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\mu} +$$
$$\frac{1}{2}\underline{t}^T\underline{\underline{W}}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{x} - \frac{1}{2}\underline{t}^T\underline{\underline{W}}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\underline{W}}\,\underline{t} - \frac{1}{2}\underline{t}^T\underline{\underline{W}}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\mu} - \frac{1}{2}\underline{\mu}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\underline{W}}\,\underline{t} +$$
$$-\frac{1}{2}\underline{\mu}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{x} + const \tag{A.4}$$

In eq. A.4 all the terms that do not involve $\underline{\mathrm{t}}$ or $\underline{\mathrm{x}}$ have been incorporated in the constant term. For the present discussion, it is more suitable to compute the conditional variance matrix. By definition of Gaussian variable, if $\underline{\mathrm{x}}$ is set as constant within eq. 4, the resulting distribution parameters refer to $\underline{\mathrm{t}}|\underline{\mathrm{x}}$. Such an operation can be performed because $\underline{\mathrm{x}}$ is regarded as constant when considering $\underline{\mathrm{t}}|\underline{\mathrm{x}}$. This means that, the reciprocal of the coefficient of the quadratic term (highlighted in red) in eq. A.4 is the variance of $\mathrm{p}(\underline{\mathrm{t}}|\underline{\mathrm{x}})$:

$$QT = \frac{1}{2}\underline{t}^T\left(\underline{\underline{I}} + \underline{\underline{W}}^T\left(\sigma^2\underline{\underline{I}}\right)^{-1}\underline{\underline{W}}\right)\underline{t} \tag{A.5}$$

From eq. A.5 it can be said that:

$$var(\underline{t}|\underline{x}) = \left(\underline{\underline{I}} + \underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\underline{W}}\right)^{-1} = \sigma^2 \left(\sigma^2 \underline{\underline{I}} + \underline{\underline{W}}^T \underline{\underline{W}}\right)^{-1} = \sigma^2 \underline{\underline{M}}^{-1} \quad \text{(A.6)}$$

Which corresponds to the variance reported in eq. 2.50. Analogously, the expected value of $\underline{t}$ given the occurrence of $\underline{x}$ can be obtained by taking into consideration the linear terms (LT) of eq. A.4:

$$LT = \frac{1}{2}\underline{x}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\underline{W}}\, \underline{t} + \frac{1}{2}\, \underline{t}^T \underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{x} - \frac{1}{2}\underline{t}^T \underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\mu} - \frac{1}{2}\underline{\mu}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\underline{W}}\, \underline{t} \quad \text{(A.7)}$$

By stating that $\underline{\mu}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\underline{W}}\, \underline{t} = \underline{t}^T \underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \underline{\mu}^T$, Eq. A.5 can be rearranged as follows:

$$LT = \underline{t}^T \left[\underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \left(\underline{x} - \underline{\mu}\right)\right] \quad \text{(A.8)}$$

The coefficient of $\underline{t}^T$ in eq. A.8 corresponds, by definition, to $[var(\underline{t}|\underline{x})]^{-1} E(\underline{t}|\underline{x})$. This means that:

$$E(\underline{t}|\underline{x}) = var(\underline{t}|\underline{x}) \left[\underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \left(\underline{x} - \underline{\mu}\right)\right] = \sigma^2 \underline{\underline{M}}^{-1} \underline{\underline{W}}^T \left(\sigma^2 \underline{\underline{I}}\right)^{-1} \left(\underline{x} - \underline{\mu}\right) = \underline{\underline{M}}^{-1} \underline{\underline{W}}^T \left(\underline{x} - \underline{\mu}\right) \quad \text{(A.9)}$$

Which is consistent to eq. 2.50.

APPENDIX B

In this Appendix, the mathematical steps needed to derive the coagulation model (Eq. 3.8) are reported.

