UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Ingegneria Industriale

SCUOLA DI DOTTORATO DI RICERCA IN INGEGNERIA INDUSTRIALE
INDIRIZZO: INGEGNERIA CHIMICA, DEI MATERIALI E DELLA PRODUZIONE
CICLO XXVI

LATENT VARIABLE MODELING TO ASSIST PRODUCT QUALITY CHARACTERIZATION IN THE FOOD AND PHARMACEUTICAL INDUSTRIES

Direttore della Scuola: Ch.mo Prof. Paolo Colombo
Coordinatore d'indirizzo: Ch.mo Prof. Enrico Savio
Supervisore: Ch.mo Prof. Massimiliano Barolo

Dottorando: Matteo Ottavian

Foreword

The realization of the work included in this Dissertation involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.
Most of the research activity that led to the results reported in this Dissertation has been carried out at CAPE-Lab, Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy), under the supervision of Prof. Massimiliano Barolo. Part of the work was carried out at Pfizer Worldwide R&D, Groton, CT (U.S.A) during a 6-month stay under the supervision of Dr. Salvador García-Muñoz, and part represents a collaboration with Dr. Luca Fasolato from the Department of Comparative Biomedicine and Food Science and Prof. Paolo Berzaghi of the Department of Animal Medicine, Production and Health, both of the University of Padova (Italy).
Financial support to this study has been provided by "Fondazione CARIPARO" (Project # PARO104725 - 2010) and by "Fondazione Ing. Aldo Gini", Padova (Italy).

All the material reported in this Dissertation is original, unless explicit references to studies carried out by other people are indicated. In the following, a list of the publications stemmed from this project is reported.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (published or in press)

Ottavian, M., M. Barolo and S. García-Muñoz. Maintenance of Machine Vision Systems. Part II: Addressing Camera Replacement. *Ind. Eng. Chem. Res*, in press. DOI: 10.1021/ie402910z.

Ottavian, M., M. Barolo and S. García-Muñoz. Multivariate Image and Texture Analysis to Investigate the Erosion Mechanism of Film-Coated Tablets: an Industrial Case Study. *J. Pharm. Innov.*, in press.

Ottavian, M., L. Fasolato, L. Serva, P. Facco and M. Barolo. Data Fusion for Food Authentication: Fresh/Frozen-Thawed Discrimination in West African Goatfish (*Pseudupeneus prayensis*) Fillets. *Food Bioprocess Tech.*, in press. DOI: 10.1007/s11947-013-1157-x.

Dalle Zotte, A., M. Ottavian, A. Concollato, L. Serva, R. Martelli and G. Parisi. Authentication of Raw and Cooked Freeze-Dried Rainbow Trout (*Onchorhyncus mykiss*) by Means of Near-Infrared Spectroscopy and Data Fusion. *Food Res. Int*., in press. DOI: 10.1016/jfoodres.2013.10.033.

Ottavian, M., L. Fasolato, P. Facco and M. Barolo (2013). Foodstuff Authentication From Spectral Data: Toward a Species-Independent Discrimination Between Fresh and Frozen-Thawed Fish Samples. *J. Food Eng.*, **119**, 765-775.

Ottavian, M., M. Barolo and S. García-Muñoz (2013). Maintenance of Machine Vision Systems. Part I: Addressing Changes in Lighting Conditions. *Ind. Eng. Chem. Res.*, **52**, 12309-12318.

Boschetti, L., M. Ottavian, P. Facco, M. Barolo, L. Serva, S. Balzan and E. Novelli (2013). A Correlative Study on Data From Pork Carcass and Processed Meat (Bauernspeck) for Automatic Estimation of Chemical Parameters by Means of Near-Infrared Spectroscopy. *Meat Sci.*, **95**, 621-628.

Ottavian, M., P. Facco, L. Fasolato and M. Barolo (2012). Multispectral Data Classification Using Similarity Factors. *Chemom. Intell. Lab. Syst.*, **118**, 13-23.

Ottavian, M., P. Facco, M. Barolo, P. Berzaghi, S. Segato, E. Novelli and S. Balzan (2012). Near-Infrared Spectroscopy to Assist Authentication and Labeling of Asiago d'Allevo Cheese. *J. Food Eng.*, **113**, 289-298.

Ottavian, M., P. Facco, L. Fasolato, E. Novelli, M. Mirisola, M. Perini and M. Barolo (2012). Use of Near-Infrared Spectroscopy for Fast Fraud Detection in Seafood: Application to the Authentication of Wild European Sea Bass (*Dicentrarchus labrax*). *J. Agric. Food Chem.*, **60**, 639-648.

CONTRIBUTIONS IN REFEREED CONFERENCE PROCEEDINGS (published or in press)

Ottavian, M., M. Barolo and S. García-Muñoz (2013). Multivariate Image and Texture Analysis for Film Coated Tablets Elegance Assessment. Edited preprints of *DYCOPS-10 - 10th IFAC International Symposium on Dynamics and Control of Process Systems* (M. Henson and G. Pannocchia, Eds.), Mumbai (India), 18-20 December, 331-336.

CONTRIBUTIONS IN UNREFEREED CONFERENCE PROCEEDINGS

García-Muñoz, S., M. Ottavian and M. Barolo (2013). Overall Elegance Assessment for a Solid Oral Dosage Form Using Multivariate Image Analysis and Multivariate Texture Analysis. *AIChE 2013 Annual Meeting*, November 3-8, San Francisco, CA (U.S.A.).

Facco, P., M. Ottavian, L. Fasolato and M. Barolo (2013). Toward Species-Independent Foodstuff Authentication From Spectral Data: the Case of Fresh and Frozen-Thawed Fish. *VIII Colloquium Chemometricum Mediterraneum*, June 30-July 4, Bevagna (PG, Italy).

Ottavian, M., M. Barolo and S. García-Muñoz (2012). Model-Based Overall Elegance Assessment for a Pharmaceutical Drug Product. *IFPAC 2012: advances in pharmaceutical innovation and manufacturing control*, October 7-10, Cortona (AR, Italy).

Ottavian, M., P. Facco, L. Fasolato, L. Serva, M. Mirisola and M. Barolo (2012). Autenticazione di Filetti di Triglia Atlantica (*Pseudopeneus prayensis*) Freschi Mediante uno Strumento NIR Portatile. *Convegno NIR Italia 2012: energia per il futuro*, September 26-28, Legnaro (PD, Italy).

Ottavian, M., P. Facco, M. Barolo and S. García-Muñoz (2012). Standardizzazione di Sistemi di Visione Artificiale per la Caratterizzazione di Prodotti Farmaceutici. *Convegno GRICU 2012. Ingegneria chimica: dalla nanoscala alla macroscala*, September 16-19, Montesilvano (PE, Italy).

Facco, P., M. Ottavian, L. Fasolato and M. Barolo (2013). Caratterizzazione della Qualità del Prodotto nell'Industria Alimentare Mediante Metodi Statistici Multivariati. *Convegno GRICU 2012. Ingegneria chimica: dalla nanoscala alla macroscala*, September 16-19, Montesilvano (PE, Italy).

Abstract

The pressure of the global competition, continuously asking for lower costs and improved productivity, is forcing companies to seek global supply chains to cut production costs down. As a result, it is becoming more and more difficult to accurately monitor each step of a production process and to protect products from economically motivated fraud, adulterations and counterfeiting. In such context, traditional methods for product quality characterization, such as lab assays, are expensive, destructive, time-consuming, and for these reasons they have become inadequate in several applications. On the other hand, other approaches, such as absorption spectroscopy and computer vision, have been gaining much attention in the last decade, successfully contributing to speed up and automate the quality assessment exercise. Statistical modeling tools, particularly latent variable models (LVMs), are usually employed to exploit the information embedded in the large amount of highly correlated data (spectra and images) that absorption spectroscopy and computer vision generate.

In the food and pharmaceutical sectors, product quality assessment still relies mainly on the judgment (of product color, odor, form, taste, etc.) of a panel of trained experts. Although the number of applications of LVMs as predictive tools for product quality monitoring is growing in these sectors, the use of LVMs for product quality assessment is usually *tailored* to each application, and *general* approaches to product quality assessment based on LVMs are lacking. The main objective of the research presented in this Dissertation is to overcome some of the limitations that hinder the diffusion of LVM tools in the food and pharmaceutical industrial practice. Three main strategies for product quality assessment are explored, namely the use of computer vision, the use of absorption spectroscopy, and the possibility of combining the information derived from different analytical instruments.

With respect to the use of **computer vision** systems, the problem of maintaining such systems is discussed. Computer vision systems are deemed to be quick, accurate, objective and able to return reproducible results. However, likewise all other measurement systems, they need to be maintained. Alterations or failures (e.g. of the illuminating system or of the camera sensors) can dramatically affect measurement reproducibility, leading to a wrong product quality characterization. The problem of how to detect and manage these alterations or failures is discussed through a pharmaceutical engineering case study. General strategies are proposed to adapt a quality assessment model, which has been calibrated under certain environmental conditions, to new conditions. Results show that long downtime periods, which may be necessary to recalibrate the quality assessment

model after a failure of the camera or of the lighting system, can be significantly reduced. Additionally, it is shown how image analysis can be effectively used not only to characterize the quality of a product, but also to improve the understanding on the production process (e.g., for troubleshooting or optimization purposes). In a specific pharmaceutical application, image analysis is used to investigate the causes leading to the erosion of tablets, allowing one to evaluate the effect of different physical phenomena occurring in the film-coating process. Additionally, the model relating the process conditions to the tablets quality is shown to be useful for process monitoring purposes.

With respect to the use of **absorption spectroscopy**, a novel methodology to preprocess and classify spectral data is proposed. Traditionally, LVMs are built after some preprocessing of the raw spectra, and the optimal preprocessing strategy is chosen trough a time consuming trial-and-error procedure. Results from three different food engineering case studies show that the proposed methodology performs similarly to other existing approaches, but it uses a sequence of totally automated preprocessing steps, with no need for trial-and-error searches.

Especially in the food industry, LVMs are usually tailored on the specific product being analyzed. For instance, for the detection of the fresh/frozen-thawed substitution fraud in fish fillets, a model is calibrated *for each* fish species possibly subject to substitution. This Dissertation considers a different approach: some strategies are proposed to design a multi-species, and possibly species-independent, classification model to detect this substitution fraud. The most promising strategy decomposes the information embedded in the spectral data using a single model, and it is shown to return the same overall accuracy of traditional approaches that employ one classification model for each species under investigation.

Finally, with respect to the use of **data fusion**, it is shown how to effectively combine the information derived from different analytical instruments (such as spectrometers, digital cameras, texture analyzers, etc.) to enhance product quality characterization. Results on two food engineering case studies show that fusing the available information, rather than using them separately, improves the ability of assessing product quality.

Riassunto

In un sistema economico globalizzato come quello in cui viviamo, garantire elevati standard in termini di qualità di prodotto costituisce per ogni azienda produttiva un fattore di successo. Monitorare in modo accurato la qualità del prodotto lungo tutti gli stadi della filiera produttiva, tuttavia, è divenuto progressivamente più complesso a causa della dimensione globale che quest'ultima ha assunto. É questo un effetto dei fenomeni di delocalizzazione della produzione, legati alla necessità delle aziende di non perdere quote di mercato a discapito di paesi emergenti caratterizzati da costi di produzione inferiori. In un tal sistema, aumenta anche il rischio di frodi, adulterazioni e contraffazioni del prodotto. Per certe categorie di prodotti, come quelli alimentari e farmaceutici, tali attività non solo danneggiano i consumatori dal punto di vista economico, ma possono anche causare seri problemi di salute.

Nonostante la grande importanza del monitorare la qualità di prodotto, a livello industriale si è ancora lontani da un sistema che permetta di caratterizzarla in modo rapido, economico, non invasivo (e quindi utilizzabile in tempo reale), riproducibile e multivariato (cioè in grado di quantificare contemporaneamente più parametri di qualità). Le tecniche che si sono dimostrate più promettenti in tal senso sono la spettroscopia d'assorbimento nella regione del visibile e del vicino infrarosso e l'analisi d'immagine. Per analizzare la moltitudine di dati (spettri e immagini) caratterizzati da forti correlazioni che queste generano, è necessario ricorrere a tecniche statistiche apposite, in particolare i modelli a variabili latenti (LVM, *latent variable models*). Tali tecniche, che sono pensate per trattare tali tipologie di dati, nascono dall'assunto che un sistema possa essere descritto mediante pochi fattori (detti anche variabili latenti) esprimibili come combinazione lineare delle variabili originali e interpretabili sulla base dei fenomeni chimico/fisici che interessano il sistema.

Il numero di applicazioni di LVM nel campo della caratterizzazione di prodotti alimentari e farmaceutici è cresciuto rapidamente negli ultimi anni. La maggior parte dei contributi pubblicati, tuttavia, offre soluzioni a specifici problemi anziché fornire approcci *generali*. L'obiettivo di questa Dissertazione è quello di superare alcune delle limitazioni esistenti al fine di favorire la diffusione di questi strumenti nella comune pratica industriale. La ricerca presentata si suddivide in tre macro aree di applicazione, che si differenziano a seconda della tecnica utilizzata per caratterizzare la qualità di prodotto, e cioè l'analisi d'immagine, la spettroscopia, e la fusione di dati (*data fusion*), cioè la combinazione delle informazioni provenienti da più strumenti analitici. Per ciascuna di queste aree, l'efficacia della

modellazione a variabili latenti viene dimostrata applicando i modelli in diversi casi studio di tipo industriale o di laboratorio.

Con riferimento all'**analisi d'immagine**, vengono proposte applicazioni nel campo farmaceutico. Nel Capitolo 3, l'analisi d'immagine viene utilizzata per il miglioramento della comprensione di un processo industriale di rivestimento di compresse. In tale processo la qualità finale del prodotto, che è legata all'omogeneità del rivestimento e al grado di erosione superficiale, viene tradizionalmente valutata da un *panel* di esperti, che necessariamente fornisce un giudizio soggettivo e poco riproducibile. Inizialmente, il Capitolo discute come, a partire da immagini del prodotto finito, sia possibile valutare in modo quantitativo e riproducibile i parametri di qualità. Le metriche sviluppate vengono quindi utilizzate per il *troubleshooting* del processo stesso, con il fine di indagare il meccanismo che porta all'erosione superficiale. A tal scopo, le metriche vengono correlate ai parametri di processo tramite un modello a variabili latenti, e i parametri del modello vengono utilizzati per definire le condizioni operative ottimali da utilizzare per garantire un prodotto in specifica.
Il Capitolo 4, usando ancora come pretesto un processo di rivestimento di compresse, discute in modo critico il problema della riproducibilità dei risultati ottenuti tramite analisi d'immagine. Tale riproducibilità, infatti, è garantita solamente se le condizioni sperimentali utilizzate per raccogliere le immagini destinate alla calibrazione del modello di stima della qualità vengono mantenute inalterate. Tali condizioni includono il sistema di illuminazione e la fotocamera stessa. Viene proposto innanzitutto un modello per il monitoraggio dello stato dell'apparato sperimentale, da utilizzare ogniqualvolta viene avviata una campagna di controllo qualità e basato semplicemente sull'acquisizione di un'immagine di standard colorati. In caso venga rilevato un cambiamento, viene proposta una strategia per adattare il modello di stima della qualità alle nuove condizioni. I risultati dimostrano l'efficacia della strategia proposta, che si basa su una tecnica già nota nel contesto della sincronizzazione vocale e dell'allineamento di traiettorie temporali in processi produttivi di tipo *batch*.

Con riferimento alla **spettroscopia d'assorbimento**, le applicazioni presentate riguardano prodotti alimentari, con particolare attenzione alle tecnologie per la rilevazione rapida di frodi di sostituzione (di un prodotto avente un certo valore di mercato con uno a valore di mercato inferiore). Nel Capitolo 5, viene presentata una nuova tecnica per la classificazione di dati spettrali, che ha l'obiettivo di razionalizzare il pretrattamento cui i dati stessi sono generalmente sottoposti. I risultati dimostrano come la tecnica proposta garantisca di ottenere la stessa accuratezza di altri metodi, senza tuttavia ricorrere a

procedure di tipo *trial-and-error*, onerose in termini computazionali, per la scelta del miglior pretrattamento.
Nel Capitolo 6, accanto a due applicazioni di autenticazione di prodotti alimentari (filetti pescati di branzino e formaggio Asiago d'allevo) tramite spettroscopia, viene presentata una tecnica multi specie per la rilevazione di una tipica frode del settore ittico, cioè la sostituzione di filetti freschi con filetti decongelati. Rispetto al tradizionale approccio di costruire un modello di rilevazione della frode per ciascuna specie, lavorare con un modello multi specie (e, magari, indipendentemente dalla specie) riduce notevolmente i tempi e i costi necessari nella fase di calibrazione. Delle tre strategie proposte, quella che fornisce risultati migliori lavora decomponendo l'informazione contenuta nei dati spettrali in due componenti, una legata alla specie e una legata allo stato fresco o decongelato. La tecnica, convalidata su un numero di spettri molto maggiore rispetto alla applicazioni riportate in letteratura, si è dimostrata efficace anche nell'autenticazione di campioni di specie non utilizzate nella fase di calibrazione.

Infine, con riferimento alla **fusione di dati**, il Capitolo 7 dimostra, attraverso due applicazioni in campo alimentare, come unire le informazioni ottenute da più strumenti analitici permetta di migliorare la caratterizzazione della qualità di un prodotto. La combinazione dei segnali a disposizione (detta *low level*, per distinguerla da altre tecniche di fusione di dati), opportunamente pesati, permette di ottenere risultati migliori rispetto all'utilizzo dei singoli segnali.

Table of contents

List of acronyms

AOAC	=	Association of Analytical Communities
DoE	=	design of experiments
DTW	=	dynamic time warping
EFSA	=	European Food Safety Authority
EMA	=	European Medicine Agency
FDA	=	Food and Drug Administration
ICH	=	International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use
*k*NN	=	*k* nearest neighbor
LDA	=	linear discriminant analysis
LV	=	latent variable
LVM	=	latent variable model
MB	=	multi-block
MB-PLS	=	multi-block projections to latent structures
MIA	=	multivariate image analysis
MWTA	=	multivariate wavelet texture analysis
NIPALS	=	nonlinear iterative partial least squares
NIR	=	near infrared
NIRS	=	near infrared spectroscopy
NMR	=	nuclear magnetic resonance
NIC	=	normal illuminating conditions
OPLS	=	orthogonal projections to latent structures
OPLS-DA	=	orthogonal projection to latent structures - discriminant analysis
PAT	=	process analytical technology
PC	=	principal component
PCA	=	principal component analysis
PLS	=	projections to latent structures
PLS-DA	=	projections to latent structures - discriminant analysis
PRESS	=	prediction error sum of squares

QbD = quality-by-design

QDA = quadratic discriminant analysis

RMSECV = root mean squared error of cross-validation

SNV = standard normal variate

SPE = squared prediction error

SPE_i = squared prediction error for sample *i*

SVD = singular value decomposition

VIP = variable importance in the projection

VIS = visible

Chapter 1

Motivation and state of the art

This Chapter provides an overview of the background and motivations of this Dissertation. First, the concept of product quality is discussed together with its relevance within the process industry. Then, some traditional approaches to product quality assessment are described. Finally, the role and the importance of latent variable modeling for product quality assessment is discussed. The objectives of the Dissertation and a roadmap to its reading conclude the Chapter.

1.1 Product quality and its role in a global economy

Defining product quality is not trivial. One of the first (and perhaps most comprehensive) definitions of quality was given by Garvin (1984), as summarized in Table 1.1. According to Garvin, five different approaches to the concept of product quality need to be taken into account, each of them representing the point of view of the discipline in which it has been rooted.

Table 1.1. *Product quality definitions according to Garvin (1984).*

Approach	Definition
Transcendent	Innate excellence
Product-based	Quantity of a desired attribute
User-based	Satisfaction of individual consumer preferences
Manufacturing-based	Conformance to requirements
Value-based	Affordable excellence

Table 1.1 clearly leads to conclude that a global definition of product quality does not exist; rather, different definitions of quality are appropriate under different circumstances (Garvin, 1984; Reeves and Bednar, 1994). Similarly, the more recent ISO 9000's quality definition, i.e. the *degree to which a set of inherent characteristics fulfils requirements* (International Standards Organization, 2005), needs to be conjugated to the specificity of the context being analyzed. Hence, the conclusion drawn by Garvin (1984), i.e. that product quality is a *multivariate* concept requiring several different product features to be taken into account simultaneously, is still appropriate nowadays.

Notwithstanding the difficulty of properly defining quality, it is widely recognized that in our global economy it plays a key role in determining the success (or failure) of a given product, as clearly reported for example in the European Union Green Paper: "*As globalisation spreads, products from emerging countries with low production costs are putting greater pressure on EU [...]. There is growing competition for both agricultural commodities and value-added products. Faced with these new commercial challenges, the [...] most potent weapon is 'quality'*" (Commission of the European Communities, 2008). As recently observed also by the US Food and Drug Administration (FDA, 2011b), the pressure of the global competition, continuously asking for lower costs and improved productivity, increased the flows of products around the world in the last two decades, as companies continue to move manufacturing activities to new and different locations, looking to global supply chains to reduce production costs. The market for contract manufacturing outsourcing in pharmaceutical production is a clear evidence of this trend: it grew linearly up to an estimated USD 46 billion in 2010, i.e. more than double its size in 2001. As an example, the cost of formulation of an active pharmaceutical ingredient can range from 15% to 40% less to produce it in India as compared to the US (FDA, 2011b).
Highly complex global supply chains represent nowadays the common practice for the majority of the existing products, even for those with a low added value. An example is given in Figure 1.1, which shows some of the steps fresh tuna goes before being sold as canned tuna in the US.