The ODE resulting from the first stage can be written as eq. B.1:

$$\begin{cases} \frac{dS_1}{dt} = -\alpha_1 (S_1 - S_{1,\infty}) \\ \frac{dS_2}{dt} = n\ \alpha_2 (S_1 - S_{1,\infty}) \\ \frac{dG'}{dt} = 0 \end{cases} \quad \text{(B.1)}$$

The equation is integrated in the [0,t] domain:

$$\begin{cases} \int_{S_{1,0}}^{S_1(t)} \frac{1}{S_1 - S_{1,\infty}} dS_1 = \int_0^t -\alpha_1 d\tau \\ \int_{S_{2,0}}^{S_2(t)} dS_2 = \int_0^t n\ \alpha_1 (S_1(\tau) - S_{1,\infty}) d\tau \\ G' = G'_0 \end{cases} \quad \text{(B.2)}$$

By solving eq. A.2, the following system of non-linear equations is obtained:

$$\begin{cases} \ln\left(\frac{S_1 - S_{1,\infty}}{S_{1,0} - S_{1,\infty}}\right) = -\alpha_1 t \\ S_2 - S_{2,0} = \int_0^t n\ \alpha_1 (S_1(\tau) - S_{1,\infty}) d\tau \\ G' = G'_0 \end{cases} \quad \text{(B.3)}$$

By solving the first equation for PC_1 and substituting it to the second equation it is possible to write down:

$$\begin{cases} S_1 = S_{1,\infty} + (S_{1,0} - S_{1,\infty}) e^{-\alpha_1 t} \\ S_2 - S_{2,0} = \int_0^t n\ \alpha_1 (S_{1,\infty} + (S_{1,0} - S_{1,\infty}) e^{-\alpha_1 \tau} - S_{1,\infty}) d\tau \\ G' = G'_0 \end{cases} \quad \text{(B.4)}$$

Eq. B.4 can be algebraically rearranged as follows:

$$\begin{cases} S_1 = S_{1,\infty} + (S_{1,0} - S_{1,\infty}) e^{-\alpha_1 t} \\ S_2 - S_{2,0} = n\ \alpha_1 (S_{1,0} - S_{1,\infty}) \int_0^t e^{-\alpha_1 \tau} d\tau \\ G' = G'_0 \end{cases} \quad \text{(B.5)}$$

Solving this equation leads to the final expression of the model in the hydrolysis step:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,0}+n\left(S_{1,0}-S_{1,\infty}\right)\left(1-e^{-\alpha_1 t}\right) \\ G'=G'_0 \end{cases} \quad \text{(B.6)}$$

In the aggregation step, the term for S_2 consumption and G' generation is considered:

$$\begin{cases} \frac{dS_1}{dt}=-\alpha_1(S_1-S_{1,\infty}) \\ \frac{dS_2}{dt}=n\,\alpha_1(S_1-S_{1,\infty})-\alpha_2(S_2-S_{2,\infty}) \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.7)}$$

The first equation of B.7 is not affected by the hydrolysis to aggregation stage transition, and therefore its solution does not change along the time span. Conversely, the solutions for S_2 and G' require further manipulations:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ \frac{dS_2}{dt}=n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t}-\alpha_2(S_2-S_{2,\infty}) \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.8)}$$

Both sides of balance equation for S_2 are multiplied times $e^{\alpha_2 t}$, and S_2 linear term is moved in the left-hand side of the equation:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ e^{\alpha_2 t}\frac{dS_2}{dt}+\alpha_2(S_2-S_{2,\infty})e^{\alpha_2 t}=n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)e^{(\alpha_2-\alpha_1)t} \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.9)}$$

By applying the integrating factor method, it is possible to obtain:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ \frac{de^{\alpha_2 t}(S_2-S_{2,\infty})}{dt}=n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)e^{(\alpha_2-\alpha_1)t} \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.10)}$$

Eq. B.10 can be integrated in both sides, giving rise to the following system:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ \int_{e^{\alpha_2 t^*}(S_2-S_{2,\infty})}^{e^{\alpha_2 t}(S_2-S_{2,\infty})} d\left[e^{\alpha_2\tau}(S_2-S_{2,\infty})\right]=\int_{t^*}^{t} n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)e^{(\alpha_2-\alpha_1)\tau}d\tau \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.11)}$$

By solving the integral, the following system is thus written down:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ e^{\alpha_2 t}(S_2-S_{2,\infty})-e^{\alpha_2 t^*}(S_{2,t^*}-S_{2,\infty})=\frac{n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_2-\alpha_1}\left[e^{(\alpha_2-\alpha_1)t}-e^{(\alpha_2-\alpha_1)t^*}\right] \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.12)}$$

Here, S_{2,t^*} is the initial condition for the second stage. Such a value can be evaluated by substituting t^* to t in S_2 balance equation from system B.6. Therefore, this term can be used to derive the following form:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ e^{\alpha_2 t}(S_2-S_{2,\infty})-e^{\alpha_2 t^*}\left[S_{2,0}+n\left(S_{1,0}-S_{1,\infty}\right)\left(1-e^{-\alpha_1 t^*}\right)-S_{2,\infty}\right]=\frac{n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_2-\alpha_1}\left[e^{(\alpha_2-\alpha_1)t}-e^{(\alpha_2-\alpha_1)t^*}\right] \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.13)}$$

S_2 can be now isolated and both sides of Eq. S.13 are multiplied times $e^{-\alpha_2 t^*}$:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,\infty}+e^{\alpha_2(t^*-t)}\left[S_{2,0}+n\left(S_{1,0}-S_{1,\infty}\right)\left(1-e^{-\alpha_1 t^*}\right)-S_{2,\infty}\right]+\frac{n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_2-\alpha_1}\left[e^{-\alpha_1 t}-e^{(\alpha_2-\alpha_1)t^*-\alpha_2 t}\right] \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.14)}$$

In the following step, the multiplication operations are performed and terms are conveniently grouped up:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,\infty}+S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)e^{\alpha_2(t^*-t)}+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]e^{\alpha_2 t^*}e^{-\alpha_2 t}-n\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t^*+\alpha_2 t^*}e^{-\alpha_2 t}-\frac{n\,\alpha_1\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_1-\alpha_2}\left[e^{-\alpha_1 t}-e^{(\alpha_2-\alpha_1)t^*-\alpha_2 t}\right] \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.15)}$$

For the sake of practicality, the first two constant terms of right hand side of S_2 balance are renamed as C_1, while the coefficient of $e^{-\alpha_2 t}$ in the third term is renamed as C_2. Furthermore, $e^{-\alpha_2 t}$ is collected from fourth and fifth terms:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=C_1+C_2e^{-\alpha_2 t}+n\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_2 t}\left\{-e^{-\alpha_1 t^*+\alpha_2 t^*}-\frac{\alpha_1}{\alpha_1-\alpha_2}\left[e^{-\alpha_1 t+\alpha_2 t}-e^{(\alpha_2-\alpha_1)t^*}\right]\right\} \\ \frac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \quad \text{(B.16)}$$

By developing the term inside the curly brackets of A.16 and performing the least common multiple it is possible to land into the following expression:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2= C_1+C_2e^{-\alpha_2 t}+\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_1-\alpha_2}e^{-\alpha_2 t}\left(\alpha_2 e^{-\alpha_1 t^*+\alpha_2 t^*}-\alpha_1 e^{-\alpha_1 t+\alpha_2 t}\right) \\ \dfrac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \tag{B.17}$$

For the sake of convenience, the whole term inside the round brackets is multiplied and divided by $e^{-\alpha_1(t+t^*)}$:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2= C_1+C_2e^{-\alpha_2 t}+\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)}{\alpha_1-\alpha_2}e^{-\alpha_2 t}e^{-\alpha_1(t+t^*)}\left(\alpha_2 e^{\alpha_1 t+\alpha_2 t^*}-\alpha_1 e^{\alpha_2 t+\alpha_1 t^*}\right) \\ \dfrac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \tag{B.18}$$

The final expression of S_2 balance can be derived out by substituting C_1 and C_2, and rearranging $n\left(S_{1,0}-S_{1,\infty}\right)$ term:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,\infty}+S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)e^{\alpha_2(t^*-t)}+e^{-\alpha_2 t}\left\{\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)\left(\alpha_2 e^{\alpha_1 t+\alpha_2 t^*}-\alpha_1 e^{\alpha_2 t+\bar{k}_1 t^*}\right)e^{-\alpha_1(t+t^*)}}{\alpha_1-\alpha_2}+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]e^{\alpha_2 t^*}\right\} \\ \dfrac{dG'}{dt}=m\alpha_2(S_2-S_{2,\infty}) \end{cases} \tag{B.19}$$

By substituting the second equation of B.19 into the ODE of G' balance, it is possible to integrate both sides of the third equation.