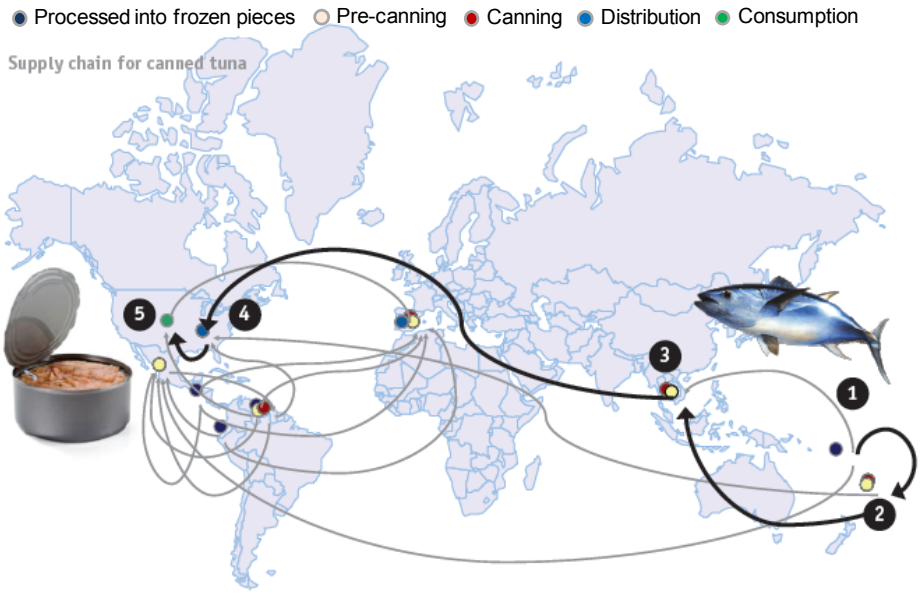


Figure 1.1. *Supply chain for canned tuna. The arrows indicate a possible path from the time the fish is caught in East Asia to the time the finished product reaches store shelves in the US. Adapted from FDA (2011).*

Spreading the stakeholders of product quality throughout the world by establishing global supply chains hinders the possibility of accurately monitoring *each* step of the production

process, due to the increased number of both processing steps and entities touching a given product. Additionally, in such complex networks it is becoming more and more difficult to prevent and detect the intentional efforts of adulteration, fraud and counterfeiting*. Such activities can not only negatively affect consumers from an economical point of view, but could also cause serious safety issues when the manipulation concerns some highly-sensitive product categories such as food and pharmaceuticals (Deisingh, 2005; Jacquet and Pauly, 2008). For example, the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) has estimated that 7% of all pharmaceuticals worldwide are counterfeited, the value of this business being worth more than USD 30 billion. Several examples of patients died after receiving adulterated medicines have also been reported (www.ifpma.org/global-health/counterfeits.html).

In such a framework, all players of a product supply chain are strongly motivated in assessing product quality, with each of them representing one (or more) of the points of view listed in Table 1.1. Manufacturers are interested in reducing scraps and defects to cut costs down; consumers look for products meeting their needs, i.e. with a specific series of characteristics; control agencies aim at ensuring product safety, preventing intentional adulteration, fraud and counterfeiting.

The problem of quantifying product quality in an unbiased and reproducible way represents the central idea developed throughout the Dissertation, with applications to two similar product categories such as foods and pharmaceuticals. Both the pharmaceutical and food industries have to face strict regulatory demands in terms of quality control, safety and traceability. Concepts like current good manufacturing practices (GMPs), quality risk assessment and management are enforced for food and pharmaceuticals producers (FDA, 2004a, 2004c and 2005; ICH 2006), with the aim of ensuring that the end products meet the desired quality standards. In addition, food and pharmaceuticals are both highly-sensitive products for which either poor quality or economically motivated fraud, adulteration and counterfeiting can easily translate into serious harms for public health.

1.2 Product quality in the pharmaceutical industry

A decade ago, the Quality by Design (QbD) initiative (ICH, 2004) revolutionized the approach of pharmaceutical practitioners towards product quality. Taking inspiration from the experiences of different industries (e.g. automotive, semiconductors, etc.), the new approach to pharmaceutical development and manufacturing had the purpose of favoring an efficient and flexible environment to produce reliably high quality products, without

* Fraud-related terms have been defined and harmonized for the global audience by one of the first working group in ISO Technical Committee 247, Fraud Countermeasures and Controls (International Standards Organization, 2010).

extensive regulatory oversight (Winkle, 2007). The QbD initiative encourages companies to the adoption of systematic science-based tools, rather than fixed traditional procedures (Tomba, 2013). The ultimate objective of this approach is to promote product and process understanding in pharmaceutical development, in order to increase manufacturing flexibility and process robustness. According to the QbD philosophy, the quality of a product cannot be assessed at the end of the product development activity or after manufacturing, but must be "built into" the product and ensured *since its design*, through a thorough mechanistic understanding of the relations between the quality of the product and the parameters that have an impact on it.

Building quality into products rather than testing it at the end of the manufacturing process implies that a comprehensive understanding of the characteristics of the drug (chemical, physical, pharmacological, pharmacokinetic, etc.), of the design and selection of the product components, and of the design of the manufacturing process and quality assurance is achieved. To reach this level of comprehension and develop well understood processes that are able to ensure consistently the predefined product quality, appropriate tools need to be employed, in order to measure and analyze effectively the relevant data. To this end, the FDA introduced the Process Analytical Technology (PAT) framework (FDA, 2004b). According to the agency definition, PAT is "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 product quality". It is important to note that the term *analytical* in PAT is viewed broadly to include chemical, physical, microbiological, mathematical and risk analysis conducted in an integrated manner (FDA, 2004b).

Through the PAT initiative, the FDA indicates the tools to be considered for an effective innovation in development, manufacturing and quality assurance. In particular, the objective of PAT is to provide support to clarify on a scientific basis typical issues that are likely to be encountered in development and manufacturing studies: for example, which the effects of product components on quality are, what sources of variability are more critical for the product, or how the process is able to manage variability.

In general, PAT includes all those tools that can provide an effective and efficient mean for acquiring valuable information to facilitate process understanding, continuous improvement through process and product monitoring and development of control and risk-mitigation strategies. In the PAT framework, these tools can be categorized according to the following (FDA, 2004b):

- multivariate tools for design, data acquisition and analysis;
- process analyzers;
- process control tools;
- continuous improvement and knowledge management tools.

The multivariate tools category includes all the multivariate mathematical approaches, such as statistical design of experiments, response surface methodologies, process simulation and pattern recognition tools, in conjunction with knowledge management systems, which allow to gain scientific understanding of the relevant multi-factorial relationships between formulation, process, and quality attributes. It includes also the means to evaluate the applicability of this knowledge to different scenarios. When used appropriately, these tools "enable the identification and evaluation of product and process variables that may be critical to product quality and performance". They may also "identify potential failure modes and mechanisms and quantify their effects on product quality" (FDA, 2004b).

Process analyzers include all the tools committed to the collection of data from the process. These measurements can be obtained at-line, i.e. by removing, isolating and analyzing the sample in proximity to the process stream; on-line, i.e. by diverting the sample from the manufacturing process and returning it to the process stream after the measurement; in-line, i.e. by keeping the sample inside the process stream, while the measurement can be made invasively or not. Process analyzers generate typically large volumes of data. For this reason, multivariate methodologies are indicated to extract critical process knowledge that can be related to product and process quality and used for process monitoring, control and end point determination. The design and installation of the analyzers on the process equipment is also identified as a critical step, as it must be ensured that the collected data are relevant and representative of process and product attributes (FDA, 2004b).

The process control tools include all the "process monitoring and control strategies intended to monitor the state of a process and actively manipulate it to maintain a desired state. Strategies should accommodate the attributes of input materials, the ability and reliability of process analyzers to measure quality attributes, and the achievement to process end points to ensure consistent quality of the output materials and the final product". In a PAT framework, the process should be continually monitored, evaluated and adjusted using in-process measurements, tests and controls in order to guarantee continuous quality assurance. This represents a way to demonstrate process validation.

Finally, the Agency encourages the adoption of PAT as continuous improvement tools, which enable a continuous learning through the data collected and analyzed over the lifecycle of the product. Approaches that support the acquisition of knowledge from these data would be valuable for manufacturing and facilitate the communication with the Agency on a scientific basis.

According to the FDA guidelines, process understanding, control strategies plus on-, in- or at-line measurement of quality attributes provides a scientific risk-based approach to justify how real time quality assurance is at least equivalent to, or better than, laboratory-based testing on collected samples (FDA, 2004b).

The QbD initiative is clearly intended to promote a novel approach to product quality within the manufacturing environment. The US FDA, however, is active also in protecting consumer from fraud, adulteration and counterfeiting. Drug counterfeiting is a vast global problem leading to deaths from untreated disease, reduced confidence in some vital drugs, large economic losses for the legitimate manufacturers and possible drug resistance (Deisingh, 2005; Feng *et al.*, 2013). According to the World Health Organization definition (http://www.who.int/medicines/services/counterfeit/overview/en/) and their chemical nature, the counterfeit drugs may include but not limited to (*i*) drugs containing no active ingredients; (*ii*) drugs containing the incorrect amount of active ingredients; (*iii*) drugs containing different but structurally related active ingredients which have lower prices than the right one; (*iv*) drugs containing different active ingredients totally unrelated to the correct one. In developed countries, counterfeit drugs harm primarily the trademark owner of expensive lifestyle drugs (Viagra® and Cialis®), while in developing countries drugs to treat life-threatening illnesses like malaria and tubercolosis are frequently counterfeited. However, while in the former case the percentage of counterfeited drugs is deemed to represent less than 1% of the market value, in the latter one it grows up to 25 or even to 50% (Mukhopadhyay, 2007). The Agency is currently promoting several strategies for securing the drug supply chain, in agreement with the US Federal Food, Drug and Cosmetic Act. These strategies are mainly based on improving the drug traceability throughout the market (FDA, 2010), also with the incorporation of physical-chemical identifiers in the case of solid oral dosage form drug products (FDA, 2011a). In addition, novel analytical techniques are being employed and developed for fast detecting of fraud, adulteration and counterfeiting, with the aim of replacing traditional laboratory analysis (see § 1.4).

The European Medicine Agency (EMA) runs roughly parallels to the US FDA and attempts to harmonize (but not replace) the work of existing national medicine regulatory bodies. EMA and FDA strictly collaborate to exchange confidential information as part of their regulatory and scientific processes, including information on advance drafts of legislation and regulatory guidance documents, as well as non-public information related to ensuring the quality, safety and efficacy of medicinal products for human and veterinary use.

1.3 Product quality in the food industry

US FDA guidance and regulations for food products are similar to pharmaceutical ones. As a consequence, QbD and PAT concepts are penetrating also in the food industry. In fact, the high degree of variation in biological processes makes it difficult to meet the strict regulatory demands in terms of quality control, safety and traceability using production

methods and technology that are primarily recipe-driven and based on post-production qualification and off-site laboratory analysis. Thanks to the introduction of real-time monitoring tools, food industries are gradually moving towards model predictive process control without the need of post-process testing (van den Berg *et al.*, 2013; Fissore *et al.*, 2013).

Likewise pharmaceutical products, one of the risks gaining attention from industry, governments, and standards-setting organizations (such as the US FDA, or the European Food Safety Authority, EFSA), is the economically motivated fraud by food producers, manufacturers, processors, distributors, or retailers. As reported by Moore *et al.* (2012), food fraud is a collective term encompassing the deliberate substitution, addition, tampering, or misinterpretation of food, food ingredients, or food packaging, or false or misleading statements made about a product for economic gain. A food fraud can damage consumers from an economical point of view. For example, in the case of seafood products, selling farmed fish as wild fish produces an economic loss for the customer, since wild fish as a higher market price. However, a food fraud can also entail safety issues. For example, in the case of seafood products, selling frozen-thawed fish as fresh fish may be harmful, since frozen-thawed fish has a lower market price and a shorter shelf life, i.e. it is more susceptible to microbial growth with respect to fresh meat (Pavlov, 2007). Moore *et al.* (2012) created recently a database of food fraud and adulterations by exploiting both literature and public media articles, and reported for olive oil, milk (and dairy products), honey and seafood the greatest number of records. Similar results were reported also by the Italian Ministry of Agricultural, Food and Forest Politics in its 2012 fraud report, which shows that on a total of 8200 food samples analyzed, nearly 12% was found to be irregular, leading to more than 20 million of kilograms of food products confiscated (for an equivalent value of EUR 0.5 billion; http://www.politicheagricole.it).

The US Food Safety and Modernization Act (FSMA) and the EU Commission Regulation No 1169, both enacted in 2011, recognized the importance of labeling as one of the most important tools to be used for fraud prevention. The EU Regulation lists the mandatory information for correct product labeling, which includes the ingredient list, the use by date, any special storage conditions, the place of the provenance, and the name and address of the food business operator responsible for all the information. Clearly, part of the Agencies activities consists of verifying whether a product is mislabeled or not. To this purpose, researchers are constantly seeking for fast, non-invasive and reliable techniques to be used for fraud screening. Some of these techniques will be discussed in Chapters 5-7 of this Dissertation.

1.4 Product quality assessment

There are no instruments capable of capturing alone the complex and multivariate nature of product quality. Consider for example a seafood product such as the European sea bass (Fasolato *et al*., 2010), which is traditionally one of the preferred finfish species by the Italian consumers (Poli *et al*., 2001). A summary of its main quality parameters is given in Table 1.2, together with the indication of the analytical technique/instrument used for quality assessment.

Table 1.2. *Product quality attributes of European sea bass and analytical technique/instrument for their assessment. The list is not exhaustive*.

Parameter	Analysis
Fat content	Method 920.39 (AOAC, 2002) as in Folch *et al*. (1957)
Moisture content	Method 934.01 (AOAC, 2002), i.e. gravimetrically oven-drying at 103°C
Protein content	Method 992.15 (AOAC, 2002), i.e. Kjeldahl method
Ash content	Method 942.05 (AOAC, 2002), i.e. gravimetrically after incineration at 550°C
Water activity	Hygrometer
Fatty acids profile	Gas chromatography of anhydrous fat transesterified according to Christie (1982)
Color	Colormeter
Mechanical properties	Texture analyzer (compression/shear stress)

Some of the attributes listed in Table 1.2 are related to nutritional aspects (e.g., fatty acids profile), some to safety issues (e.g., water activity), and some to the consumers' acceptance (e.g., color). As implicitly suggested in Table 1.2, measuring *all* these attributes is an expensive and time consuming activity, which requires several dedicated instruments and trained personnel. Some measurements can take many hours to be completed, preventing from in-/on-line use in high-throughput production chains. In addition, most of the techniques are destructive, and some of them require reagents, which entails environmental issues. Furthermore, it should be taken into account that the list proposed in Table 1.2 is not exhaustive. Information about freshness, geographical origin and status (i.e., whether wild or farmed), which are mandatory for correct product labeling (see § 1.3), are also highly appreciated by consumers. However, to date there are no accepted methods for the unequivocal determination of these quality attributes (Martinez *et al*., 2003; Jacquet and Pauly, 2008).

A similar situation exists also for many other food and pharmaceutical products. The quality assessment procedures in the food and pharmaceutical industries rely often on inspections carried out by a panel of trained experts that is intended to reproduce the consumers' perception (of odor, taste, color, etc.). Such evaluations are necessarily subject to human judgment, and hence barely reproducible. For instance, visual analysis can be used to assess tablet surface erosion (which is a key quality parameter for controlled-

release products as it might affect the absorption of the active pharmaceutical ingredient), or the color of fresh and processed food (fruits, beverages, fish, meat; Pathare *et al.*, 2013), or the odor and taste of both food (Lawless and Heymann, 2010) and pharmaceuticals (Baldwin *et al.*, 2011).

Despite the importance of monitoring product quality and the progresses made in the last decade as a result of the QbD initiative (Matero *et al.*, 2013; Tomba *et al.*, 2013; van den Berg *et al.*, 2013), the industrial practice is still far from employing measurement systems that are simultaneously cost-effective, fast, non-destructive (hence suitable for in-/on-line applications), objective and multivariate.

Several techniques have been proposed in the last decade to help assessing product quality, and two of them (absorption spectroscopy and computer vision) have been demonstrated to be particularly attractive since they encompass the majority of the abovementioned requirements. Basic information on these techniques is provided in the following.

1.4.1 Absorption spectroscopy

Absorption spectroscopy is the science that studies how light is absorbed when it interacts with matter. Given the electromagnetic spectrum, absorption spectroscopy techniques consider light sources covering the whole range between 1 nm and 10^9 nm (Figure 1.2), i.e. X-ray, visible and ultraviolet (UV-VIS), infrared (IR), which is further divided into near- (NIR), mid- (MIR) and far-infrared (FIR), and radio wave. Among them, NIR is by far the technique with more applications, particularly in the food and pharmaceutical industries: for this reason its use will be explored considerably throughout this Dissertation.

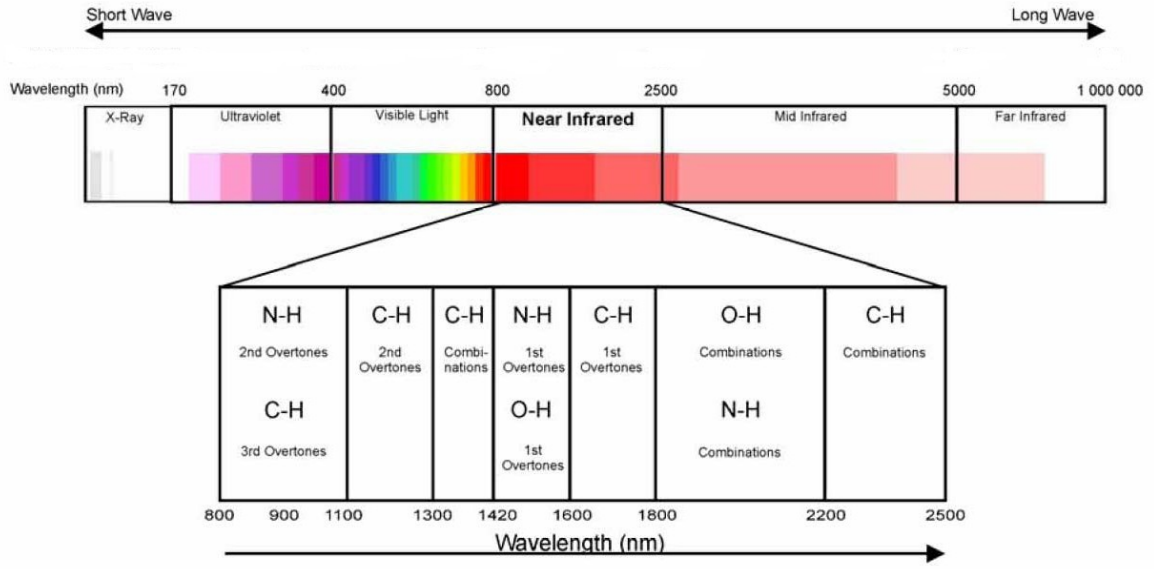
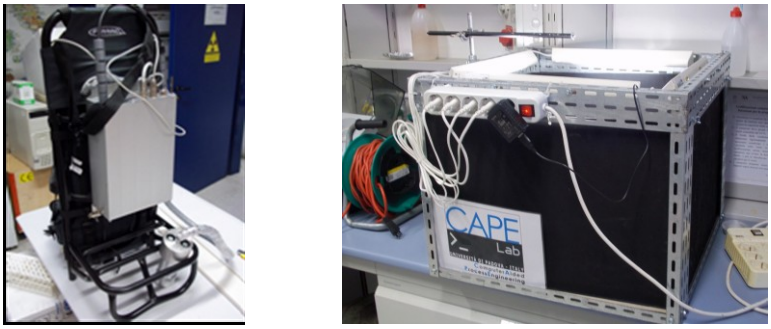


Figure 1.2. *Electromagnetic spectrum: light sources in absorption spectroscopy. For NIR radiation, the main vibrational and rotational transitions are indicated.*

The way the radiation interacts with matter is directly dependant on the energy (i.e., the wavelength) of the radiation. The higher energy UV-VIS wavelengths affect the energy levels of the outer electrons. IR radiation is absorbed by matter resulting in rotation and/or vibration of molecules. Radio waves are used in nuclear magnetic resonance and affect the spin of nuclei in a magnetic field. When the radiation is absorbed, it causes a transfer of energy from a lower energy state (the ground state) to a higher energy one (the excited state), with the amount of energy absorbed being exactly equal to the energy difference between the two states. Molecules can have many possible states and can absorb many different wavelengths, since their energy states are made up of three components, i.e. electronic, vibrational, and rotational. The electronic component is characterized by the energy states of bonding electrons (outer shell electrons). Vibrational states are associated with interatomic vibrations present in molecules. A molecule generally has many more vibrational levels than it does electronic levels and the energy difference between these states is generally much smaller than the differences between electronic states. There are also a number of rotational states for each of the vibrational states and these also have lower energies of transition between states.

By employing a low energy radiation, NIR spectroscopy (NIRS) explores the vibrational and rotational transitions of the main functional groups (-OH, -CH, $-CH_2$, $-CH_3$, -C=O, etc.) of the molecules constituting the product of interest. Clearly, the wavelengths absorbed by a molecule are specific of *that* molecule, and hence the spectrum of a product, which reports the absorbance of its molecules at different wavelengths, can be used to investigate qualitatively and quantitatively its chemical structure and composition. Spectroscopic analysis are carried out using spectrophotometers. Essential elements of a spectrophotometer are the light source, the monochromator containing a diffraction grating to produce the analytical spectrum, and the light detector. The instrument can be installed in a laboratory, in a production line or it can be made portable (Figure 1.3a). More details on NIR spectroscopy can be found in dedicated literature (Chalmers and Griffiths, 2002; Burns and Ciurczak, 2008).



(a) *(b)*

Figure 1.3. *(a) Portable NIR instrument and (b) in-house developed photographic box for automated inspection of product visual quality.*

1.4.2 Computer vision

Computer vision is the construction of explicit and meaningful descriptions of physical objects from images (Ballard and Brown, 1982). The term, which is synonymous with machine vision, embodies several processes. Images are acquired with a physical image sensor and dedicated computing hardware and software are used to analyze them with the aim of performing a predefined visual task. The technology aims to duplicate the effect of human vision by electronically perceiving and understanding an image (Sonka *et al*., 1999; Brosnan and Sun, 2004). Following its origin in the 1960s, computer vision has experienced a tremendous growth with applications expanding in very diverse fields.

Key elements of a computer vision apparatus are the illuminating system and the camera. As with the human eye, vision systems are strongly affected by the level and quality of illumination: the adjustment of the lighting can radically change the appearance of an object, with the features of interest sharpened or blurred. Details of various aspects of illumination including location, lamp type and color quality can be found elsewhere (Bachelor, 1985; Jahr, 2007; Valous *et al*., 2009). The in-house developed photographic box used to collect the images of Chapter 7 is shown in Figure 1.3b. The 60×60×60 cm^3 box is screened from the environmental light with opaque walls and equipped with four fluorescent tubes (Lumilux XT, Osram GmbH, Munich, Germany) with the aim of controlling the illumination.

The core of a computer vision system is represented by image processing and analysis. Image processing involves a series of image operations that enhance the quality of an image in order to remove defects. Image analysis is the process of distinguishing the objects (regions of interest) from the background and producing quantitative information, which is used in subsequent control systems for decision making. The information of interest is typically related to the *color* of the product being imaged or to its surface characteristics (coarseness, presence of defects, erosion, etc., i.e. *surface texture*). A thorough description of the techniques used for the extraction of this information is provided in Chapter 2.