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,\infty}+S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)e^{\alpha_2(t^*-t)}+e^{-\alpha_2 t}\left\{\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)\left(\alpha_2 e^{\alpha_1 t+\alpha_2 t^*}-\alpha_1 e^{\alpha_2 t+\alpha_1 t^*}\right)e^{-\alpha_1(t+t^*)}}{\alpha_1-\alpha_2}+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]e^{\alpha_2 t^*}\right\} \\ \int_{G_0'}^{G'} dG_d' = \int_{t^*}^{t} m\alpha_2 S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)e^{\alpha_2(t^*-\tau)}+e^{-\alpha_2\tau}\left\{\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)\left(\alpha_2 e^{\alpha_1 \tau+\alpha_2 t^*}-\alpha_1 e^{\alpha_2 \tau+\alpha_1 t^*}\right)e^{-\alpha_1(\tau+t^*)}}{\alpha_1-\alpha_2}+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]e^{\alpha_2 t^*}\right\}d\tau \end{cases} \tag{B.20}$$

Again, dummy integration variables are used to avoid any conflicts of notation. After solving the integrals and isolating for G', the following equation is obtained:

$$\begin{cases} S_1=S_{1,\infty}+\left(S_{1,0}-S_{1,\infty}\right)e^{-\alpha_1 t} \\ S_2=S_{2,\infty}+S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)e^{\alpha_2(t^*-t)}+e^{-\alpha_2 t}\left\{\dfrac{n\left(S_{1,0}-S_{1,\infty}\right)\left(\alpha_2 e^{\alpha_1 t+\alpha_2 t^*}-\alpha_1 e^{\alpha_2 t+\alpha_1 t^*}\right)e^{-\alpha_1(t+t^*)}}{\alpha_1-\alpha_2}+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]e^{\alpha_2 t^*}\right\} \\ G'=G_0'+m\left\{S_{2,0}\left(1-e^{-\alpha_1 t^*}\right)\left(1-e^{\alpha_2(t^*-t)}\right)+\dfrac{\alpha_2}{\alpha_1-\alpha_2}n\left(S_{1,0}-S_{1,\infty}\right)\left(e^{\alpha_1 t^*}-e^{-\alpha_2(t^*-t)-\alpha_1 t^*}\right)+\dfrac{\alpha_2}{\alpha_1-\alpha_2}n\left(S_{1,0}-S_{1,\infty}\right)\left(e^{-\alpha_1 t}-e^{-\alpha_1 t^*}\right)+\left[n\left(S_{1,0}-S_{1,\infty}\right)-S_{2,\infty}\right]\left(1-e^{-\alpha_2(t^*-t)}\right)\right\} \end{cases} \tag{B.21}$$

It is possible to rearrange the terms in equation B.21 to obtain a more compact form, which corresponds to the final expression of the aggregation model presented in this work (Eq. 3.8):

$$\begin{cases} S_1 = S_{1,\infty} + \left(S_{1,0} - S_{1,\infty}\right) e^{-\alpha_1 t} \\ S_2 = S_{2,\infty} + S_{2,0}\left(1 - e^{-\alpha_1 t^*}\right) e^{\alpha_2 (t^* - t)} + e^{-\alpha_2 t} \left\{ \frac{n\left(S_{1,0} - S_{1,\infty}\right)\left(\alpha_2 e^{\alpha_1 t + \alpha_2 t^*} - \alpha_1 e^{\alpha_2 t + \alpha_1 t^*}\right) e^{-\alpha_1 (t + t^*)}}{\alpha_1 - \alpha_2} + \left[n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty}\right] e^{\alpha_2 t^*} \right\} \\ G' = G'_0 + m\, S_{2,0}\left(1 - e^{-\alpha_1 t^*}\right)\left(1 - e^{-\alpha_2 (t^* - t)}\right) + m \left\{ n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty} + \frac{n\,\alpha_2 \left(S_{1,0} - S_{1,\infty}\right) e^{-\alpha_1 t}}{\alpha_1 - \alpha_2} - \frac{n\,\alpha_2 \left(S_{1,0} - S_{1,\infty}\right) e^{-\alpha_2 t - \alpha_1 t^* + \alpha_2 t^*}}{\alpha_1 - \alpha_2} - \left[n\left(S_{1,0} - S_{1,\infty}\right) - S_{2,\infty}\right] e^{-\alpha_2 (t + t^*)} \right\} \end{cases} \quad \text{(B.22)}$$

For which the initial condition is $S_{1,0}=1$, $S_{2,0}=0$ and $G'_0 = G'_{0,exp}$.

APPENDIX C

C.1 Fault detection algorithms

Four additional algorithms were investigated for identifying faulty batches among the coagulation assays:

PC scores confidence interval (SCI) algorithm: the PCA model was trained using calibration samples, and the upper control limit (UCL) and lower control limit (LCL) for PC1 were determined. These limits were calculated as the mean PC1 trajectory from calibration set plus or minus the confidence interval derived from the t-distribution with a confidence level 1-α. The calibration dataset was also used to compute the loading values, which were subsequently employed for projecting Raman spectra from validation samples (3 NOC+5 faults) into the principal component subspace. If all projected values in a pre-selected time window exceeded the UCL or fell below the LCL, the sample was classified as a fault. The window was moved until a fault was detected or all acquisitions of the same experiment were considered. Different consecutive number of points to raise the alarm (window size) and significance levels were employed to identify the combination that maximizes the accuracy.

T^2-based fault detection: T^2 statistic was computed after projecting each validation sample onto the PC space using loading values derived from training sample spectra. The operation was iterated across the time after rennet addition, enabling the T^2 values to reflect the local data variance, which was compared to the limit value. If T^2 values exceeded the threshold for the predefined consecutive acquisitions, a fault was detected.

SPE-based fault detection: SPE statistics were computed after projecting each validation sample onto the PC space using loading values derived from training sample spectra. The operation was repeated for each acquisition time, providing an assessment of the unexplained local data variance. SPE value was therefore compared with SPE_{lim} for the predefined consecutive acquisitions. Again, if the threshold values were exceeded for all the consecutive points, the fault alarm was triggered.

SPE and T2-based algorithm: SPE and T^2 statistics are merged in a single fault detection algorithm. In this scenario the alarm is triggered if both SPE and T2 values fall above the control limit.

C.2 Results: Score confidence interval algorithm for detecting milk coagulation faults

The fault detection rationale adopted in the SCI algorithm (1) refers to Figure 1 in the main text, illustrating how the batch faults deviate from the family of trajectories associated with NOC. The results of the detection algorithm (1) are displayed in figure C1, which highlights the accuracy in cross validation obtained under different hyperparameters combinations.

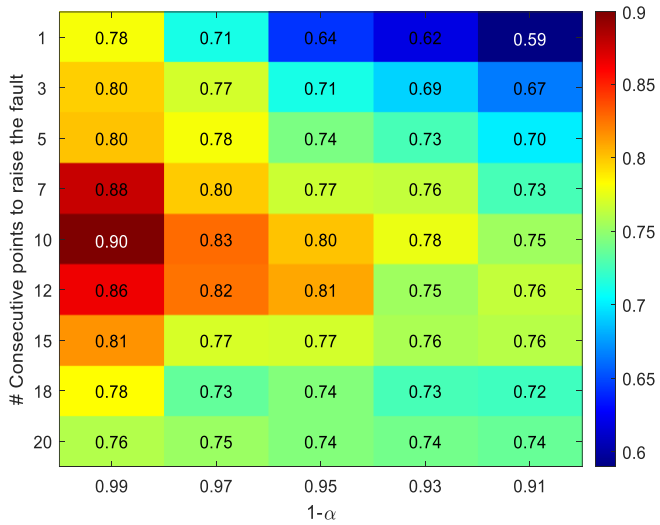


Figure C1 Accuracy as a function of hyperparameters for SCI using one PC to calibrate the PCA model. The accuracy is indicated by a colorbar.

As reported, the optimal hyperparameters combination reveals that the maximum accuracy (90%) is achieved when the algorithm (1) was tuned with a confidence level of 99% and a number of points to raise the fault of 10, which corresponds to a delay of 5 minutes after the trajectory starts falling outside the NOC area. Under these conditions, the FDR resulted to be equal to 96%, along with a FAR of 18%. Such an outcome indicates that this algorithm holds high sensitivity to faults.

Particular attention is paid to the thermal fault batch, as it allows for a direct comparison of the temporal behavior of Raman spectra with the reactor temperature profile. In this regard, figure C2 displays the dynamic evolution of the reactor's temperature and PC1 for the thermal fault batch. Despite the qualitative trend of the score values resembles the ones observed for NOC batches, an abnormal behavior is observed in the first 15 minutes after enzyme addition, which corresponds to the steep drop of temperature. Probably, the variability explained by PC1 is the result of a

combined effect of temperature drop and casein hydrolysis. Interestingly, after approximately 15 minutes an elbow in temperature profile can be remarked along with an abrupt slope variation of PC1 scores. This finding probably indicates that once the system is approaching the thermal equilibrium, the greatest source of variability turns out to be only related to the renneting effect.