1.5 Latent variable modeling and product quality

Latent variable models (LVMs) are statistical model purposely designed to handle large amount of (highly) correlated data. The idea behind LVM is that a system can be described by using few underlying factors (called *latent variables*, LVs), which can be expressed as linear combinations of the original variables and can be interpreted based on the knowledge of the physical and chemical phenomena involved. Figure 1.3 reports the geometrical interpretation of an LVM on a dataset of 11 samples characterized by 3 measured variables x_n (n = 1, 2, 3). As can be seen, the LVM transforms the original three-

dimensional variable space into a two-dimensional space (the *latent space*) defined by the orthogonal directions LV1 and LV2, representing the directions of maximum variability of the data. The projections (called *scores*) of the original variables onto the latent space become the new variables defining the state of the system.

LVMs are useful not only for modeling single spaces as depicted in Figure 1.3, but also for relating data from different datasets (Burnham *et al*., 1996). These models are commonly associated to analytical instruments to relate highly correlated input variables (e.g., spectroscopic variables) to response variables such as product quality. Generally, their use is suggested when the number of measured variables is large compared to the number of samples, since traditional least-squares approaches can lead to ill-conditioned problems (Burnham *et al*., 1996).

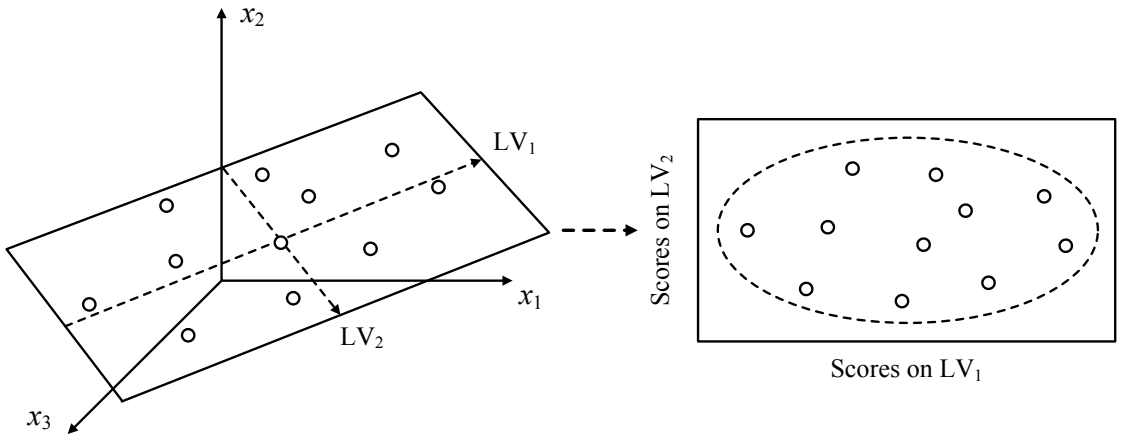


Figure 1.3. *Geometrical interpretation of an LVM model (from Tomba et al., 2013).*

Spectroscopy and machine vision, which both generate highly correlated data, are usually coupled with LVM tools such as principal component analysis (PCA; Jackson, 1991) or projection to latent structures (PLS; Geladi and Kowalski, 1986) in order to leverage the information embedded in spectra and images (respectively) and relate them to the quality features of interest. LVMs have found wide application both in the pharmaceutical and in the food industries as predictive tools, as demonstrated by recently published reviews (Karoui and De Baerdemaeker, 2007; Huang *et al.*, 2008; Rajalahti and Kvalheim, 2011; Pomerantsev and Rodionova, 2012; Matero *et al*., 2013).

1.5.1 Latent variable models in the pharmaceutical industry

Within the pharmaceutical industry, the use of latent variable models coupled to UV-VIS, NIR, Raman, acoustic emission, X-ray, nuclear magnetic resonance, chemical and digital imaging analysis has become common practice under the PAT framework. Though applications can be found for several dosage form (liquids, Kim *et al*., 2007; pellets, Mantanus *et al.*, 2010; syrups, Ziémons *et al*., 2010), the majority of them is concentrated on the solid one. In fact, the preferred way of administering active pharmaceutical

ingredients (APIs) via the oral route is in solid dosage forms, such as tablets (Matero *et al.*, 2013). PAT tools have been extensively used during manufacturing of solid oral dosage forms, with applications of real-time analyzers (especially NIR ones) in several key steps such as powder mixing, granulation, compaction and testing of the final preparation. Application range from the design of predictive model to the design of models for monitoring a unit operation. In particular:

- powder mixing is of primary importance in pharmaceutical manufacturing because formulations seldom consist of one ingredient only. There are several particle properties responsible for powder organization during the mixing process, and the most important ones (particle size, shape and density), which change during the process, can be monitored using NIRS, since it is sensitive to these changes. The traditional approach for monitoring the API and excipient(s) distribution during powder mixing consists in an off-line HPLC analysis of a sample taken (invasively) with a thief and dissolved in a buffer. With respect to this approach, LVM applications (using either PCA, PLS, or other models) based on NIR or Raman spectra or on hyperspectral images allow assessing in real-time blend uniformity (Wargo and Drennen, 1996; Ma and Anderson, 2008), mixture homogeneity (El-Hagrasy et al., 2006; Amigo et al., 2008) and mixing end point (Sekulic et al., 1998; El-Hagrasy et al., 2001; Shi et al., 2008; De Beer et al., 2008) with no sample representativeness problems due to segregation phenomena (El-Hagrasy et al., 2006);
- granulation, i.e. powder agglomeration, is carried out to ensure powder processability. Currently, the quality attributes of the granules are tested post-process by Karl Fischer tritation or weight loss on drying (moisture) and by sieving or laser diffractometry (particle size), while the process is controlled based on the univariate detection of individual process variables such as processing time, temperature, impeller tip speed and power and binder liquid consumption (Matero *et al*., 2013). Skibsted *et al*. (2007) proposed a monitoring model based on NIR spectra using the multivariate control charts described in Chapter 2 (Section § 2.1.3), showing its effectiveness in detecting bad batches with too many fine or coarse particles. Particle size estimation using either NIR or images was proposed by Rantanen *et al*. (2005), Alcala *et al.* (2010) and Laitinen *et al*. (2003), while Rantanen *et al*. (2001) used a four-wavelengths near-infrared sensor for in-line moisture evaluation;
- tablet formation through powder compaction requires the quality of the end product to be evaluated in terms of uniformity of content and mass, mechanical properties (hardness) and API dissolution profile. While traditional testing methods are time demanding, destructive and inefficient (Matero *et al*., 2013), several applications have been proposed using NIR, Raman and chemical imaging analysis (Ravn *et al*., 2008; Gordon and McGoverin, 2011; Gray *et al.*, 2009). Additionally, García-Muñoz and Carmody (2010)

and García-Muñoz and Gierer (2010) showed how to employ easy-to-operate digital cameras to estimate some tablets quality attributes such as texture and color uniformity.

Despite the prominent use of LVMs as PAT tools, García-Muñoz and Oksanen (2010) and Tomba *et al.* (2013) stressed on the much more valuable role these models can play in the practical implementation of the QbD paradigm. The effectiveness of LVMs in such activities as process control (Flores-Cerrillo and MacGregor, 2003, 2004 and 2005; Yu and Flores-Cerrillo, 2013), process design (Jaeckle and MacGregor, 1998, 2000a and 2000b), product design (Lakshminarayanan *et al.*, 2000) and optimization (Yacoub and MacGregor, 2004) has been proved in several industrial sectors. Pharmaceutical applications in these areas, which are outside the aim of this Dissertation, have been recently revised by Tomba *et al.* (2013).

LVMs as PAT tools are largely employed in the detection of counterfeited drugs (Deisingh, 2005; Feng *et al*., 2013). Simple PCA models built on NIR or Raman spectra have been shown to easily detect counterfeited Combiron® (ferrous sulfate), Aldomet® (methyldopa), Floxacin® (norfloxacin), Tylenol® (acetaminophen), and Viagra® (Deisingh, 2005; Moffat *et al.*, 2010).

1.5.2 Latent variable models in the food industry

PAT tools, especially NIR spectroscopy, have been around in the food industry for more than 40 years. For example, Vornheder and Brabbs (1970), for instance, used NIRS for moisture determination in potato chips, honey and soybean oils; Begley *et al.* (1984) for the detection of sodium chloride in meat samples; and Kennedy *et al.* (1985) for the analysis of milk and dairy products. The number of applications of PCA, PLS or of other LVM-based techniques combined with different analytical instruments (UV-VIS, NIR, Raman photometers, GC-MS analyzers, digital cameras, etc.) have been steadily increasing in the last decade. For instance, PCA and/or PLS combined to NIR spectra have been used:

- in the meat industry, for monitoring on-line fat, water, and protein content (Tøgersen *et al.*, 2003) and proximate composition (Hildrum *et al.*, 2004) in semi-frozen raw beef and pork and ground meat, and for the on-line grading of poultry carcasses (Chen *et al.*, 2003);
- in the fruit and vegetables industry, for monitoring on-line the sugar content (He *et al*., 2001; Miller and Zude, 2002) in apples and oranges or to replace human operators in sorting apples (McGlone *et al*., 2005) and stone fruits (Golic and Walsh, 2006);
- in the grain and grain products industry, for monitoring on-line price-related features such as protein and starch content and grain hardness (Berardo *et al.*, 2005) and for rice sorting (Kawamura *et al.*, 2003a);
- in the dairy industry, for predicting the three major milk constituents (fat, protein and lactose), somatic cell count and urea nitrogen (Kawamura *et al.*, 2003b) and for process

control purposes during the production of traditional feta cheese (Adamopoulos *et al.*, 2001) and yogurt (Navrátil *et al.*, 2004);

- in the beverages industry, for monitoring on-line the alcohol content during wine fermentation (Zeaiter *et al.*, 2006; Cozzolino *et al*., 2006) and for soluble solids and total solids/total moisture in processing apple, grape, pear juices (Singh *et al.*, 1996).

LVMs built on images collected from digital cameras are also common in the food industry. For instance, Yu and MacGregor (2003) showed how PCA-based image analysis can be effectively used to predict the coating content and distribution of snack food, and how to use the predictions in a closed loop control scheme (Yu *et al.*, 2003). Image analysis has been proposed in several other applications. For example, to assess on-line the brightness and surface kinetics during coffee roasting (Hernández *et al.*; 2008), to determine hazelnut peeling (Pallottino *et al.*; 2010), to classify or grade different food products such as bananas (Mendoza and Aguilera, 2004), ham (Valous *et al*., 2009), salmon (Misimi *et al.*, 2008), to monitor milling quality during rice production (Yadav and Jindal, 2001), etc.

PAT tools are also widely employed as anti-fraud systems. Several reviews are available on this topic (Rodriguez-Saona and Allendorf, 2011; Elmasry *et al.*, 2012; Liu *et al.*, 2013), and additional references are provided in Chapters 5-7.

1.6 Objectives of the research

Despite the continuously growing number of applications of LVMs as PAT tools for product quality monitoring and control both in the pharmaceutical industries and in the food industries, in most published contributions *tailored* solutions to specific problems have been provided. The main objective of the research presented in this Dissertation is to overcome some of the existing limitations that hinder the diffusion of PAT tools in common industrial practice. Namely, applications are presented in the following areas.

- **Machine vision systems for product quality assessment**. Machine vision systems have been gaining much attention in the last decades for the characterization of several products whose quality can be related to some visual features. With respect to the visual inspection traditionally carried out by a panel of trained experts, machine vision systems are deemed to be quick, accurate, objective and able to return reproducible results. However, likewise all other measurement systems, they need to be maintained. Alterations or failures (e.g. of the illuminating system or of the camera sensors) can dramatically affect reproducibility, leading to a wrong product quality characterization. The problem of how to detect and manage these alterations or failures, which is usually not covered in the open literature, is discussed in details in the Dissertation, and strategies are provided to avoid (or at least reduce) downtime periods in case of changes

of the lighting conditions and/or camera replacement (due to either failure or multiple installations). Additionally, the Dissertation shows how image analysis can be effectively used not only to characterize the quality of a product, but also to gain process understanding of the production process (for troubleshooting or optimization purposes).

- **Absorption spectroscopy for product quality characterization**. Traditionally, LVMs are built on spectral data after some preprocessing on the raw spectra. While the preprocessing is intended to reduce the variability in the regressor matrix **X** that is not related to the response matrix **Y**, the optimal preprocessing strategy is chosen through a time consuming trial-and-error procedure. In addition, models are usually tailored on the specific product being analyzed requiring, at least in the food industry, an unfeasibly large number of calibration samples. To detect the fresh/frozen-thawed substitution in fish fillets using UV-VIS/NIR spectra, for instance, a model is calibrated *for each* fish species. Both the abovementioned issues are discussed in the Dissertation. A novel approach is proposed to automate the preprocessing of the raw data, at least for classification applications, and several strategies are evaluated for the first time for calibrating multi-products models.
- **Sensor and data fusion**. Data fusion approaches are oriented toward the simultaneous use of the information arising from data of different natures (i.e., data derived from different analytical instruments). The joint (i.e., multivariate) evaluation of the analytical results allows to better describe the investigated system and to answer different questions pertaining to which information it is expected (it is possible) to gain from the different sets/blocks of data (Van Mechelen and Smilde, 2010). Data fusion techniques can be classified into three main groups: *i*) low-level data fusion consists of the simple concatenation of the data of different nature; *ii*) mid-level data fusion is based on features extraction or variable selection prior to multivariate analysis; *iii*) high-level or hierarchical data fusion is based on the concatenation of the scores, extracted by means of multivariate projection techniques such as PCA, PLS, etc. In this Dissertation, examples of low-level data fusion are provided, showing how to effectively combine the available information to enhance the characterization of product quality.

The effectiveness of the strategies proposed in this Dissertation is demonstrated by applying each of them to experimental case studies. The next section presents a roadmap to the Dissertation.

1.7 Dissertation roadmap

After a thorough description of the LVM techniques and the of the statistical background in Chapter 2, the following Chapters of this Dissertation, whose roadmap is sketched in

Figure 1.4, can be divided according to the three research areas identified in the previous section.

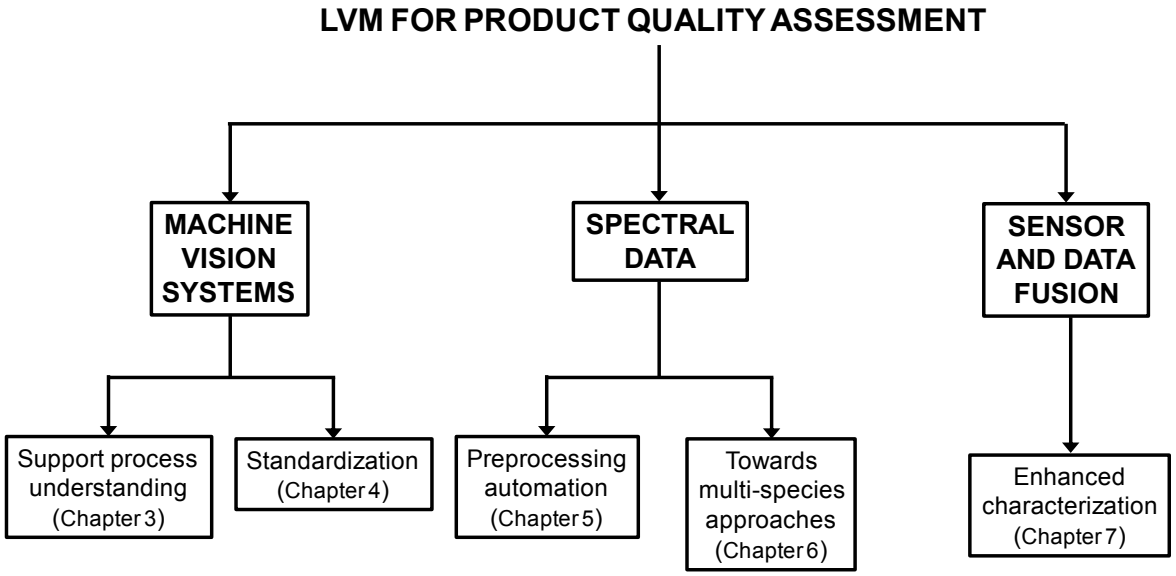


Figure 1.4. *Dissertation roadmap.*

With respect to machine vision systems, Chapter 3 shows how image analysis can be employed not only in replacement of a panel of trained experts (to improve the characterization of the visual quality of the product), but also to troubleshoot the production process. The design space of an industrial tablet film-coating process is determined in the reduced space of a PLS model by regressing the process operating conditions against the quality metrics derived from the digital images of the tablets, and the model itself is used to investigate the root causes leading to surface erosion. In Chapter 4, the problem of how to standardize machine vision systems is discussed with respect to *i*) changes in lighting conditions and *ii*) camera replacement. Namely, given a machine vision system, novel approaches are presented to detect and manage changes in its conditions. The proposed approaches, which are based on adapting the PCA scores of the images to the new conditions, are shown to be effective also when transferring the quality assessment model across different cameras.

With respect to modeling spectral data for product quality assessment, Chapter 5 proposes a novel strategy (for classification problems) to handle automatically the data preprocessing step. The strategy, which is based on similarity factors defined from PCA models, is intended to avoid the trial-and-error procedure that is usually employed to accomplish the preprocessing operations, and aims at classifying the samples (i.e., spectra) based on the similarity among their shape. In Chapter 6, the use of spectral data for fast fraud detection is discussed. In particular, a novel approach based on the orthogonal PLS algorithm developed by Trygg and Wold (2002) is presented to allow calibrating species-independent anti-fraud systems. Stemming from the problem of detecting the fresh/frozen-

thawed substitution fraud in fish fillets, a multi-species approach is proposed and evaluated over more than a thousand fish samples and using two NIR instruments.

Eventually, Chapter 7 focuses on the problem of fusing the information derived from different analytical instruments (digital cameras, spectrometers, texture analyzers, etc.) to authenticate a product. The effectiveness of concatenating into a multiblock (Westerhuis *et al.*, 1998) framework the available information is shown through two different case studies, considering both linear (PLS) and non-linear (*k*NN) models.

A conclusive Section summarizes the major findings of the Dissertation and outlines future perspectives.

Chapter 2

Latent variable modeling background*

This Chapter provides a general overview of the statistical and mathematical techniques applied in this Dissertation. First, a background on latent variable models (LVMs) is presented. Techniques like PCA and PLS are discussed from both an algorithmic and a practical point of view. Furthermore, details on the use of LVMs for image analysis applications are provided.

2.1 Latent variable modeling approaches

Latent variable models (LVMs) are statistical models that are intended to analyze large amounts of (usually correlated) data. The idea behind LVM is that a system can be described by using few underlying factors (called *latent variables*, LVs), which can be expressed as linear combinations of the measurements taken on the system and can be interpreted based on the knowledge of the physical and chemical phenomena involved. LVMs find the driving forces acting on the system and responsible for the variability that can be observed in the data, under the assumption that the essential information does not lies in any individual variable, but in how the variables change with respect to one another, i.e. in how they co-vary. Hence, the objective of an LVM analysis is twofold: *i*) data compression and *ii*) data interpretation.

The available data can be organized into either a matrix $\mathbf{X}$ [$I \times N$], in which the N variables have been observed per I samples (or observations). Alternatively, the observations can be organized in a matrix $\mathbf{X}$ of regressors and a matrix $\mathbf{Y}$ [$I \times M$] of response variables. In the first case, the objective of an LVM analysis is to explain the correlation structure of the N variables, in order to understand the relationships among them. Principal component analysis (PCA; Jackson, 1991) is one of the most useful techniques to this purpose. In the second case, the objective of an LVM analysis is to explain the cross-correlation structure of the variables in $\mathbf{X}$ and in $\mathbf{Y}$, in order to study and quantify the relationships between regressors and response variables. Projection to latent structures (PLS, also called partial

* Ottavian, M., E. Tomba and M. Barolo. Advanced Process Decision Making Using Latent Variable Methods. In: *Handbook in Process Simulation and Data Modeling in Solid Oral Drug Development and Manufacture* (S. García-Muñoz and M. Ierapetritou, Eds.). Submitted.

least-squares regression; Höskuldsson, 1988) can be used for this purpose. Basic theory of PCA and PLS is reported in the following.

2.1.1 Principal component analysis (PCA)

Principal component analysis (PCA; Jackson, 1991) is a multivariate statistical method that summarizes the information embedded in a dataset $\mathbf{X}$ of N correlated variables, by using a linear transformation to project the data onto a new coordinate system of latent orthogonal variables, which optimally capture the variability of the data and the correlation among the original variables. Each of these new coordinates identifies a latent direction in the original data and is called principal component (PC).
The search for the directions of the new coordinate system can be formulated as an optimization problem (Burnham *et al*., 1996). For one PC:

$$\begin{aligned} &\max_{\mathbf{p}} \left(\mathbf{p}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{p} \right) \\ &\text{subject to} \quad \mathbf{p}^{\mathrm{T}} \mathbf{p} = 1 \end{aligned} \quad , \tag{2.1}$$

where the superscript $^{\mathrm{T}}$ indicates the transpose of a matrix, and $\mathbf{p}$ is the $[N \times 1]$ vector of the coefficients of the linear combination (called *loadings*). The loadings of a given PC represent the director cosines of the PC. Given the vector $\mathbf{t}$ $[I \times 1]$ of the projections (called *scores*) of the original data onto the PC direction:

$$\mathbf{t} = \mathbf{X}\mathbf{p} \quad , \tag{2.2}$$

the problem (2.1) can be reformulated as in (2.3), representing the maximization of the score vector length:

$$\begin{aligned} &\max_{\mathbf{p}} \left(\mathbf{t}^{\mathrm{T}} \mathbf{t} \right) \\ &s.t. \quad \mathbf{t} = \mathbf{X}\mathbf{p} \\ &\qquad \mathbf{p}^{\mathrm{T}} \mathbf{p} = 1 \end{aligned} \quad . \tag{2.3}$$

The analytical solution of problem (2.3) is given by the eigenvector decomposition of the covariance (or correlation, according to the preprocessing of the data, see § 2.1.1) matrix of the original variables (Burnham *et al*., 1996; López-Negrete de la Fuente *et al*., 2010):

$$\mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{p} = \lambda \mathbf{p} \quad . \tag{2.4}$$

Vector $\mathbf{p}$ is the eigenvector of the covariance matrix of $\mathbf{X}$ ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) corresponding to the eigenvalue λ, and λ is a measure of the variance explained by the given PC.