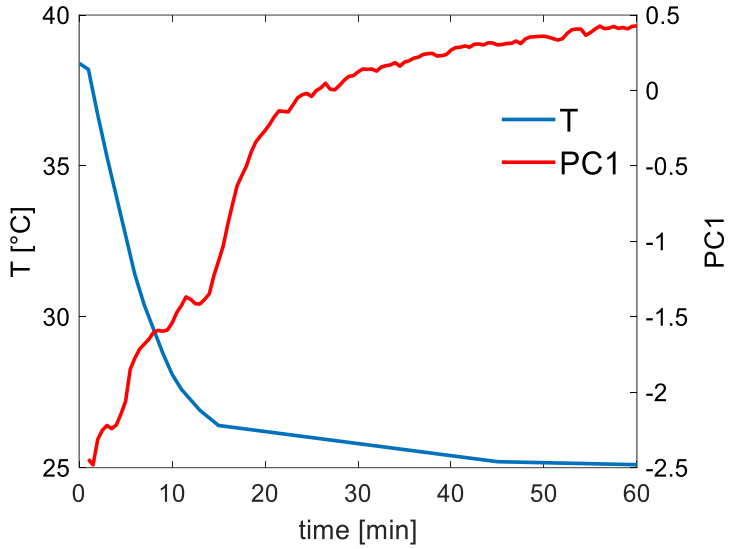


Figure C2 Time profile of reactor temperature and the corresponding PC1 trajectory for the thermal fault batch.

C.3 Results: Local PCA-based fault detection algorithms

Algorithms (3), (4) and (6) mentioned in the main text were explored, and the results are here reported (Figure C3).

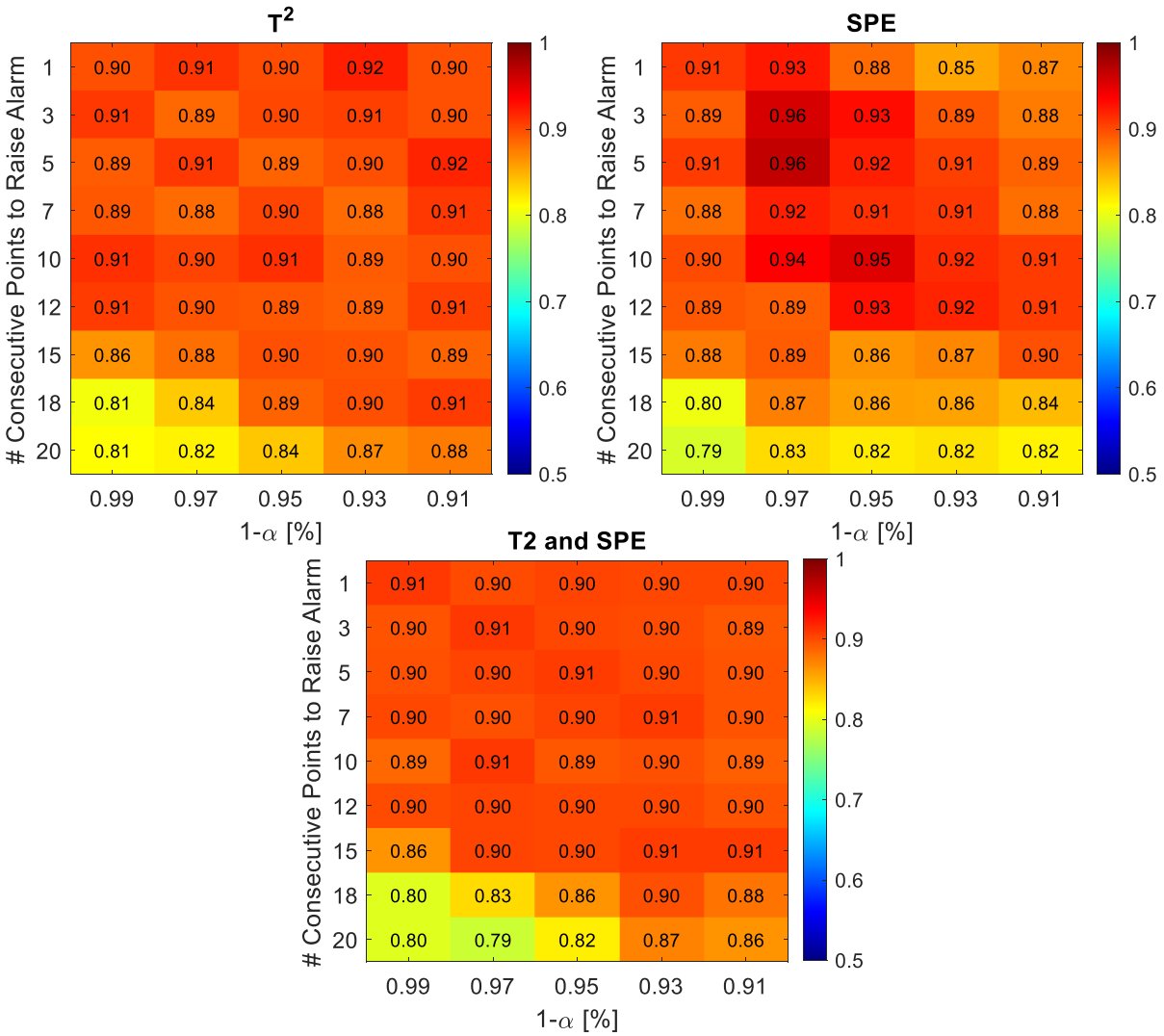
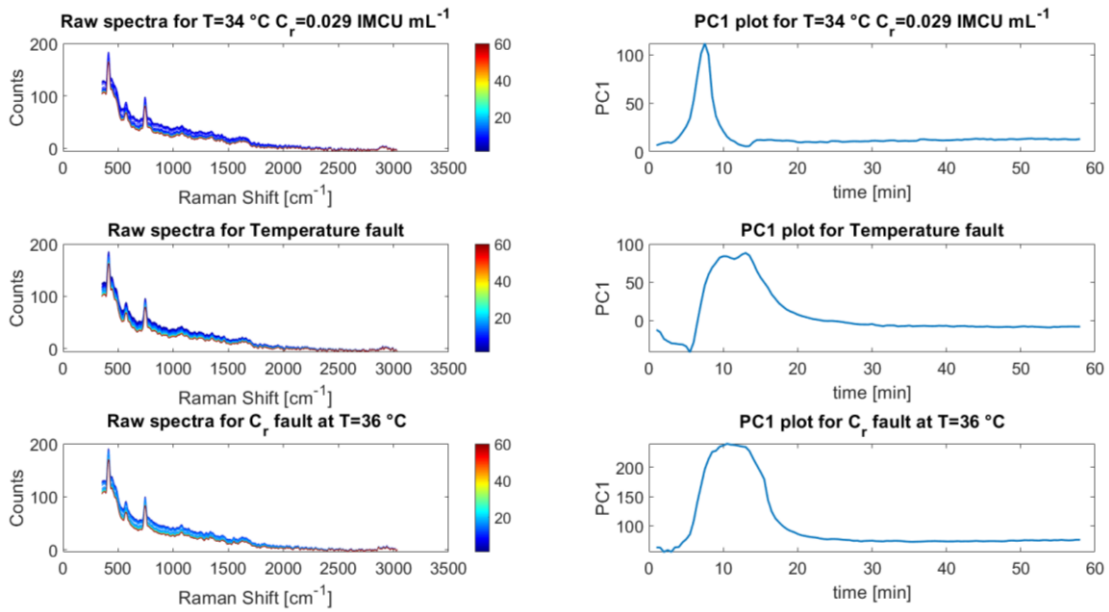


Figure C3 Accuracy as a function of the number of consecutive points and the confidence level for Algorithms (3), (4) and (6).

The best performance for the T^2-based detection algorithm was achieved using a significance level of 7% and a requirement of only one consecutive point to trigger an alarm, resulting in an accuracy of 92%. In contrast, the SPE-based detector reached its optimal performance with an α value of 3% and three consecutive points, yielding an accuracy of 96%. When combining both detectors using a logical AND approach, an alarm was triggered by a single point with a confidence level of 99%, leading to a maximum accuracy of 91%.

For the sake of completeness, the results depicted in Figure 4.8 in the main text are supplemented with the corresponding analysis of the raw spectra, along with the local-PC1 values (Figure C4). From a visual inspection of the raw spectra, it is possible to denote that the bigger portion of variability is observed in the earliest times after rennet addition. Such an outcome is consistent



with the PC1 plots, where the peak values are reached within the first 15 minutes.

Figure C4 Dynamic profile of Raw spectra (of which time is denoted with the color gradient) and their corresponding PC1 value.