The eigenvector problem (2.4) can be used to determine the N loadings $\mathbf{p}_n$ of the PCA model, which are orthonormal (i.e., $\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_r = 0$ if $n \neq r$, $\mathbf{p}_n^{\mathrm{T}}\mathbf{p}_r = 1$ if $n = r$, with n, r = 1,..., N). The scores $\mathbf{t}_n$, instead, are orthogonal (i.e., $\mathbf{t}_n^{\mathrm{T}}\mathbf{t}_r = 0$ if $n \neq r$, with n, r = 1,..., N).
The dataset $\mathbf{X}$ can be represented as the sum of the N scores-loadings vectors outer products:

$$\mathbf{X} = \sum_{n=1}^{N} \mathbf{t}_n \mathbf{p}_n^{\mathrm{T}} . \tag{2.5}$$

As a result of the eigenvector problem (2.4), the PCs are ordered according to the variance of the original dataset $\mathbf{X}$ that they capture. Usually, $A << N$ principal components are sufficient to adequately describe $\mathbf{X}$, because if two or more original variables are correlated, they identify a common direction of variability that can be described by a single PC. Hence, assuming that only the first A PCs are retained, and defining the score matrix $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$ and the loading matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_A]$, Eq. (2.5) can be rewritten as:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \sum_{a=A+1}^{N} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \quad , \tag{2.6}$$

where $\mathbf{E}$ is the $[I \times N]$ matrix of the residual generated by the $(N - A)$ discarded PCs of the PCA model when $\mathbf{X}$ is reconstructed (i.e., approximated) by using only the first A PCs (i.e., $\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$).
For the computation of the model scores and loadings, the singular value decomposition (SVD; Meyer, 2000) of the covariance matrix of $\mathbf{X}$ ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) or the nonlinear iterative partial least-squares algorithm (NIPALS; Wold, 1966) can be used. The latter should be preferred as it tolerates a certain amount of missing data into the dataset $\mathbf{X}$ and does not require the calculation of all the N PCs. Details of the algorithms are given in Appendix A.
Figure 2.1 reports the geometrical interpretation of the PCA model parameters in a simplified case, i.e. a [7×2] dataset $\mathbf{X}$. As it can be seen, the data follow a trend in the (bidimensional) space of the original variables. If a PCA model is applied, the direction of maximum variability of the data is identified by PC1. The model loadings (p_1, p_2) represent the director cosines of PC1. The scores represent the coordinates of the data samples of matrix $\mathbf{X}$ in the new reference system represented by PC1. In Figure 1, the second principal component (PC2) is reported as a dashed line. It can be seen that PC2 is orthogonal to PC1, but accounts for a very limited variability in the data compared to PC1 (in fact, it accounts for the orthogonal distance of each projection from the PC1 direction). In this case, it can be therefore concluded that PC1 is sufficient to adequately describe $\mathbf{X}$.

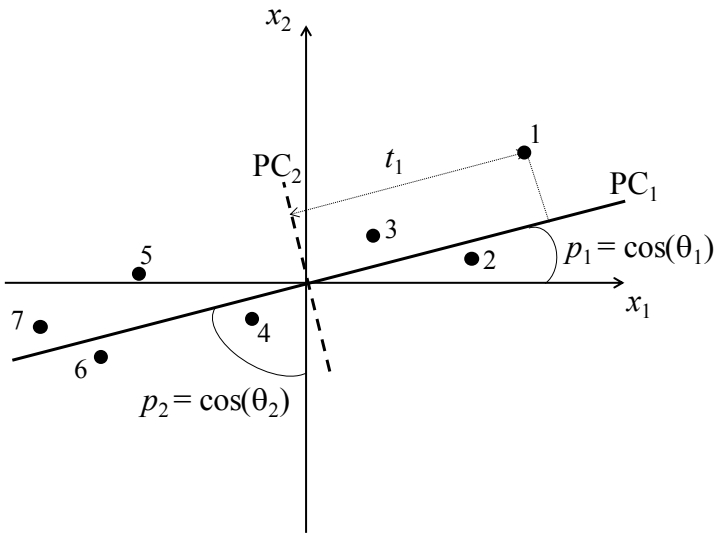


Figure 2.1. *Geometrical interpretation of the PCA scores and loadings for a dataset with 2 variables (x_1 and x_2) and 7 samples.*

A graphical representation of the PCA model scores and loadings is often used to gain understanding on the correlations among samples (through the scores) and variables (through the loadings). Additional details on the interpretation of scores and loadings plots are provided in Appendix B.

2.1.1.1 PCA data pretreatment

Before building a PCA model, the **X** data are usually pretreated. The appropriate pretreatment depends on the characteristics of the data and on the objectives of the analysis, and it may include filtering, denoising, transformations (e.g., logarithmic ones), advanced scaling and data compression (Eriksson *et al*., 2006).

In general, when managing simultaneously data with heterogeneous origin and different measurement units (e.g., process measurements such as compositions, temperatures, pressures, torques, etc., formulation details, etc.), it is important that all variables are given the same weight. To this purpose, the data are auto-scaled, i.e. the variables are mean-centered and scaled to unit variance (Wise *et al*., 1996). Mean-centering (i.e., subtracting to each column $\mathbf{x}_n$ of **X** its mean values) avoids to detect the differences among the mean values of different variables as significant directions of variability; scaling to unit variance (i.e., diving each column $\mathbf{x}_n$ of **X** by its standard deviation) makes the analysis independent of the measurement units, thus enabling the simultaneous analysis of variables with values of very different magnitudes.

2.1.1.2 Selection of the number of PCs

An additional issue to be considered in building a PCA model is how to select the number of PCs to be used in the model, i.e. the dimensionality of the latent space. Several methods have been proposed in the literature to this purpose (Valle *et al.*, 1999).

The selection of the appropriate number of PCs should take into account different issues, such as the number of samples, the total variance explained, the relative size of the eigenvalues (i.e. the variance explained per component), and the subject-matter interpretations of the PCs (Johnson and Wichern, 2007). The following approaches have been applied in this Dissertation:

- the scree test (Jackson, 1991);
- the eigenvalue-greater-than-one rule (Mardia *et al.*, 1979);
- the cross-validation based on the prediction error sum of squares (Wold, 1978).

The scree test is an empirical and graphical procedure, which is based on the analysis of the profile of an index indicating the variability of the original data captured by the PCA model per PC (e.g., the explained variance per PC, the eigenvalues or the residual percent variance). The method is based on the idea that the variance described by the model should reach a "steady-state", when additional PCs begin to describe the variability due to random errors. When a break point is found in the curve or when the profile stabilizes, that point corresponds to the number of PCs to be included in the model. The implementation of the method is relatively easy, but if the curve decreases smoothly it can be difficult to identify an "elbow" on it.

The eigenvalue-greater-than-one rule is a simple rule according to all the PCs whose corresponding eigenvalues are lower than one are not considered in the model. The basic idea behind this method is that, if data are auto-scaled, the eigenvalue corresponding to a PC represents roughly the number of original variables whose variability is captured by the PC itself. If so, a PC capturing less than one original variable should not be included in the model. Although this method is very easy to implement and automate, in some cases PCs are discarded even if their eigenvalue is very close to one and their contribution to explain the systematic variability is significant. In these cases, it may be reasonable to lower the threshold in order to include PCs whose eigenvalue may be (slightly) lower than one.

Cross-validation (Wold, 1978) is another technique which can be employed in the determination of the number of PCs. The basic idea of cross-validation is that the number of PCs to be selected to build the model is the one for which the error in reconstructing new samples through the model is minimum. When no independent validation data are available, the data in the **X** matrix itself are used to evaluate the reconstruction error (or prediction error sum of squares, *PRESS*). Different cross-validation algorithms can be employed. The general cross-validatory scheme is the following:

1. divide the **X** dataset in S subgroups $\mathbf{X}^s$ of C samples (with $s = 1,\ldots, S$);

2. delete the samples in one of the $\mathbf{X}^s$ groups from the original dataset $\mathbf{X}$;
3. build a PCA model with the reduced dataset $\mathbf{X}$;
4. project the data in $\mathbf{X}^s$ onto the PCA model built in step 3.;
5. compute $PRESS_s$ for the reconstruction of $\mathbf{X}^s$ and store it:

$$PRESS_s = \sum_{c=1}^{C}\sum_{n=1}^{N}\left(x_{c,n} - \hat{x}^s_{c,n}\right)^2 \quad , \tag{2.7}$$

being $\hat{x}^s_{c,n}$ the reconstructed element of $\mathbf{X}^s$ in the c-th row and n-th column;

6. go back to step 1 to select the next subset s until all the S subsets have been considered;
7. repeat the procedure by increasing the number of PCs used to build the PCA model.

By summing all the partial *PRESS* values per subgroup, eventually a total *PRESS* per PC is obtained. The evaluation of the *PRESS* profile can be useful to select the number of PC to build the model. Namely, a PC is included if it increases the predictive power of the model. Therefore, the number of PCs for which the minimum value of *PRESS* is found or for which a steady state in the profile is reached, should be considered (Wold, 1978).

2.1.1.3 PCA model diagnostics

Given a PCA model, several diagnostics can be used to assess its performance. In general, model, variable and sample diagnostics can be distinguished (Eriksson *et al.*, 2006).

Among model diagnostics, the most common one is the coefficient of determination R^2, which quantifies the amount of variability of the original data that the model explains:

$$R^2 = 1 - \frac{\sum_{i=1}^{I}\sum_{n=1}^{N}(x_{i,n} - \hat{x}_{i,n})^2}{\sum_{i=1}^{I}\sum_{n=1}^{N}(x_{i,n})^2} \tag{2.8}$$

where $\hat{x}_{i,n}$ is the PCA reconstruction of the element $x_{i,n}$ of the original matrix $\mathbf{X}$. If in (2.8) the summations are extended only to the I rows of the original matrix, the explained variance per variable is evaluated (R^2_{pv}). If the *PRESS* value as defined in (2.7) is used, (2.8) returns the Q^2 statistics. Q^2 can be used in alternative to *PRESS* for the selection of the PCs to include in the model and can be seen as a measure of the "predictive" power of the model. Usually $Q^2 < R^2$.

Beside diagnostics on the model performances, when a PCA model is built, it allows to calculate statistics on the data used for its calibration, in order to discover potential outliers or data that have a strong influence on the model. Two statistics are used to this purpose: the squared prediction error (SPE) and the Hotelling's T^2.

SPE measures how well the sample conforms to the PCA model, and is defined for the *i*-th sample as:

$$\mathrm{SPE}_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i = \mathbf{x}_i(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathrm{T}})\mathbf{x}_i^{\mathrm{T}} \quad , \tag{2.9}$$

where $\mathbf{e}_i$ is the [N×1] residual vector for the reconstruction of the *i*-th observation $\mathbf{x}_i$ (the *i*-th row of the residual matrix **E**), and **I** the identity matrix of appropriate size [N×N]. SPE_i measures the orthogonal distance of the *i*-th observation from the latent space identified by the model, thus accounting for the model mismatch: samples with a high value of SPE are characterized by a different correlation structure with respect to the one described by the PCA model and, as a consequence, are not represented well.

The Hotelling's T^2 statistic (Hotelling, 1933) measures the overall distance of the projections of a sample of the **X** dataset from the PC latent space origin. Since each PC of the model explains a different percentage of variance of the data, the Mahalanobis distance is used to calculate it:

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{a,i}^2}{\lambda_a} \quad , \tag{2.10}$$

where $t_{a,i}$ represents the projection of the *i*-th observation on the *a*-th PC used to build the model. The T^2 statistic is used to assess the deviation of a sample from the average conditions (the PC space origin) represented in the dataset. A sample with a large Hotelling's T^2 has a large influence to the model (high *leverage*) and should be handled with care: if it is represented well by the model (small SPE), the information it provides can be useful to expand the data space and improve model robustness.

The graphical interpretation of both the Hotelling's T^2 and the SPE is shown in Figure 2.2 for a model with one PC.

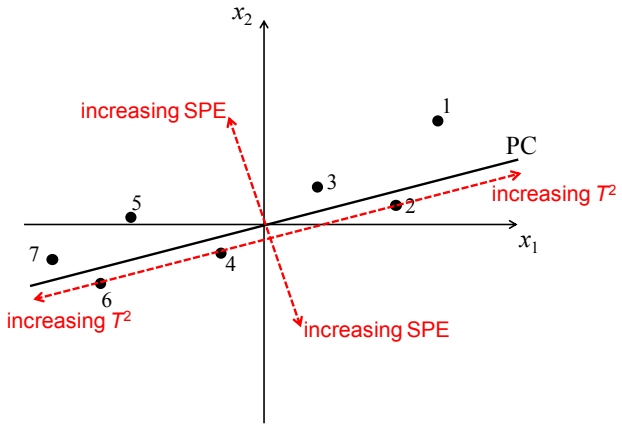


Figure 2.2. *Hotelling's T^2 and SPE geometric meaning in the 1 PC PCA model of Figure 2.1.*

Samples moving far from the center of the model, but within the model space, increase their Hotelling T^2 value, while samples moving far from the model space, independently from their position within the model space, increase their SPE value.
For both the SPE and the T^2 statistics, the contribution of each variable of a given sample to their values can be evaluated. These contributions can reveal which variables mainly determine the sample position in the score space (or outside it). A thorough discussion on the use of the variable contributions can be found in Conlin *et al*. (2000).

2.1.2 Projection to latent structures (PLS)

Projection to latent structures (PLS; Wold *et al.*, 1983; Geladi and Kowalski, 1986; Höskuldsson, 1988; Burnham *et al.*, 1996) is a regression modeling technique that relates a dataset of regressors **X** to a dataset of response variables **Y** through the projections onto their latent structure. PLS aims at finding a linear transformation of the **X** data that maximizes the covariance of its latent variables with the **Y** dataset variables. As for PCA, the optimization problem formalizing the search for the LVs can be converted into an eigenvector problem, namely the eigenvector decomposition of the joint covariance matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$:

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \quad , \tag{2.11}$$

being **w** the [$N\times1$] vector of weights representing the coefficient of the linear combination of **X**-variables determining the PLS scores **t**:

$$\mathbf{t} = \mathbf{X}\mathbf{w} \quad . \tag{2.12}$$

As shown for the PCA model, assuming that A LVs have been retained, the **X** and **Y** datasets are decomposed and related through their latent structures:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.13}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F} \tag{2.14}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \quad , \tag{2.15}$$

where **T** is the [$I\times A$] score matrix, **P** and **Q** are the [$N\times A$] and [$M\times A$] loading matrices, **E** and **F** are the [$I\times N$] and [$I\times M$] residual matrices, and $\mathbf{W}^*$ is the [$N\times A$] weight matrix, which is calculated from the weights **W** to allow interpretation with respect to the original **X** matrix:

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \quad . \tag{2.16}$$

As (2.13)-(2.15) clarify, PLS modeling returns a model for the correlation structure of **X**, a model for the correlation structure of **Y,** and a model for their mutual relation. Therefore, PLS modeling is suitable to handle reduced-rank datasets including highly correlated and noisy measurements. The basic assumption is that the spaces identified by **X** and **Y** have a common latent structure, which can be employed to relate them. Note that oftentimes in (2.14) of the PLS model the score matrix **T** is substituted by the **Y** space score matrix **U** [$I \times A$].

Figure 2.3 provides a geometrical interpretation of the PLS model. As it can be seen, the **X** data arrange mainly on a plane, defined by two latent directions. The latent directions are identified in the **X** and in the **Y** spaces in order to best approximate the directions of maximum variability of the points and to provide a good correlation between their projections along these directions. Note that, while weights **W** are orthogonal, the loadings **Q** may not necessarily be (Eriksson *et al.*, 2001).

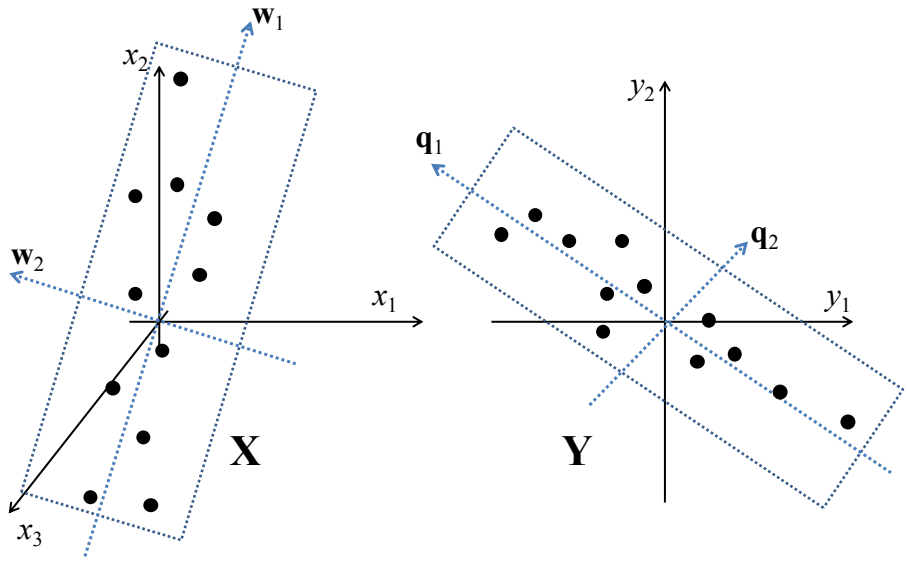


Figure 2.3. *Geometric interpretation of the PLS decomposition of data matrixes **X** and **Y**.*

More details on the algorithm used for PLS decomposition are provided in Appendix A.

2.1.2.1 Selection of the number of LVs and PLS model diagnostics

Similarly to PCA modeling, several strategies have been proposed in the literature for the selection of the number of PLS components, with cross-validation being the most popular one. The procedure is the same described in Section 2.1.1.2 for PCA and is repeated increasing at each iteration the number of LVs included in the model, obtaining a profile of *PRESS* or alternatively the root mean square error of cross-validation per LV ($RMSECV_a$):

$$RMSECV_a = \sqrt{\frac{PRESS_a}{I \cdot N}} \quad . \tag{2.17}$$

In the regression case, the *PRESS* values are calculated on the basis of the predictions of the response variables in **Y**. Accordingly, the number of LVs to consider should be the one for which $PRESS_a$ (or $RMSECV_a$) is minimum. The profile of explained variance in cross validation (Q^2) provides a similar information. However, it must be underlined that a PLS model provides a model both for the **X** and for the **Y** datasets. Hence, indices accounting for the variance captured for each dataset would be preferable to use (Tomba *et al*., 2012).
Once a PLS model is built, the diagnostics to evaluate its performances are the same as for the PCA model (Eqs.(2.8)-(2.10)). In this case, they can be applied to both **X** and **Y**. Furthermore, for a successful calibration of the PLS model and for model interpretation, it may be useful to understand which are the regressor variables that most affect the projections and that are most appropriate to build the PLS model. This can be quantified by the VIP index (variable importance in the projection; Chong and Jun, 2005; Andersen and Bro, 2010), which is defined as:

$$\mathrm{VIP}_n = \sqrt{\frac{N \sum_{i=1}^{A} R_{y,a}^2 w_{n,a}^2}{\sum_{i=1}^{A} R_{y,a}^2}} \quad , \tag{2.18}$$

where N is the total number of variables considered, $R_{y,a}^2$ is the **Y**-variance explained by the a-th LV of the model, while $w_{n,a}$ is the weight of the n-th variable on the a-th LV calculated from the PLS model. Since the sum of squares of all the N VIPs is equal to the number of terms in the model, the average VIP would be equal to 1. Variables with $\mathrm{VIP}_n \geq 1$ are therefore considered valuable predictors of the variables in **Y**.

2.1.2.2 PLS discriminant analysis (PLS-DA)

Oftentimes the data in **Y** are not in the form of quantitative responses; rather, they could be categorical variables representing the class membership of the I samples in **X**. In this case, regressing **Y** against **X** using the decomposition given by (2.8)-(2.10) means using the PLS model for a discriminant analysis (PLS-DA; Barker and Rayens, 2003), i.e. to classify the I samples or, stated differently, to determine the maximum directions of variability in **X** which are responsible for the partition of the I samples into the M classes in **Y**. The class membership of the i-th sample in **X** is encoded in M binary variables in **Y**. If the i-th sample belongs to the m-th class, then:

$$\begin{cases} y_{i,j} = 0 & \text{for } j \neq m \\ y_{i,j} = 1 & \text{for } j = m \end{cases} \quad \text{with } j = 1,...,M \quad . \tag{2.19}$$

Since the output of the PLS-DA model is not in the form of 0's and 1's, but instead a real number spanning a range wider than [0,1], a threshold is usually chosen to define the class membership. Following a Bayesian approach (with the assumption that the predictions within each class are approximately normally distributed), the threshold value is determined in such a way as to return the best possible split among classes with the least probability of false classification of future predictions (Fawcett, 2006).

Binary classification results are usually expressed in terms of a confusion matrix such as the one reported in Figure 2.4. Considering a classification problem with i_1 samples belonging to Class 1 and i_2 samples belonging to Class 2 ($i_1 + i_2 = I$), the PLS-DA classifier assigns p_1 samples to Class 1 and p_2 samples to Class 2 ($p_1 + p_2 = I$). Given a reference class, the samples belonging to that class can be either correctly assigned to it (true positives, TP) or erroneously attributed to the other one (false negatives, FN). Similarly, samples not belonging to the reference class can be correctly assigned to the other class (true negatives, TN), or erroneously attributed to the reference class (false positives, FP).

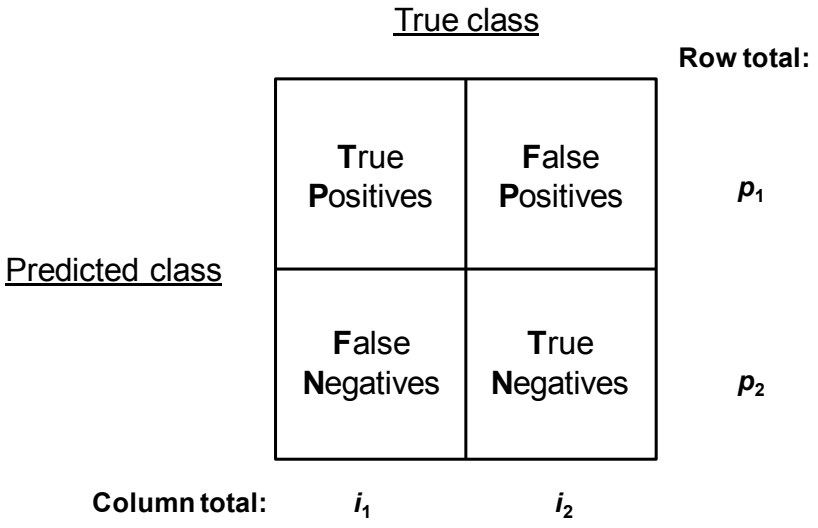


Figure 2.4. *Confusion matrix for a binary classification problem (adapted from Fawcett, 2006).*

From the confusion matrix of Figure 2.4, three metrics are usually generated:

- sensitivity (toward one class), i.e. the percentage of samples that belong to the reference class and are correctly classified, i.e. TP/i_1;
- specificity (toward one class), i.e. the percentage of samples that belong to the other class and correctly classified, i.e. TN/i_2;
- accuracy, i.e. the percentage of samples correctly classified, i.e. (TP + TN)/I.