APPENDIX D

Bagging methods are known to provide stable estimates as they mitigate variance through repeated bootstrap resampling and aggregation. Bagging exploits bootstrapped replicates from the training set to train multiple models and return as output the average of all the predictions of out of bags samples, which are used to test the predictive ability of the learners. This approach allows to reduce overfitting of a learning method. In bootstrap aggregating, each tree is trained on a bootstrap sample of the same size as the original dataset, drawn with replacement. Within this framework, predictor importance was derived from the reduction in out-of-bag prediction error associated with each feature. This procedure serves as a complementary assessment to verify the consistency of the features identified by the main method. The goal of this supplementary section is to show the common set metabolites identified by both the algorithm presented in chapter 8 and the bagged tree protocol. The investigated feature was identified as significant if the absolute value of the Out of Bag Permuted Predictor Delta Error was higher than 0.02. The bagged tree was tuned with 500 trees to ensure the convergence of the performance indices, while the number of predictor variables to select at random for each decision split was 32. Table D.1 shows the subset of metabolites consistently identified across the different selection methods. Concerning the rennet type classification, 354 metabolites were considered as informative, out of which 14 were also included within the algorithm reported in figure 8.2, while 365 were included when the bagged tree algorithm was tasked to classify different ripening times, out of which 8 were identified from the ANOVA-PLS-DA based model, shown in the main body of the text. The recurrence of these predictor variables suggests that they are informative biomarkers for discriminating among rennet types and ripening stages.

Table D.1 Common feature between the algorithm presented in chapter 8 and the ones outcoming the bagged tree-based feature selection method.

Rennet type		Ripening time	
FA (18:2)	PS 18:1_20:0	FA 11:0	LysoPE 18:1
U3	FA 24:2	FA 12:0	U2
Cer(d18:1/23:0)	Cer(d18:0/23:0)	hydroxy palmitic acid	hydroxy-octadecadienoic acid
PE (P_18:1_18:1)	SM (d 36:1)	FA 18:1	FA 15:1
PC (P_16:0_18:4)	SM d37:1		
PS (18:1_20:0)	PS 18:0_22:0		
PC 14:0	TG 51:5		

Copyright and permissions

Acknowledgements

I would like to express my deepest gratitude to my Supervisor, Professor Stefania Tronci, and my Co-supervisor, Professor Massimiliano Grosso, for their constant support and openness, qualities that have played a key role throughout my doctoral studies. Their experience has been deeply instructive in shaping my understanding of how rigorous scientific research is conducted.

Thanks to Professor Massimiliano Errico and Professor Martin Aage Barsøe Hedegaard for hosting me at the University of Southern Denmark, offering me the opportunity to work with highly advanced equipment while sharing their high-level knowledge. I would like to extend my thanks to the personnel I worked with during my time abroad, especially Ron Hajrizaj, with whom I had fruitful discussions.

I would like to thank Professor Pierluigi Caboni and Dott. Cristina Manis for their valuable contribution in extending my knowledge in the field of food chemistry and for their extremely valuable support during the experimental phases of the research reported in this dissertation.

I would like to thank Agris Sardegna for conducting the cheesemaking trials that made this work possible.

Special thanks to my family, Lello, Virginia, and Benedetta, who have always encouraged me to pursue my dreams and passions with humility and respect.

Thanks to Claudia, my wonderful nutritionist girlfriend who has been tolerating my Mind-bending mathematical journeys for five years.

Thanks to the persons I met at the University of Cagliari, including my colleagues, with a special mention to Nicola, Giacomo and Andrea (I am still waiting for you to play padel with me), my friends from the “Gruppo Mensa”, I Castellet y Ballarà, Daniele, Davide and those who every year join to our “pre-Christmas lunch”. You always kept my spirit high.

Thanks to Camilla. Thank you for listening to my countless PowerPoint presentations.

Thanks to Francesco Ortu and “Efisio Carta”, who gave me the best commentaries on Cagliari Calcio football matches ever.

Thanks to the Catholic International Group from St. Albani Church of Odense for including me when I was too shy to take a step forward. Keep spreading your joy and faith. I hope one day to be cooking *Pasta alla Campidanese* for you again, while boringly showing off my Sardinian pride. Mange Tak!

Thanks to Padre Emmanuel, a source of wisdom and a mentor during this experience.

Thanks to the relatives in my extended family that always cheered me on.

I apologize if somebody has not been mentioned in this section, but I hope you understand that a text of this thesis's length is not sufficient to list all the people and experiences that enriched me during these eight years at university.

Thanks to Paola, to whom I dedicate this thesis. A colleague and friend whose absence leaves a bitter void every day. May your spirit of joy remain an everlasting mark for all of us.