2.1.3 Monitoring charts

Once a PCA or a PLS(-DA) model has been calibrated on the available dataset, the model can be used to assess whether a new sample ($\mathbf{x}^{NEW}$) conforms to the data used to build it or not. The scores ($\hat{\mathbf{t}}^{NEW}$) of the new sample can be calculated by projecting $\mathbf{x}^{NEW}$ onto the reduced latent space of the model:

$$\begin{cases} \hat{\mathbf{t}}^{NEW^T} = \mathbf{x}^{NEW}\mathbf{P} & \text{for PCA} \\ \hat{\mathbf{t}}^{NEW^T} = \mathbf{x}^{NEW}\mathbf{W}^* & \text{for PLS} \end{cases} . \tag{2.20}$$

After evaluation of the scores, the reconstruction of $\mathbf{x}^{NEW}$ can be computed as:

$$\hat{\mathbf{x}}^{NEW^T} = \mathbf{P}\hat{\mathbf{t}}^{NEW} , \tag{2.21}$$

and, in the case of the PLS model, a prediction of the response variables ($\hat{\mathbf{y}}^{NEW}$) can be obtained as:

$$\hat{\mathbf{y}}^{NEW^T} = \mathbf{Q}\hat{\mathbf{t}}^{NEW} . \tag{2.22}$$

Given the values of $\hat{\mathbf{t}}^{NEW}$ and $\hat{\mathbf{x}}^{NEW}$, the Hotelling's T^2 and the SPE statistics of the new sample can be evaluated in order to quantify its deviation from the average conditions of the data used to build the model and its deviation from the model space, respectively. Based on the values of the statistics for the data used in the calibration step, confidence limits can be set (Mardia *et al.*, 1979).

In particular, the scores have zero mean, variance equal to their associated eigenvalues and are orthogonal. Assuming that the data used to build the model are independent and identically distributed, the scores are normally distributed. Therefore, for the scores on the *a*-th LV, an univariate confidence limit can be calculated from the critical value of the Student's t-distribution, with $I - 1$ degrees of freedom at significance level α:

$$t_{(1-\alpha)\text{lim}}(a) = \pm t_{I-1,\alpha/2} \cdot \sqrt{\lambda_a} . \tag{2.23}$$

Under this assumption, the Hotelling's T^2 can be well-approximated as a Fisher's *F*-distribution, being it computed from the ratio of approximately normal variables. Its relevant confidence limit can therefore be estimated as (Mardia *et al.*, 1979):

$$T^2_{(1-\alpha)\text{lim}}(A, I) = \frac{A \cdot (I^2 - 1)}{I \cdot (I - A)} \cdot F_{A,I-A,\alpha} , \tag{2.24}$$

where $F_{A,I-A,\alpha}$ is the critical value of the F distribution with A and $I-A$ degrees of freedom at significance level α. This determines an ellipsoidal confidence region in the A-dimensional score space, whose semi-axes are:

$$sa_a = \sqrt{\lambda_a T^2_{(1-\alpha)\text{lim}}(A, I)} \quad \text{with } a = 1, \ldots, A \quad . \tag{2.25}$$

In particular, to allow a visual representation, confidence ellipses can be determined through (2.25) for the projections of the scores of data in bi-dimensional planes.

The SPE statistic is a sum of squared errors, which can be assumed to follow a normal distribution (Jackson, 1991). As a consequence, SPE can be approximated as a χ^2-distribution, and its relevant limit calculated as follows:

$$\text{SPE}_{(1-\alpha)\text{lim}} = [\nu/(2\mu)] \cdot \chi^2_{2\mu^2/\nu,\alpha} \quad , \tag{2.26}$$

where $\chi^2_{2m^2/\nu,\alpha}$ is the critical value of the χ^2-distribution with $2\mu^2/\nu$ degrees of freedom at the significance level α; μ and ν are respectively the mean and the variance of the SPE values of the data used to build the model (Nomikos and MacGregor, 1995).

On the basis of the computed confidence limits, monitoring charts can be built for the scores, the Hotelling's T^2 and SPE. In particular, when a new sample is available, the mentioned statistics are compared with the relevant confidence limits to judge the similarity and the adherence of $\mathbf{x}^{\text{NEW}}$ to the data used to build the model. Being multivariate indices, charts on T^2 and SPE are more effectively used to this purpose, by observing that:

$$\begin{cases} T^2_{\mathbf{x}^{\text{NEW}}} \le T^2_{(1-\alpha)\text{lim}} \\ \text{SPE}_{\mathbf{x}^{\text{NEW}}} \le \text{SPE}_{(1-\alpha)\text{lim}} \end{cases} . \tag{2.27}$$

If conditions in (2.27) are satisfied, $\mathbf{x}^{\text{NEW}}$ is considered in a state of statistical control with a $100(1-\alpha)\%$ probability; otherwise an occurrence of *special cause variability* is detected. This occurrence may be due to a change in the mean conditions ($T^2_{\mathbf{x}^{\text{NEW}}} > T^2_{(1-\alpha)\text{lim}}$) or in the representativeness of the model ($\text{SPE}_{\mathbf{x}^{\text{NEW}}} > \text{SPE}_{(1-\alpha)\text{lim}}$) compared to the *common cause variability* characterizing the data used to build the model. The procedure in (2.27) is equivalent to test the hypothesis that $\mathbf{x}^{\text{NEW}}$ complies with the calibration (i.e. historical) data (Johnson and Wichern, 2007).

2.2 Latent variable modeling for image analysis applications

Images are three-way arrays $\underline{\mathbf{X}}$ [$N_{row} \times N_{col} \times C$] whose first two dimensions are the number of pixel rows (N_{row}) and columns (N_{col}), and the third dimension (C) represents the light intensity of the spectral channels (i.e. wavelengths) excited by the camera. Images can be classified according to the value of C:

- $C = 1$: grayscale images, allowing one to explore only the textural features of the product being imaged;
- $C = 3$: RGB (red, green, blue) images, such as those collected with a common compact digital camera, allowing one to explore also the color characteristics of the product being imaged;
- $C > 3$: multi- or hyperspectral images, where for each of the ($N_{row} \cdot N_{col}$) pixels a spectrum in the visible or infrared region is collected, allowing one to analyze also the chemical structure of the product.

Due to the prominent use of RGB images (and of their corresponding grayscale images) in the Dissertation, only the case $C = 3$ will be discussed in the remaining. The following sections introduce the use of LVMs to effectively extract color- and texture-related information from RGB images.

2.2.1 Color analysis

The use of RGB images to explore the color characteristics of a product was pioneered by Geladi and coworkers (Geladi *et al.*, 1989; Geladi and Esbensen, 1989; Esbensen and Geladi, 1989), who introduced the concept of multivariate image analysis (MIA). MIA relies on the PCA decomposition of an image $\underline{\mathbf{X}}$ properly unfolded into a two-way matrix of size [($N_{row} \cdot N_{col}$)×3] prior to the factorization, as depicted in Figure 2.5 for an image of four color standards (red, green, blue and yellow; Edmund Optics T56-079, USA).

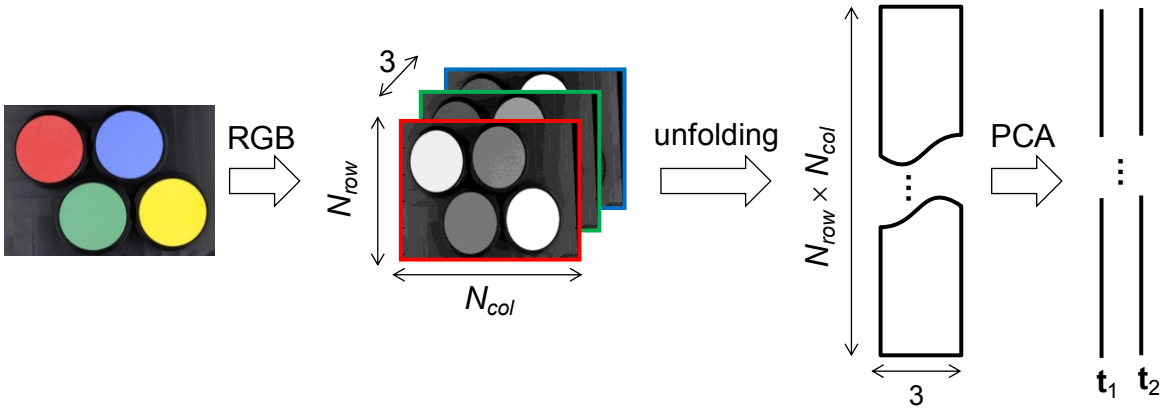


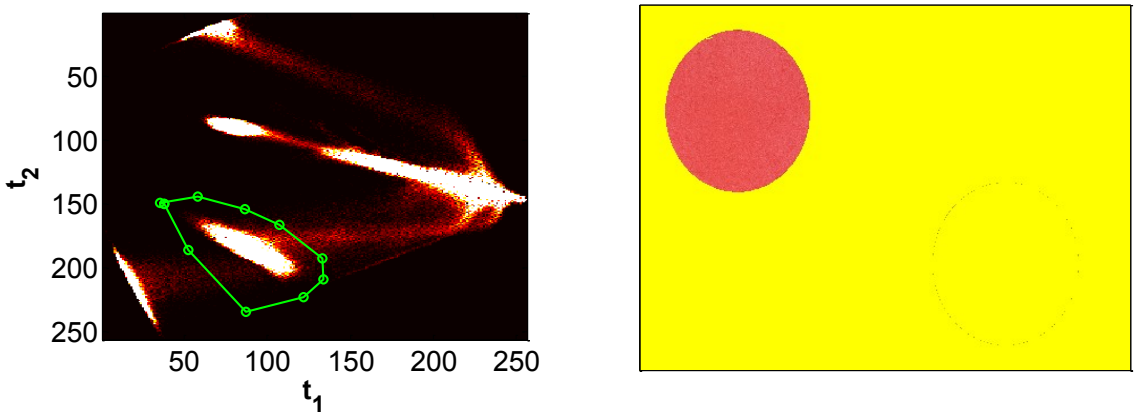
Figure 2.5. *Schematic of multivariate image analysis (MIA).*

As shown in Figure 2.5, an MIA model is usually built on 2 PCs, as they account for more than 90% of the pixel variability of an RGB image. The scores $\mathbf{t}_1$ and $\mathbf{t}_2$ are scaled within

the range [0-255], which is the same range spanned by the light intensity values (Yu and MacGregor, 2004):

$$\mathbf{t}_i = \text{round}\left[255 \cdot \frac{\mathbf{t}_i - \min(\mathbf{t}_i)}{\max(\mathbf{t}_i) - \min(\mathbf{t}_i)}\right] \quad . \tag{2.28}$$

The score space is usually represented in terms of a two dimensional (2D) histogram-scatter (or density) plot, as shown in Figure 2.6 for the image of Figure 2.5 (the color is proportional to the number of pixels having specified $\mathbf{t}_1$ and $\mathbf{t}_2$ coordinates; i.e. the lighter the color, the larger the number of pixels). Pixels having projections close to each other in the score space are similar in terms of color structure, regardless of their spatial arrangement in the original image. This can be easily seen by defining a mask (i.e. a geometrical shape highlighting a certain region of the $\mathbf{t}_1\mathbf{t}_2$ space, Figure 2.6a) and looking at the pixels in the original image that project within it. An example is given in Figure 2.6, where the mask is used to highlight the red pixels in the original image.



(a) *(b)*

Figure 2.6. *(a) 2D histogram-scatter plot of the score space of the MIA model built on the image of Figure A1, with a mask defined on it. (b) pixels (in the original image) whose projection lay underneath the mask defined in (a).*

In order to predict a color-related property from an image, the image itself has to be represented in terms of feature variables to be regressed using PCA or PLS. Yu and MacGregor (2003) demonstrated that the 2D histogram-scatter plot is an excellent starting point for features extraction, as it allows one to (partially) filter out the effect of a non-homogeneous illumination, which can instead affect other parameters (potential regressors of the property of interest) such as the average color of the red, green and blue channels. The most effective method proposed by Yu and MacGregor (2003) is the *covariance mask* method. The aim of the method is that of lumping the histogram elements of the $\mathbf{t}_1$-$\mathbf{t}_2$ space (discretized into a 256×256 grid as in Figure 2.6) into histogram bins that are expected to have the same value of the property of interest $\mathbf{y}$ [I×1]. The method is based on the

covariance properties between the counts of the histogram elements in the $\mathbf{t}_1$-$\mathbf{t}_2$ space and two variables $\mathbf{z}_1$ and $\mathbf{z}_2$ related to $\mathbf{y}$ and defined as:

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{y} \\ \mathbf{z}_2 &= \left| \mathbf{z}_1 - \frac{1}{2}[\max(\mathbf{y}) + \min(\mathbf{y})] \right| \end{aligned} \quad . \tag{2.29}$$

For an image i (i = 1, 2, …, I) the relative histogram can be represented as:

$$\mathbf{h}_i = [h_i(B_1) h_i(B_2) \cdots h_i(B_J)] \quad , \tag{2.30}$$

where $J = 256^2 = 65536$ for a 256×256 grid and $h_i(B_j)$, j = 1, 2, …, J, is the pixel count for element B_j divided by the total number of image pixels, i.e. an estimate of the probability of pixels falling into the j-th element. A matrix $\mathbf{\Gamma}$ can be constructed by stacking on the top of each other the relative histograms of all available I images:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_I \end{bmatrix} = \begin{bmatrix} h_1(B_1) & h_1(B_2) & \cdots & h_1(B_J) \\ h_2(B_1) & h_2(B_2) & \cdots & h_2(B_J) \\ \vdots & \vdots & \ddots & \vdots \\ h_I(B_1) & h_I(B_2) & \cdots & h_I(B_J) \end{bmatrix} = [\mathbf{H}(B_1)\mathbf{H}(B_2)\cdots\mathbf{H}(B_J)] \quad , \tag{2.31}$$

where $\mathbf{H}(B_j)$ is the j-th column of matrix $\mathbf{\Gamma}$ for each bin B_j. It can be shown that $\mathbf{H}(B_j)$ and $\mathbf{H}(B_k)$ have the same direction if bins B_j and B_k represent the same level of the property in $\mathbf{y}$. This property can be used to lump histogram bins with similar values of the property. From a general point of view, one can assume that differences in color among different pixels in an image can be attributed to two independent factors, i.e. the local value of the property of interest and the illumination. Considering two histogram elements B_j and B_k having similar average property value y, then for the i-th image:

$$\begin{aligned} \mathbf{h}_i(B_j) &= \mathbf{h}_i(y) \cdot s_j \\ \mathbf{h}_i(B_k) &= \mathbf{h}_i(y) \cdot s_k \end{aligned} \quad , \tag{2.32}$$

with s_j and s_k representing the local average imaging conditions. Assuming that lighting conditions do not change, these values can be considered constant and hence:

$$\begin{aligned} \mathbf{H}(B_j) &= \mathbf{H}(y) \cdot s_j \\ \mathbf{H}(B_k) &= \mathbf{H}(y) \cdot s_k \end{aligned} \quad \ldots, \tag{2.33}$$

i.e. for two histogram elements B_j and B_k with the same property value, $\mathbf{H}(B_j)$ and $\mathbf{H}(B_k)$ will have the same relationship with $\mathbf{y}$, but of different magnitude. Computing the covariance between $\mathbf{H}(B_j)$ and $\mathbf{z}_1$ and $\mathbf{z}_2$:

$$\begin{aligned} \mathrm{cov}_1(B_j) &= \mathrm{cov}[\mathbf{H}(B_j), \mathbf{z}_1] = \mathrm{cov}[\mathbf{H}(y), \mathbf{z}_1] \cdot s_j \\ \mathrm{cov}_2(B_j) &= \mathrm{cov}[\mathbf{H}(B_j), \mathbf{z}_2] = \mathrm{cov}[\mathbf{H}(y), \mathbf{z}_2] \cdot s_j \end{aligned} \quad , \tag{2.34}$$

allows one to cancel out the scalar s_j by evaluating the phase angle of the observed point in the space of cov_1 versus cov_2:

$$\theta(B_j) = \arg[\mathrm{cov}_1(B_j), \mathrm{cov}_2(B_j)] = \theta(y) \, . \tag{2.35}$$

If $\theta(y)$ and y have a one-to-one mapping relationship, θ can be segmented into a finite number of bins and each bin should represent a different level of the property of interest. The application of this method will be shown in Chapters 3 and 4 in the case of film-coated tablets.

2.2.2 Texture analysis

Texture is defined as a descriptor of local brightness variation from pixel-to-pixel in a small neighborhood (Russ, 1999). Texture analysis is intended to quantify surface characteristics such as coarseness, roughness, smoothness and the presence of defects. It requires the RGB image $\underline{\mathbf{X}}$ to be converted into a grayscale image $\mathbf{X}$ using:

$$\mathbf{X} = 0.2989\mathbf{R} + 0.5870\mathbf{G} + 0.1140\mathbf{B} \quad , \tag{2.36}$$

where $\mathbf{R}$, $\mathbf{G}$ and $\mathbf{B}$ are $[N_{row} \times N_{col}]$ matrices representing the red, green and blue color channels of $\underline{\mathbf{X}}$.

Several methodologies for texture analysis of grayscale images have been proposed in the literature, such as statistical methods (Haralick *et al.*, 1973), model-based methods (Cross and Jain, 1983) and transform-based methods (Geladi, 1992). A combination of multivariate techniques (PCA) and transform-based methods has been proposed as well (Bharati *et al*., 2004; Liu and MacGregor, 2007). Wavelet texture analysis (WTA), which is based on wavelet transform (Addison, 2002), is considered to be the current state of the art for the extraction of textural features (Liu and Han, 2011; Duchesne *et al*., 2012). Details on wavelet transform and its use in image analysis are given in the following.

2.2.2.1 Wavelet transform

Wavelet transform analysis uses little wavelike functions known as *wavelets*. Wavelets are used to transform the signal under investigation into another representation that presents the signal information in a more useful form. This signal transformation is known as *wavelet transform* and, mathematically speaking, is a convolution of the wavelet function with the signal. The wavelet can be manipulated in two ways: it can be moved to various location of the signal, and it can be stretched or squeezed. If the wavelet matches the shape of the signal well at a specific scale and location, then a large transform value is obtained (Mallat, 1989).

Considering a generic time-dependent (t) continuous signal $x(t)$, and given an analyzing wavelet $\psi(t)$ such as the Mexican hat wavelet:

$$\psi(t) = (1-t^2)\exp\left(-\frac{t^2}{2}\right) \quad , \tag{2.37}$$

and its dilated and scaled version:

$$\psi\left(\frac{t-b}{a}\right) = \left[1-\left(\frac{t-b}{a}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{t-b}{a}\right)^2\right] \quad , \tag{2.38}$$

with a and b respectively the dilation and the location parameters, the continuous wavelet transform $T(a, b)$ of x is defined as:

$$T(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt = \int_{-\infty}^{+\infty} x(t)\psi^*_{a,b}dt \quad , \tag{2.39}$$

where the asterisk indicates that the complex conjugate of the wavelet function is used in the transform. In practice, only discrete values of a and b are considered, resulting in the discretized wavelet:

$$\psi_{m,n}(t) = \frac{1}{a_0^m}\psi\left(\frac{t-nb_0a_0^m}{a_0^m}\right) \quad , \tag{2.40}$$

where the integers m and n control the wavelet dilation and translation respectively, a_0 is a specified fixed dilation step parameter (generally 2), and b_0 is the location parameter (generally 1). The wavelet transform of a continuous signal x using the discrete wavelet (2.40) is then:

$$T_{m,n} = \int_{-\infty}^{+\infty} x(t)\psi_{m,n}dt \quad . \tag{2.41}$$

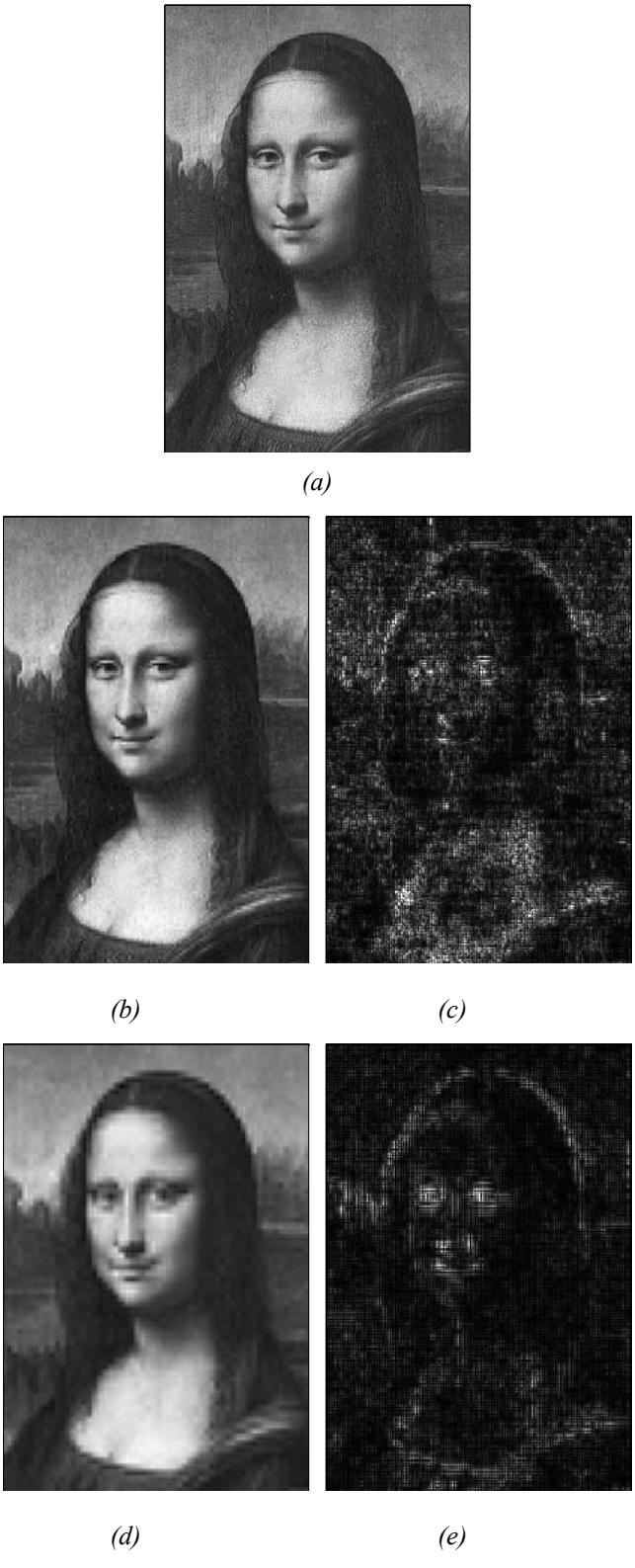
The values $T_{m,n}$ are known as *detail coefficients*. Orthonormal dyadic discrete wavelets such as those arising from the choice $a_0 = 2$ and $b_0 = 1$ are associated with scaling function $\phi_{m,n}$ that are in turns associated with the smoothing of the signal. The scaling functions can be convolved with the signal as in (2.41) to produce the *approximation coefficients* $S_{m,n}$ (Addison, 2002). The signal $x(t)$ can be eventually represented using a combined series expansion as:

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m_0,n}\Phi_{m_0,n} + \sum_{m=-\infty}^{m_0}\sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t) \quad , \tag{2.42}$$

i.e the original signal can be expressed as a combination of an approximation of itself, at arbitrary scale index m_0, added to a succession of signal details from scales m_0 down to negative infinity. From a practical perspective, the wavelet transform is equivalent to a sequence of low-pass and high-pass filters that split, at each scale, the signal (or its approximation) into its detail (high-frequency element) and its approximation (low-resolution element to be further decomposed). Hence, (2.42) returns a multiresolution representation of the original signal.

The discrete wavelet transform can be applied also to bi-dimensional signals such as images. In this case, since the decomposition of (2.41) is applied along two directions, at each decomposition stage an approximation (**A**) and three details images (horizontal $\mathbf{D}^h$, vertical $\mathbf{D}^v$, and diagonal $\mathbf{D}^d$) are generated. An example of wavelet decomposition is given in Figure 2.7, where a unique detail image **D** (= $\mathbf{D}^h + \mathbf{D}^v + \mathbf{D}^d$) is presented. The approximations (Figs. 2.7b and d) are blurred versions of the original image (Fig. 2.7a), while the details (Figs. 2.7c and e) capture distinctive features of the image (such as the head contour, the eyes or the mouth).

Textural features such as the mean, standard deviation, skewness, kurtosis, energy and entropy of the light intensity values can be extracted from either the approximations or the details (Liu and MacGregor, 2007; Facco *et al.*, 2010). The textural features of the available images, properly organized into a feature matrix, can be analyzed using correlative techniques such as PCA, PLS or PLS-DA.



(a)

(b) *(c)*

(d) *(e)*

Figure 2.7. 2D *discrete wavelet transform applied to the grayscale image of Leonardo's Mona Lisa. (a) Original image. (b-d) approximations and (c-e) details at the first and second scale of decomposition.*

Chapter 3

Multivariate image and texture analysis to support pharmaceutical process understanding*

The use of multivariate image and texture analysis is proposed in this Chapter to quantitatively characterize the elegance of film-coated tablets. Four unsupervised metrics are developed to quantify both the color uniformity of tablet faces/bands and the surface erosion. Latent variable modeling is used to regress the measured elegance against coating operating conditions to investigate the driving forces acting on the system and guide pharmaceutical manufacturing.

The Chapter is organized as follows. In Section 3.1, the problem investigated is briefly outlined. In Section 3.2, the available data and the techniques used for their manipulation are described. Results are presented in Section 3.3, showing the effectiveness of image analysis for process understanding purposes. Section 3.4 presents the conclusions to the Chapter.

3.1 Introduction

The Quality by Design (QbD) initiative was launched a decade ago with the aim of introducing a novel approach to pharmaceutical development and manufacturing. One of the main objectives of this initiative is ensuring the desired product quality *since its design*, rather than by final inspection. Within this initiative, companies are encouraged to design and validate their processes within a multidimensional domain of processing conditions, referred to as the process *design space* (ICH, 2009). The proposed design space should be supported by the scientific understanding of the driving forces acting on the system and governing the complex network of interactions between materials, process and product. Particularly, the framework should be based on metrics that are robust and reproducible, and not on qualitative ones that might be easily biased by human perception (García-Muñoz and Carmody, 2010).

* Ottavian, M., M. Barolo and S. García-Muñoz. Multivariate Image and Texture Analysis to Investigate the Erosion Mechanism of Film-Coated Tablets: An Industrial Case Study. *J. Pharm. Innov.*, in press.

Film coating is a common processing step in the manufacture of tablets (Libermen and Lachman, 1981). Tablets are loaded inside a rotating pan and sprayed with an aqueous or non-aqueous solution, and air is used to evaporate the solvent. Tablet coating is carried out for several reasons. It can enhance the tablet stability, because the tablet core may contain a substance that is not compatible with light and/or subject to atmospheric oxidation. Also tablet mechanical integrity (i.e., higher resistance to mishandling) can be enhanced by coating. Furthermore, coating can cover a bitter taste or an unpleasant odor of a substance within the tablet core, or it can modify the drug release profile (e.g., enteric coating, osmotic pump, pulsatile delivery; Libermen and Lachman, 1981, Cole *et al.*, 1995).
The elegance of film-coated tablets have been usually related to color uniformity and surface finish (roughness/erosion), and several techniques have been proposed for its characterization (Ruotsalainen *et al.*, 2003; Seitavuopio *et al.*, 2006; Zeitler *et al.*, 2006 and 2007; Ho *et al.*, 2007; Roggo *et al.*, 2005; Zhang *et al.*, 2005; Sasic, 2007). Among them, the use of multivariate image and multivariate wavelet texture analysis has been suggested by García-Muñoz and Gierer (2010) and García-Muñoz and Carmody (2010). The use of MIA and MWTA is attractive for practical industrial applications (Simon *et al.*, 2010 and 2012), where the elegance assessment is still typically performed by a trained panel of experts and, as such, may suffer from reproducibility issues.
In this Chapter, it is shown how MIA and MWTA can be combined together to quantify the elegance of film-coated tablets through several indices with the aim of supporting pharmaceutical manufacturing. In order to develop robust statistics, more than 7000 tablets coated in nine different pilot-scale batches have been considered. Latent variable modeling is used to regress the measured elegance (in terms of color uniformity and surface erosion) against the coating operating parameters and to investigate the main driving forces governing the process. Namely, the model diagnostics are used to investigate the mechanism leading to tablets erosion, allowing to evaluate the effect of different physical phenomena.

3.2 Materials and methods

3.2.1 Imaging conditions and available data

The in-house developed imaging station used in this study was equipped with a digital single-lens reflex Canon EOS 40D camera (10.1 megapixel resolution) with a Canon EF-S 60 mm f/2.8 USM Macro lens. LED lights illuminated the subject, and the system was isolated from the outside and operated through a computer. The settings of both the station (i.e. camera and lights elevation) and the camera (i.e. shutter speed, lens aperture and ISO

sensibility) were selected to ensure proper exposure of the subject and to highlight the features of interest on the tablets surface.

Images were collected off-line for nine different batches, indicated as Batches A-I. For color uniformity assessment, the tablets were withdrawn from the coater during the coating process, and tablet faces and bands were imaged separately. Several tablets were photographed per each single image; furthermore, three images of faces and three images of bands were taken for each pull point. For tablet erosion assessment, the tablets were imaged at the end of the coating process (one tablet per image), using an angle of incidence of the light onto the tablets surfaces that allowed to highlight the defects (if any). Details on the number of images collected for each batch are given in Table 3.1, together with a rough judgment on the surface finish. This judgment is only intended to help interpreting the PLS model diagnostics (see Figure 3.4).

Table 3.1. *Images collected for each batch.*

Batch	Number of pull point (color uniformity)	Number of tablets imaged (erosion analysis)	Surface finish
A	4	1308	Not eroded
B	6	1142	Mildly eroded
C	7	994	Mildly eroded
D	6	765	Not eroded
E	6	1148	Not eroded
F	5	234	Mildly eroded
G	7	768	Mildly eroded
H	6	669	Eroded
I	5	768	Eroded

Batches A-H were used for PLS model calibration, while batch I for validation of the proposed elegance metrics. Considering the main intent of proposing an image-based methodology to investigate on the erosion mechanism of film-coated tablets, the PLS model was built on the highest number of available data in order to span the entire variability observed experimentally in terms not only of the elegance metrics, but also of processing conditions.

The coating process was run in two phases, exploring different duration/spray rate combinations (the duration of the first phase was typically much lower than that of the second phase). A detailed list of the process operating conditions used to characterize each batch is given in Table 3.2.

Table 3.2. *Film coating operating parameters.*

Name	Description
Duration 1	Phase 1 duration
Duration 2	Phase 2 duration
Spray 1	Phase 1 spray rate
Spray 2	Phase 2 spray rate
Exhaust T1	Phase 1 exhaust air temperature
Exhaust T2	Phase 2 exhaust air temperature
Pan load	Amount of loaded tablets
Pan speed	Pan rotational speed
Hardness	Average tablets hardness

Note that other variables that might have affected either the coating uniformity or the surface erosion (such as the gun-to-bed distance, air pressure, air flowrate, etc.; García-Muñoz and Carmody, 2010; Cole *et al.*, 1995) were not changed in the batches considered here, and hence they did not represent a source of variability to be included in the regression model.

3.2.2 Multivariate image and texture analysis

MIA and MWTA were used in the characterization of tablets elegance. In the following, the metrics developed to assess color uniformity and surface erosion and defects are presented. The color uniformity and surface erosion metrics are unsupervised ones, i.e. there are no explicit target values for them.

3.2.2.1 Color uniformity assessment

The color signature was used to characterize tablets color uniformity (Yu and MacGregor, 2003). The color signature is an unbiased metric that evolves as long as coating material is applied to the tablets, until a certain cosmetic end-point is reached. For each lot of tablets, the following procedure was used:

1. training of an MIA model on the composite image obtained from the concatenation of all the images available for the lot;
2. background removal, by fitting a PCA model on the scores obtained from the projection of some background images onto the MIA model;
3. manipulation of the 2D histogram-scatter plots according to the covariance mask method, using the time of each withdrawn as the dependent variable;

4. training of a 1-PC PCA model on the matrix obtained by stacking on the top of each the features vectors extracted from the 2D histogram-scatter plots, the score $\mathbf{t}_1$ representing the color signature;
5. application of the covariance mask and of the PCA model defined in steps 3 and 4 (respectively) to the scores resulting from the projection of the sub-images (approximately of the size of one tablet) extracted from each available image on the MIA model of step 1.

The last step allowed to extract the color signature for both tablets faces and bands (separately), as shown in Figure 3.1 for one of the five available lots (batch D).

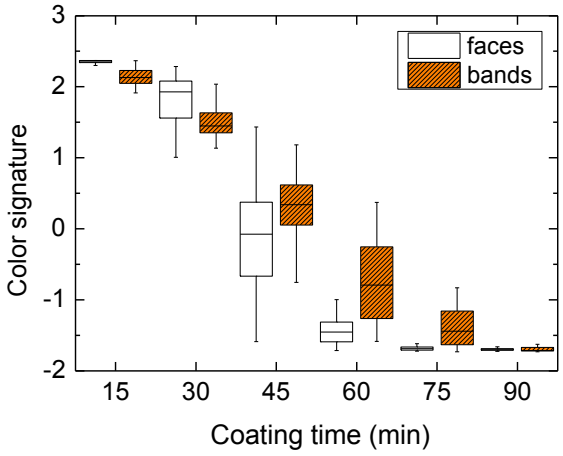


Figure 3.1 *Color signature evolution of faces (open boxes) and bands (dashed boxes) for batch D.*

Color uniformity was monitored through the color signature range (interquartile range, i.e. the difference between the 25th and 75th percentiles, to avoid including outliers): at the end of the coating process, the lower the range, the better the color uniformity across tablets.

3.2.2.2 Erosion assessment

The application of MWTA for erosion quantification required several preprocessing steps on the images prior to the analysis, in order to perfectly align and properly cut all tablet images to the greatest area around the commercial logo. The preprocessing was automated: tablets were localized within each image by using the derivative of the summation of the grayscale intensities along the two spatial directions, while the cropping operation involved simple trigonometric calculations (García-Muñoz and Carmody, 2010). After cropping, images of tablets surface where converted to grayscale. An example of two tablets surfaces characterized by a different level of erosion is given in Figure 3.2a-b.

With respect to the images of Figure 2, wavelet texture analysis (as potentially all texture analysis techniques) was complicated by the presence of the commercial logo. As suggested by Russ (1999) and reported by Bharati *et al.* (2004), texture can be defined as a descriptor of local brightness variation from pixel-to-pixel in a small neighborhood. Hence,

texture analysis was found to be biased towards the detection of the commercial logo, since it localizes the greatest pixel-to-pixel brightness variations. For this reason, the erosion quantification exercise was split into two separate problems, namely erosion quantification inside the logo and outside the logo.

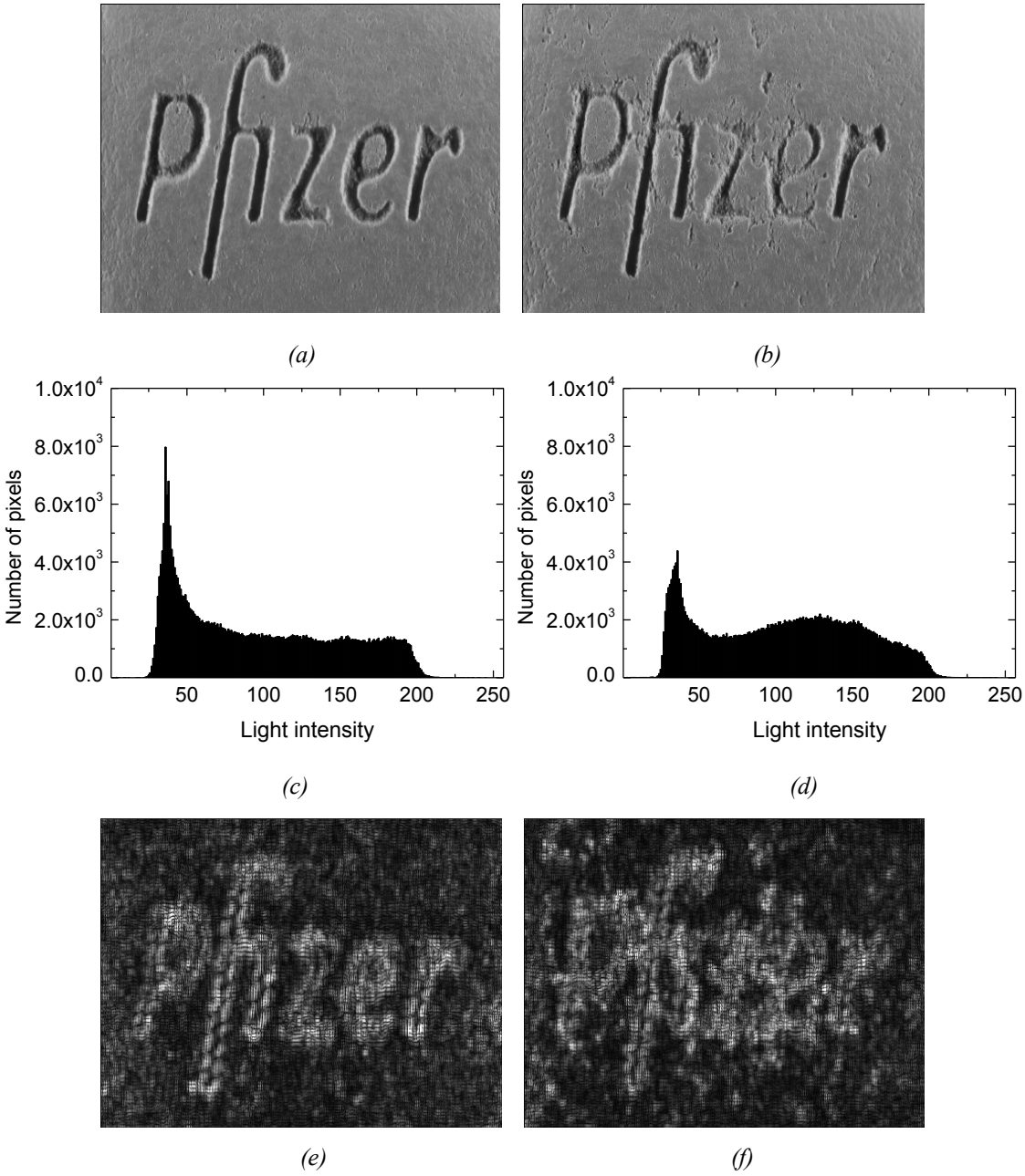
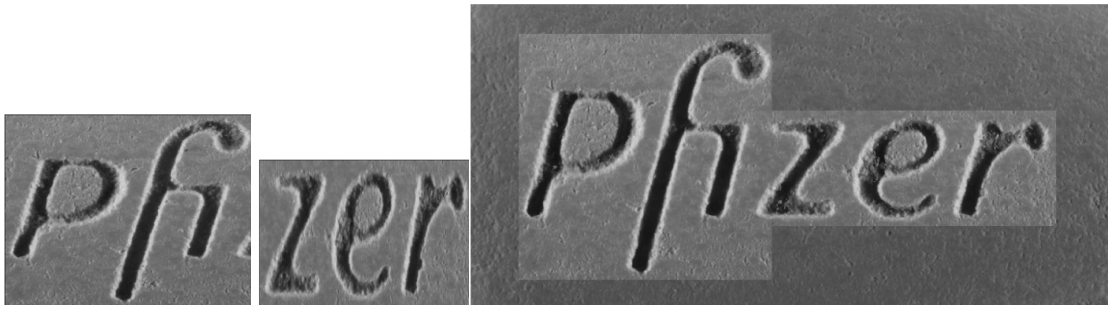


Figure 3.2 *Sample images of two tablets characterized by different levels of erosion. (a) and (b) cropped images, (c) and (d) intensity distributions of the logo pixels (inside-logo erosion assessment) and (e) and (f) 4th stage wavelet details (outside-logo erosion assessment) for images (a) and (b), respectively.*

A template matching technique (Lewis, 1995) was used to extract the commercial logo from the images of the tablets surface. The template matching algorithm is based on the

fast normalized cross correlation, and aims at determining the position of a (small) child image (e.g., the commercial logo) into a (big) parent one (e.g., the surface of a tablet).
For the case study investigated here, an image with a clearly readable commercial logo was manually cropped to the area around the logo, and then used as child image (to reduce the computational burden, the image was split into two, as the time require for the fast normalized cross correlation evaluation is proportional to the size of the child and parent images). Every time a new tablet surface image was processed, the matching algorithm localized exactly the position of the logo by using the child image as the template. Once the position was detected, a mask (i.e. a binary image of 0's and 1's) was used to separate the logo from the tablet surface. An example of the overall procedure is shown in Figure 3.3.



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

Figure 3.3 *Template matching for commercial logo removal. (a)-(b) child images of the commercial logo and (c) their superimposition on the tablet surface for the logo localization.*

The erosion quantification inside the logo relied on the rationale that the darker is the logo, the more readable it is (see Figure 3.2a and 3.2b for a comparison). Hence, the light intensity distribution of the logo pixels (i.e. the distribution of the values of the light intensity for each pixel of the logo) was used to characterize each tablet. For the two tablets of Figure 3.2a-b, the extracted light intensity distributions are shown in Figure 3.2c-d. In order to develop a synthetic descriptor of the erosion inside the logo (which will be indicated in the following as "logo index"), the following procedure was used:

1. for each lot, evaluation of the average light intensity distribution;
2. training of a 1-PC PCA model on the matrix built from the average light intensity distributions, the score $\mathbf{t}_1$ being the average logo index for the batch;
3. for each lot, projection of the light intensity distribution of each tablet onto the PCA model, generating a $\mathbf{t}_1$ distribution.

With respect to the scores of the PCA model calibrated at step 2, it was observed that higher values of the logo index indicated higher (average) erosion of the logo itself.
Erosion outside the logo was evaluated by means of the wavelet transform (using the Coiflets 5 wavelet; Addison, 2002). Only the details at the fourth decomposition stage (see

Figure 3.2e-f for the tablets of Figure 3.2a-b) were considered, since higher stages extracted irrelevant information. The pixels highlighted in white in Figure 3.2e-f identify the discontinuities in the tablets surfaces, i.e. the defects. Thus, the lighter the detail image, the higher its energy index (E_4), the more eroded the tablet. For the computation of the energy index, the pixels belonging to the logo were excluded.

3.3 Results and discussion

3.3.1 Elegance metrics

The unsupervised elegance metrics for Batches A-E are shown in Figure 3.4. Only five out of eight batches are presented for graphical reasons (conclusions drawn for batches A-E hold true also for batches F-H).

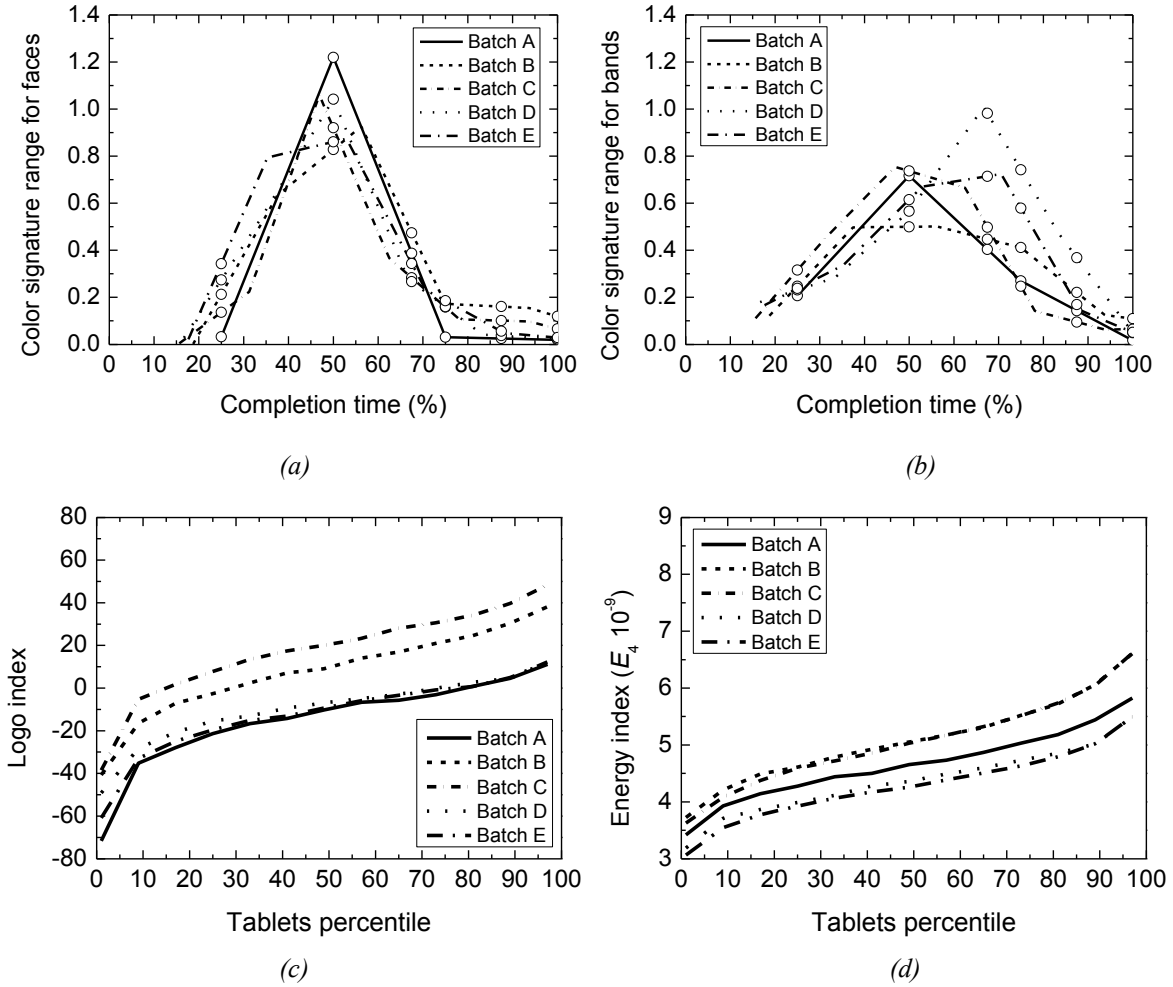


Figure 3.4. *Elegance metrics for batches A-E. (a), (b) color signature ranges for (a) faces and (b) bands. (c), (d) erosion indices: (c) logo and (d) energy.*

Figure 3.4 highlights that bands require longer coating time than faces, because the peaks of the distributions are shifted to higher fractions of the completion time. In fact, although the end-points of the distributions were found to be the same (see Figure 3.1), the color signature range of faces reaches its lowest value at around 75% of the total batch time, whereas for bands it evolves until the end of the batches. Differences between the coating evolution for tablets faces and bands have been reported by other authors using a camera-equipped microscope (Wilson and Crossman, 1997), near-infrared (Pérez-Ramoz *et al*., 2005) and Raman (Romero-Torres *et al*., 2006) spectroscopy, and discrete element method simulations (Freireich *et al*., 2011; Ketterhagen, 2011). These differences have been mainly ascribed to the exposed surface area during the coating operation, as well as to the way the tablets align themselves as they tumble and cascade in the pan (Ketterhagen, 2011).

The indices in Figure 3.4c-d highlight that batches B and C are characterized by a higher erosion level, consistently with the quality assessment carried out by a trained panel of experts. Furthermore, it can be observed that the indications obtained from the two indices are somewhat similar, though there some differences in the ranking of batches A, D and E with respect to the erosion inside and outside the commercial logo.

3.3.2 Regressing tablet elegance against coating conditions

The four elegance metrics presented in Figure 3.4 were used to build the quality matrix **Y** to be regressed against the [8×9] regressor matrix **X** of the coating operating conditions listed in Table 3.2. Namely, **Y** was defined as $\mathbf{Y} = [\mathbf{Y}_{faces}\ \mathbf{Y}_{bands}\ \mathbf{Y}_{logo}\ \mathbf{Y}_{energy}]$, i.e. **Y** was obtained from the horizontal concatenation of the color signature range of faces ($\mathbf{Y}_{faces}$ [8×6]) and bands ($\mathbf{Y}_{bands}$ [8×6]), the erosion logo index ($\mathbf{Y}_{logo}$ [8×13]) and the energy index ($\mathbf{Y}_{energy}$ [8×13]). Prior to the concatenation into **Y**, however, the matrices of the four metrics were further simplified:

- $\mathbf{Y}_{logo}$ and $\mathbf{Y}_{energy}$ were both reduced to [8×3] matrices considering the high correlation among the variables (verified using a PCA model); for each distribution, only the first, mid and last points were retained;
- $\mathbf{Y}_{faces}$ and $\mathbf{Y}_{bands}$ were reduced to a [8×2] matrix and a [8×1] matrix respectively, retaining only the final points of each distributions; in fact, the appropriate color uniformity needs to be ensured only at the end of the coating process.

The diagnostics of the 4-LV PLS model relating **X** and **Y** are given in Table 3.3 in terms of coefficient of determination per component (R^2) and cumulated (R^2_{CUM}) for both the predictor and the response matrices. It should be noted that while 2 or 3 LVs were probably sufficient to explain the tablet elegance, the fourth LV was included in the model since it greatly contributed to the description of the **X** space. The importance of modeling the **X** space has been discussed by Tomba *et al*. (2012) in the context of PLS model

inversion. By inverting the PLS model that relates the processing conditions **X** to the tablet quality **Y**, a set $\mathbf{x}^{NEW}$ of processing conditions can be determined that lead to a desired tablet quality $\mathbf{y}^{DES}$. Hence, model inversion can be used to provide model-based recommendations on how to operate the coating process.

Table 3.3. *PLS model diagnostics*.

LV	R^2**X**	R^2_{CUM}**X**	R^2**Y**	R^2_{CUM}**Y**
1	34.8	34.8	49.1	49.1
2	29.2	64.0	18.2	67.3
3	16.3	80.3	11.1	78.4
4	14.2	94.5	3.4	81.8
5	4.2	98.7	6.7	88.5

The PLS model scores (**t**) and loadings (**w***, **q**) for the first two latent variables and the modified VIP index are shown in Figure 3.5.

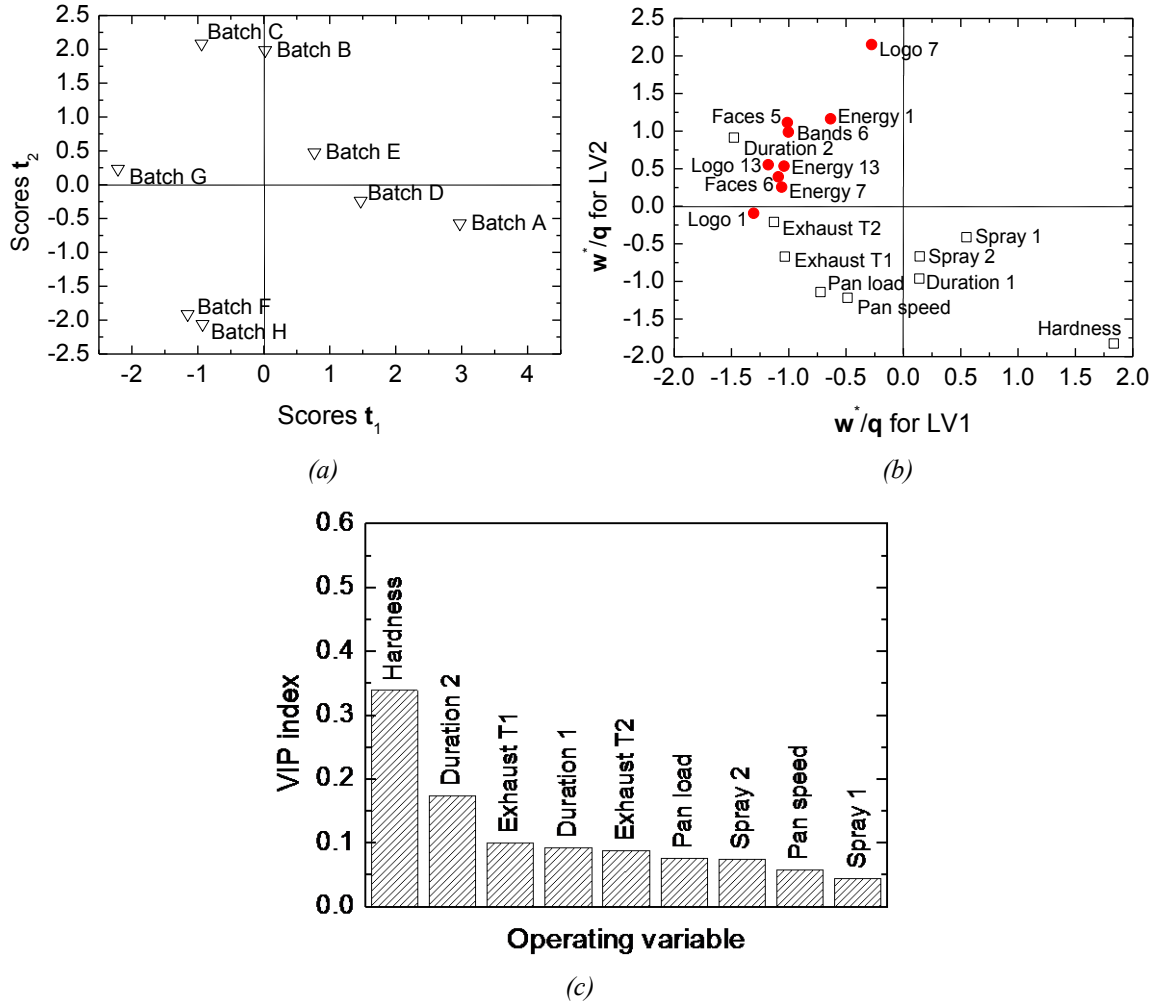


Figure 3.5. *PLS model (a) scores, (b) loading biplot (W*/Q overlay) and (c) modified VIP index.*

In the score plot (Figure 3.5a), the eight batches of the calibration set are separated along the first latent variable. From the right to the left, one can find: batches A, D and E with positive $\mathbf{t}_1$ values; batches B and C (recall that they were characterized by unsatisfactory surface finish) in the center, together with batches F and H; batch G (whose elegance metrics are given in Figure 3.6 and exhibited the worst surface quality) is relegated at negative $\mathbf{t}_1$ values. Hence, Figure 3.5a can be used as an unsupervised elegance monitoring chart: when a new batch is available, it can be projected onto the PLS model and, according to the position of its scores, either accepted or rejected.

The $\mathbf{W}^*/\mathbf{Q}$ loadings plot (Figure 3.5b) clarifies the main driving forces of the process and how they relate to the elegance metrics of the product, particularly surface erosion. Analysis of the plot leads to conclude that higher hardness leads to less erosion, and that the longer the coating process (particularly Duration 2), the more erosion the tablets will exhibit. Additionally, there is a positive relation between the pan load and pan speed and the appearance of erosion in the tablets. In fact, poor surface finish batches are located at negative values of the first latent variable, i.e. the same direction of the pan speed and pan load in the loading plot. Hence, the higher the pan speed/load, the worse the surface finish. These diagnostics lead us to conclude that the erosion in the tablet surface occurs mainly as a consequence of the physical collisions among tablets and with the baffles of the pan (and under the weight of the other tablets). However, other additional mechanism, such as the effect of water activating the disintegrant in the formulation as an effect of a too long residence time of the liquid water film on the tablet surface, should not be excluded. The loadings corresponding to the spray rate, in fact, are not negligible and, combined with the positive correlation between the duration of the coating operation and erosion, support the hypothesis that also water played a role in the surface erosion, though secondary with respect to the physical collisions.

The VIP index (Figure 3.5c) agrees with the abovementioned conclusions about the mechanisms leading to surface erosion. Tablet hardness and the duration of the coating process, in fact, were found to be the most important variables of the model, while the spray rates were deemed as less important.

3.3.3 Validating the elegance metrics

The unsupervised elegance metrics were validated using the images of the tablets of batch I. According to the panel test, the elegance of this batch was judged to be insufficient (excessive humidity during the coating process caused a high erosion level). Hence, the erosion metrics were expected to exhibit values at best similar to those of batches B-C (Figure 3.5), or possibly to those of batch G. This can be clearly seen in Figure 3.6 that compares the logo and energy indices of batches G and I.

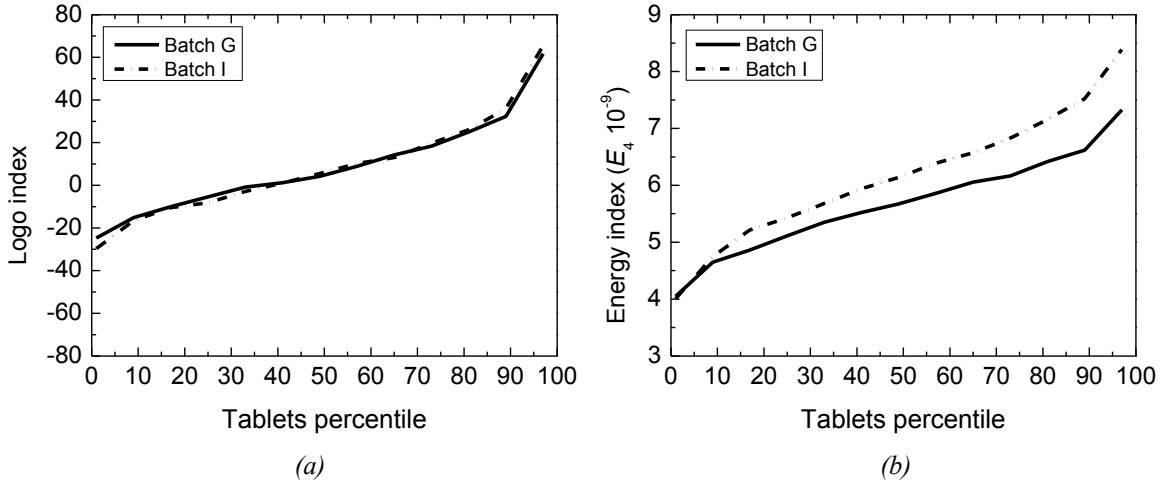


Figure 3.6. *(a) Logo and (b) energy indices for batches G and I.*

It was found that color uniformity of faces and bands for batch I was similar to that of the other batches. However, the developed erosion metrics highlighted the significant surface erosion characterizing the tablets. The inside-logo metric, in fact, is almost the same of the one characterizing batch G (Figure 3.6a), whereas the outside-logo metric (Figure 3.6b) is even worse (recall the position of batch G in the PLS model space; Figure 3.5a).

3.3.4 Surface defects detection: an alternative algorithm

This Section presents a supervised surface defects detection algorithm and metric that is intended to objectively mimic the elegance assessment exercise traditionally performed by an expert panel by analyzing, for each tablet imaged, the area of each defect. Though it was not explicitly considered, it can be of practical interest for the readers.

The algorithm consists of a sequence of morphological operations on the images, namely opening and closure, applied on the images preprocessed and cropped as described previously for surface erosion. Morphological operations are typically used to add or remove pixels from the boundaries of the objects within an image; for more details, the reader is referred to the work by Gonzales *et al*. (2009). The algorithm comprises the following steps, whose effect is shown in Figure 3.7 on a sample image:

1. evaluation of the negative of the image and masking (i.e., removal) of the commercial logo (see Figure 3.7a-b);
2. thresholding of the image, selecting only the pixels having light intensity values above a given threshold (see Figure 3.7c);
3. conversion from grayscale to black and white image (see Figure 3.7d);
4. morphological background opening (see Figure 3.7e);
5. morphological image closure and filling of the holes (see Figure 3.7f);

6. evaluation of the area of each highlighted defect, generating for each tablet a defect area distribution.

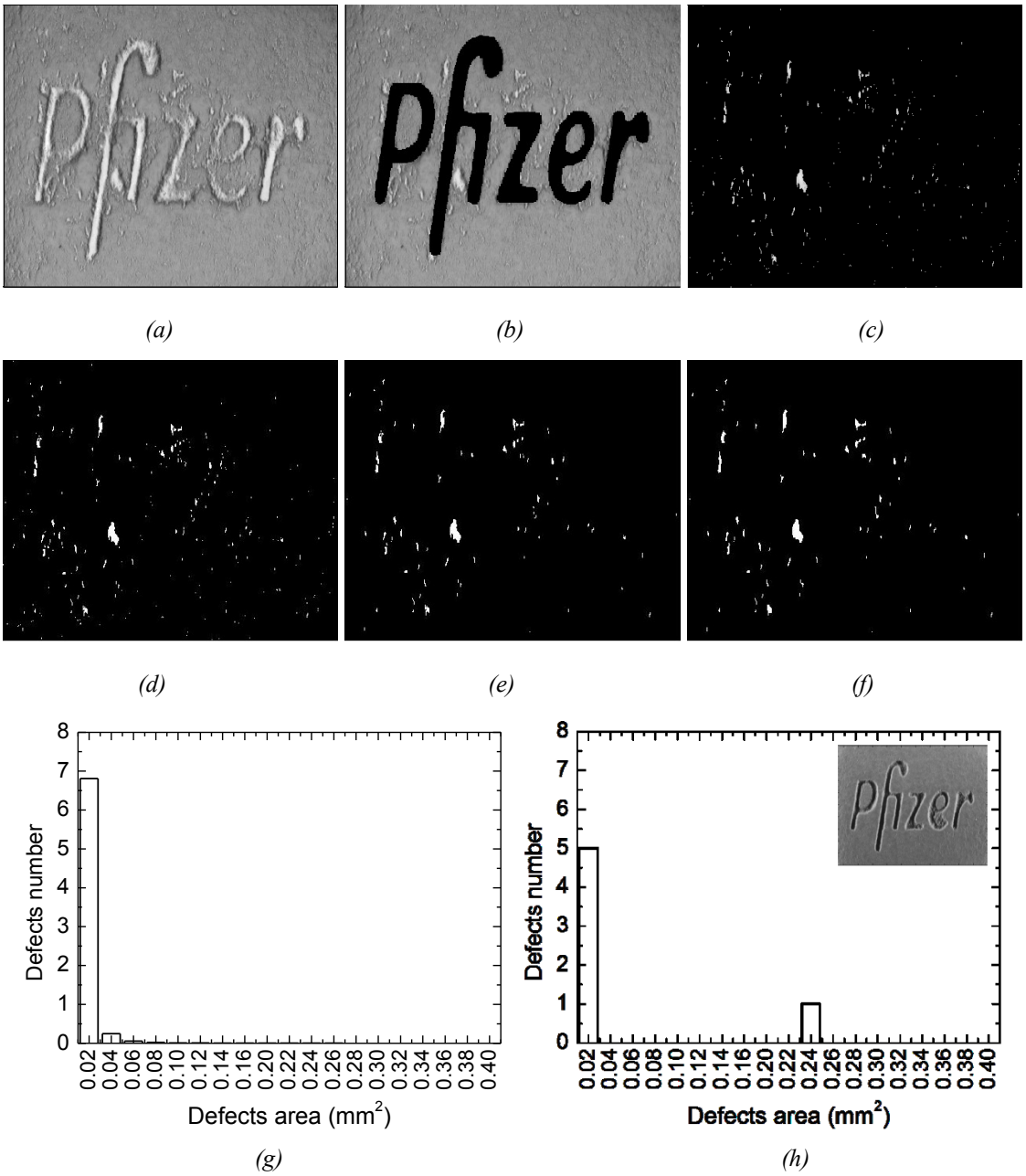


Figure 3.7. *Supervised defect detection algorithm. Image (a) negative, (b) after removal of the commercial logo, (c) after thresholding (d), after the conversion from grayscale to binary, (e) after the background opening and (f) after the closure and holes filling. (g)-(h) surface defects area distributions for batch A: (g) average and (h) tablet with the greatest defect.*

For steps 1-6 to be completed, the threshold for the grayscale light intensity distribution and the opening and closure operations need to be tuned from the user. In Figure 3.7, the threshold was set equal to 175, whereas the morphological operations were intended to

leave in the image only defects with an area of at least 50 pixels (i.e. 3.2 10^{-3} mm^2 according to the resolution of the images).
The output of the algorithm can be given in the form of the average defects area distribution per batch (Figure 3.7g) or per tablet (Figure 3.7h). Traditionally, the defects area is evaluated by visual inspection from a trained panel of experts, and results are strongly dependent on human judgment and therefore may be affected by reproducibility issues. Despite this strong limitation, one of the typical acceptance criteria for a batch is the number of tablets exhibiting defects with an area greater than an assigned value: if this number exceeds a given threshold, the entire batch is rejected. Hence, the algorithm 1-6 is important to objectively quantify the visual quality of the tablets obtained from a batch.

3.4 Conclusions

The use of multivariate image (MIA) and texture analysis (MWTA) was proposed in this Chapter to characterize film-coated tablets elegance, which is traditionally assessed by a trained panel of experts. Four unsupervised metrics have been developed and validated using images of more than 7000 tablets coated in nine different pilot-scale batches. Two metrics were intended to quantify the coating uniformity of both tablets faces and bands, while the other two the surface erosion both inside and outside the commercial logo embossed on the tablets.
A PLS model was used to regress the elegance metrics against the coating operating conditions in order to investigate the main driving forces of the process. Particularly, the analysis of the model loadings successfully allowed to investigate the mechanism leading to tablets erosion, concluding that the erosion in the tablet surface could be mainly ascribed to the physical collisions among tablets and with the baffles occurring inside the coating pan, though a water-related effect due to a too long residence time was also observed. Additionally, batches are shown to align according to the surface roughness of the tablets in the model space, which can be effectively used as a monitoring chart for the overall tablets elegance.
This case study demonstrates the effectiveness of using of simple imaging devices (RGB cameras) with powerful mathematics in providing a tool that replaces human perception with a quantitative measure useful for process development, troubleshooting and continuous quality assurance efforts.

Chapter 4

Standardization of machine vision systems for product quality assessment*

In the last decades, a growing number of image analysis applications in the process and food industries have been reported. This is because artificial vision systems can return a quick, accurate and objective indication of the quality of the manufactured end product. However, reproducibility of the image analysis results is ensured only as long as the conditions, under which the images used for the calibration of the quality assessment model were collected, do not change during normal operation of the manufacturing process. These conditions include the status of the artificial vision illuminating system and of the camera sensor. In this Chapter, the issues related to the aging or failure of the illuminating system and to the replacement of the camera are addressed.

The Chapter is organized as follows. In Section 4.1, the problem is presented with reference to the state of the art. In Section 4.2, the techniques used are described in details. Results are discussed in Section 4.3 considering a pharmaceutical engineering case study. In Section 4.4, some general conclusions are drawn about the standardization of machine vision systems.

4.1 Introduction

Oftentimes, the quality of a product or a process can be directly related to some visual features, like color or surface texture. Machine vision systems are based on image analysis and can return an objective and reproducible measurement of these features. Therefore, machine vision systems are suited for those applications where the quality assessment exercise is subject to human judgment (e.g., from a panel of experts). Furthermore, they can reduce the number of expensive and time-consuming laboratory analysis required for quality assessment, leading to related cost reductions.

* Ottavian, M., M. Barolo and S. García-Muñoz (2013). Maintenance of Machine Vision Systems for Product Quality Assessment. Part I: Addressing Changes in Lighting Conditions. *Ind. Eng. Chem. Res*., **52**, 12309-12318.

Ottavian, M., M. Barolo and S. García-Muñoz. Maintenance of Machine Vision Systems for Product Quality Assessment. Part II: Addressing Camera Replacement. *Ind. Eng. Chem. Res.*, in press. DOI: 10.1021/ie402910z.

The reported applications of image analysis for product/process quality assessment range over a broad domain of industrial sectors (Liu, 2004). For example, image analysis has been used to characterize flames from industrial boilers (Yu and MacGregor, 2004) and flares (Castiñeira *et al.*, 2012), fiber diameter distribution of polymeric membranes (Tomba *et al.*, 2010), erosion (García-Muñoz and Carmody, 2010) and color uniformity of tablets (García-Muñoz and Gierer, 2010), nucleation and crystals in a crystallization process (Simon *et al.*, 2010 and 2012), lumber grades (Bharati and MacGregor, 2003), steel (Bharati *et al.*, 2004) and paper quality (Reis and Bauer, 2009), foodstuff defects (Brosnan and Sun, 2004; Prats-Montalbán and Ferrer, 2007; Pereira *et al*., 2009), seasoning content of snack food (Yu and MacGregor, 2003), alumina content of anode cover materials (Tessier *et al*., 2008) and run-of-mine ore composition (Tessier *et al*., 2007). Two reviews on this topic (which the reader is referred to for additional examples) have been published recently (Prats-Montalbán *et al*., 2011; Duchesne *et al.*, 2012). Machine vision applications outside the process industry are innumerous (Waskewitz, 2007; Mitra and Acharya, 2007; Zhang and Gao, 2009), and focus mainly on problems of object recognition and motion tracking (i.e. pattern recognition).

Machine vision systems are made of hardware components (e.g. the illuminating system and the camera) and software components (the underlying image-based model for quality assessment). Therefore, likewise all measurement systems, also machine vision ones need to be maintained. In fact, if not promptly detected and corrected, failures (e.g. of the illuminating system or of camera sensors) can dramatically affect results reproducibility (Waskewitz, 2007), leading to a wrong characterization of the product quality.

Data-driven modeling (such as image-based modeling) usually requires a significant data collection effort to calibrate the quality assessment model, and the resulting model itself can be regarded as "valid" only if the conditions, under which the calibration data were collected, do not change during normal use of the model. Therefore, to assess whether the underlying model is valid or not, the state and the performance of a machine vision system should be monitored as well. If a component failure is detected, either the faulty component has to be replaced, or the model has to be adapted to the new imaging conditions until enough data is collected to perform a recalibration of the monitoring model. Though the former solution is easier, the latter is preferable, since there are no guarantees that the new component (though identical to the previous one, assuming it is still available on the market) will have the same behavior as that used during the collection of the calibration images. Furthermore, a well performed model adaptation will provide a shorter downtime of the system until recalibration is performed.

In this Chapter, the issue of a failure of the illuminating system is addressed first. Namely, after presenting a way to monitor its the status, two alternative strategies to deal with changes in the lighting conditions are defined. It should be noted that some techniques

exist to cope with changes in the illuminating conditions, but they usually refer to pattern recognition problems (Sigal *et al.*, 2004; Tu, 2009; Jahr, 2007), which address issues that are quite different from those encountered in product quality assessment problems. Although monitoring the light conditions is considered of utmost importance in machine vision systems (Jahr, 2007), how to achieve this in practice and how to adjust the system accordingly is seldom (if ever) discussed.

Then, the problem of transferring a quality assessment model between two cameras (from camera A to camera B) is discussed. The problem is addressed as a technology transfer problem, where "technology" includes both hardware components (the camera) and software components (the quality assessment model). Clearly, the issue under investigation does not refer only to the maintenance of an *existing* machine vision system. In fact, the technology transfer scenario applies not only when a camera B is bought to replace camera A (e.g. after a failure), but also when, based on the results obtained with a given machine vision system centered on camera A, camera B is installed in a different plant. Note that the transfer can be carried out between two cameras of the same type or of different types. In the latter case, camera B can be either of a better quality than camera A (e.g. when replacing an older camera after a failure or to improve the technology of the system) or of a lower quality (e.g. when moving the quality assessment system from a research environment to on-line manufacturing, where low resolution cameras may be employed). In all cases, the technology transfer is intended to avoid periods of downtime until model recalibration is performed. From a very general point of view, the issues addressed represent research areas pertaining also to the machine vision community (Porikli, 2003; Ilie and Welch, 2005). Within this community, several strategies have been proposed to cope with a somewhat similar problem, namely ensuring the color consistency of images collected from multiple cameras. However, the applications onto which the strategies have been tested are mainly related to pattern recognition problems. To the authors' knowledge, the technology transfer problem addressed here has never been discussed in the open literature so far, despite its importance in ensuring the reproducibility of the results of a machine vision system subject to possible camera failures or replicated installations.

4.2 Materials and methods

4.2.1 Imaging station and available data

The in-house developed imaging station used in this study was equipped with three digital cameras: two single-lens reflex Canon EOS 40D camera (10.1 megapixel resolution; camera A1 and camera A2 in the following) with a Canon EF-S 60 mm f/2.8 USM Macro lens and single-lens reflex Fujifilm S9100 (9 megapixel resolution; camera B in the

following). LED lights illuminated the subject (two racks mounting 64 LEDs each), and the system was isolated from the outside and operated through a computer. The user was allowed to set camera elevation, lights elevation (hence changing the angle of incidence of the light on the subject surface) and camera settings. Namely, shutter speed t (in seconds), lens aperture N (in f-stop) and ISO sensibility S (in ISO number) were used to ensure proper exposure. Unless differently specified, the settings were set as [t, N, S] = [1/13, f/5.6, 125] for all three cameras.

For the light monitoring exercise, images of a set of four color standards (red, green, blue and yellow; Edmund Optics T56-079, USA) were used. The light correction strategies, instead, were tested on images of film coated tablets with different percentage of coating material applied (0.5, 0.75 and 1% of coating material relative to the core weight of the tablet), with the aim of analyzing the color progression along the coating process (García-Muñoz and Gierer, 2010). All images were collected with camera A1.

For all camera-to-camera transfer exercises, two datasets were collected. The first dataset consisted of 20 images of two subjects of four different colors, and was used to draw some general conclusions on the feasibility of the transfer. The subjects were either the color standards or pharmaceutical tablets (red, green, pink and white). The second dataset consisted of the images of film coated tablets with different percentage of coating material applied previously discussed. This dataset was used to further test the transfer strategies proposed Sample images of the two datasets are shown in Figure 4.1.

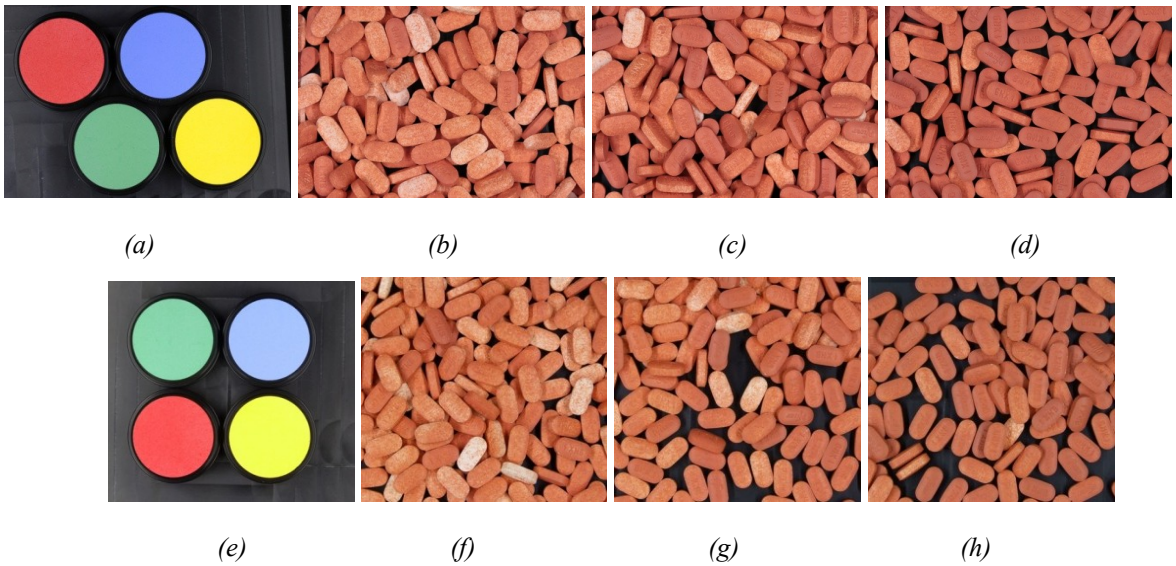


Figure 4.1. *Example of the available datasets: (a) color standards and (b)-(d) film-coated tablets with different percentage of coating applied (0.5%, 0.75% and 1%, respectively) collected with camera A1. (e)-(h) same images collected with the camera B.*

4.2.2 On the scaling of the scores in MIA modeling

The scores obtained from the PCA decomposition of an image are recommended to be scaled within the range [0-255] (Yu and MacGregor, 2004):

$$\mathbf{t}_i = \text{round}\left[255 \cdot \frac{\mathbf{t}_i - \min(\mathbf{t}_i)}{\max(\mathbf{t}_i) - \min(\mathbf{t}_i)}\right] \quad . \tag{4.1}$$

Once max(**T**) = [max(**t**$_1$); max(**t**$_2$)] and min(**T**) = [min(**t**$_1$); min(**t**$_2$)] are evaluated on the calibration images, they are used also to scale the scores of the validation images, i.e. the images projected onto the available MIA model. However, when the image projected has been collected under different lighting conditions (e.g. dimmed light) or whit a different camera, a problem arises, since for this image (4.1) may return meaningless score values, i.e. scores outside the [0-255] range. To overcome this problem, a novel scaling method for the scores is proposed. This novel "color scaling" will be used throughout the remaining of the Chapter.

The rationale of the color scaling is fairly simple: instead of using the values of the observed scores for the definition of max(**T**) and min(**T**), the scores obtained from the projection onto the MIA model of selected pure color pixel (chosen among white [255; 255; 255], black [0, 0; 0], red [255; 0; 0], green [0; 255; 0] and blue [0; 0; 255]) are used. The loadings of the MIA model drive the selection of the colors to be used. Figure 4.2 presents the loadings of a MIA model built on the image of Figure 4.1c, showing that **p**$_1$ extracts the information related to the average color of the image (same loading signs for the three spectral channels, with the red one slightly prevailing the others), whereas **p**$_2$ highlights the contrast between the red channel and the green and blue ones. Accordingly, score **t**$_1$ was scaled between the maximum and minimum scores obtained from the projections onto the MIA model of a pure white and a pure black pixel (as all colors lay between these two), whereas score **t**$_2$ was scaled between the maximum and minimum scores obtained from the projections onto the MIA model of a pure red and a pure blue pixel (as these channels exhibited the greatest contrast).

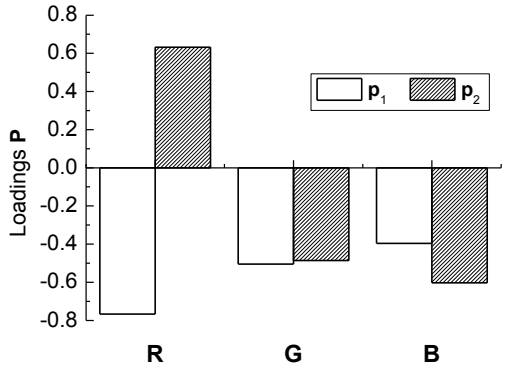


Figure 4.2. *MIA model loadings for the red (R), green (G) and blue (B) channels*.

A comparison between score spaces obtained with and without the proposed color-scaling is shown in Figure 4.3.

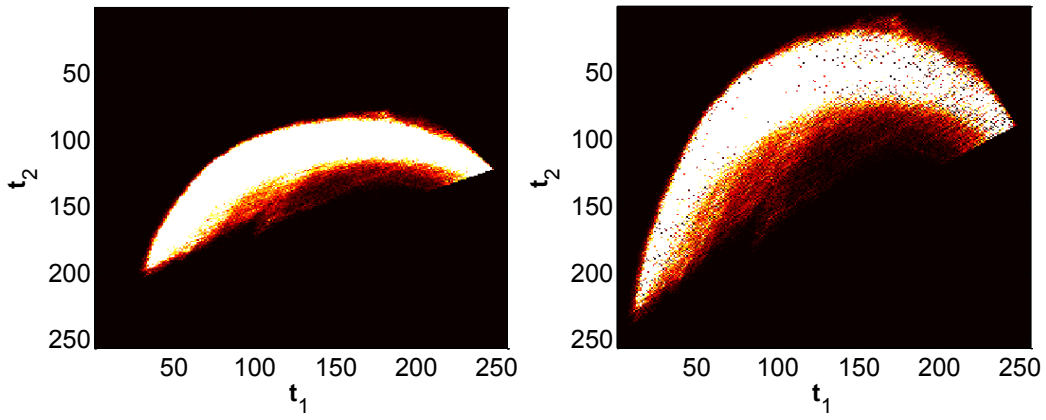


Figure 4.3. *2D histogram-scatter plot for Figure 1c obtained with the score scaling given by Equation (5) (a) with and (b) without the application of the color scaling.*

The comparison between Figure 4.3a and 4.3b reveals that the use of the color scaling could cause a reduction in the resolution of the score space if the same binning resolution is used in producing the 2D histograms, since the cloud of points is confined in a smaller region within the [0-255] range. Though not reported here for the sake of conciseness, it was found that the results presented here are not dependent on this change in the density distribution and that increasing the resolution of the binning does not improve the results (a [256×256] grid was used in producing the 2D histograms shown here).

4.2.3 Alignment of the score spaces of different images

The use of the dynamic time warping (DTW) was suggested by Kassidas and coworkers (Kassidas *et al.*, 1998) for the synchronization (alignment) of batch trajectories through a series of compressions and dilations. DTW operates by solving an optimization problem to evaluate the warping path, which is the sequence of points that have to be matched for two trajectories to be aligned. For more details on the DTW algorithm, the reader is referred to the paper by Kassidas *et al*. (1998).

The use of the DTW is proposed in this Chapter to match the 2D histogram-scatter plot of two images collected under different light conditions and to handle the technology transfer problem. The results of the DTW implementation are shown in Figure 4.4.

The score spaces of Figure 4.4a and 4.4b are generated by projecting the images of Figure 4.1g and 4.1c onto a MIA model calibrated on images collected with camera B (for more details on the analysis of the film-coated tablet images, see Section 4.3.2.1). In order to match the two score spaces, the DTW algorithm was applied in two steps, i.e. first by rows and then by columns. In the first step (Figure 4.4c), each row of the score space in Figure

4.4b was aligned against the corresponding row of the score space in Figure 4.4a. In the second step (Figure 4.4d), the synchronization was carried out by columns, considering the score spaces of Figures 4.4c and 4.4a. The cumulative results obtained for the score $\mathbf{t}_1$ are shown in Figures 4.4e and 4.4f. Further alignment of the score space of Figure 4.4d with the original score space of Figure 4.4a by re-application of the DTW algorithm was found not to improve the results. If necessary, the optimal number of cycles of DTW applications (by rows and by columns) can be determined by minimizing the difference between the $\mathbf{t}_1$ and $\mathbf{t}_2$ score spaces (Figures 4.4e and 4.4f) after the alignment.

In MIA applications, the shape of the clusters in the 2D histogram-scatter plot usually resemble those shown in this Chapter. In such situations, the DTW correction is expected to always return a good match. Nevertheless, the capability of the algorithm to work properly in all circumstances cannot be guaranteed. For example, for large differences between the resolution of the camera for which the quality assessment model is calibrated and the resolution of the new camera (i.e., the peaks in Figure 4.4e are very different), the DTW might fail, since stretching and compressing might not be enough to ensure a proper synchronization.

4.2.4 Camera transfer by color consistency matching

The machine vision literature has thoroughly studied the problem of ensuring color consistency across multiple cameras, mainly for pattern recognition (Khan and Shah, 2003; Javed *et al.*, 2003; Nummiaro *et al.*, 2003; Hu *et al.*, 2006). One of the most common applications encountered is the surveillance of a given area through multiple cameras, which involves the problem of recognizing as identical an object (e.g., a person) simultaneously imaged from different cameras. In general, since the segmentation of the object (i.e. its identification and separation from the background) is obtained by thresholding the color intensity histograms, color differences among cameras can generate mistakes, even if the cameras are of the same type. Similarly, in the technology transfer problem considered in this study, if the color structures of two images of the same product are different because the images are taken with different cameras, then the output of the quality assessment model would be (erroneously) different.

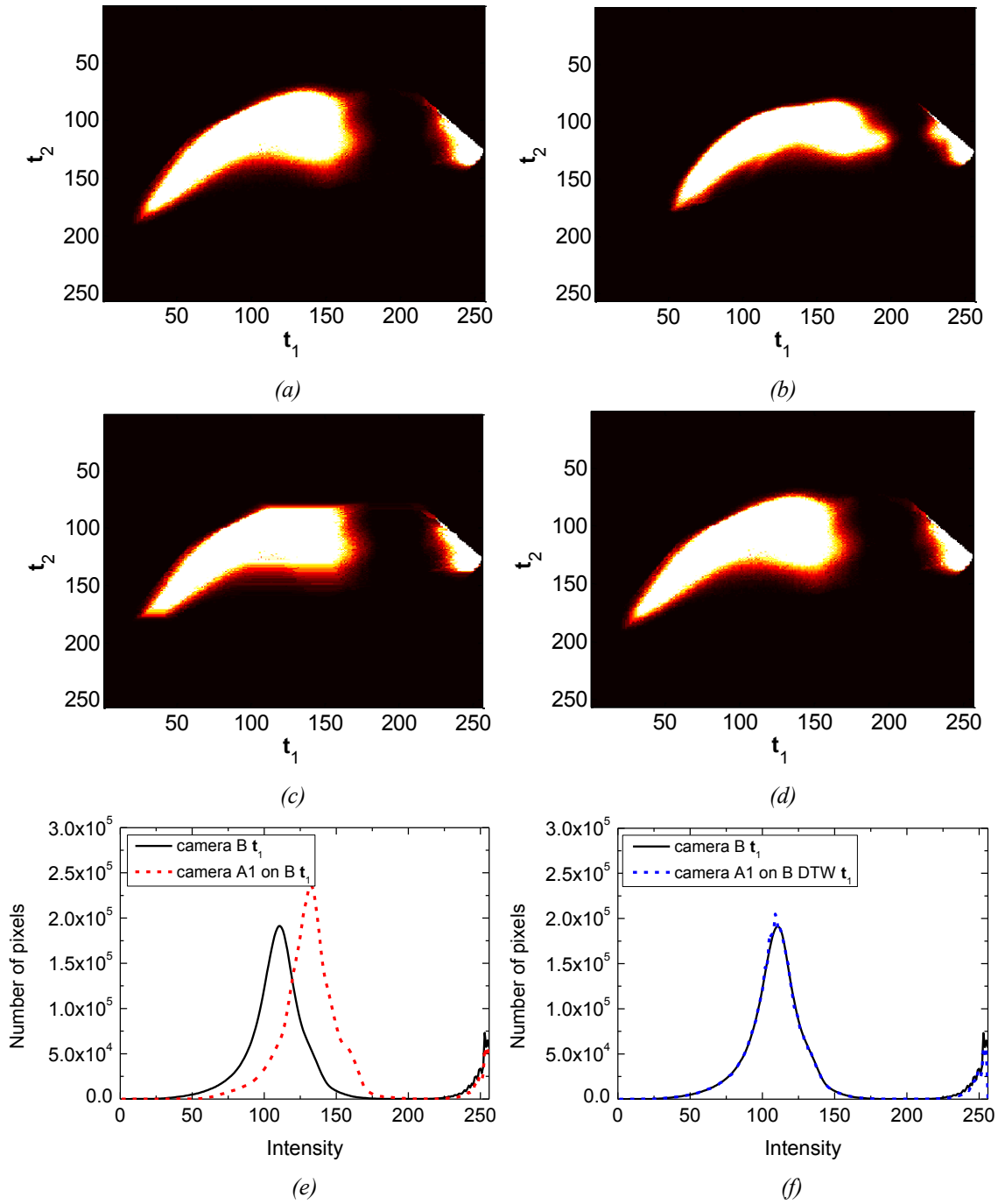


Figure 4.4. *Implementation of the DTW algorithm for the alignment of the score spaces of two images of the same tablets taken with different cameras. Score spaces of the image taken with camera B (a) and camera A1 (b) projected onto an MIA model defined on the camera B image. DTW application to align (b) on (a): synchronization by rows (c) and by columns (d). $\boldsymbol{t}_1$ scores before (e) and after (f) the DTW alignment.*

Given the similarity of the two above problems, one popular technique proposed in the machine vision literature to ensure color consistency was tested also in the context of the transfer problem addressed in this study. Namely, the approach presented by Ilie and Welch (2005) was considered. This approach aims at matching the RGB spaces of two

cameras through a transformation between the R, G and B values of selected colors, and it will be referred as "RGB transfer".

Given the two [$K \times J$] matrices $\mathbf{C}_A$ and $\mathbf{C}_B$, representing J features extracted from the RGB values of K different colors for camera A and B (respectively), the matching of the two color spaces is obtained via the [$J \times J$] transformation matrix $\mathbf{S}$ defined as

$$\mathbf{S} = (\mathbf{C}_A^T \mathbf{C}_A)^{-1} \mathbf{C}_A^T \mathbf{C}_B \quad , \tag{4.2}$$

assuming that camera B has to be aligned against camera A. Two alternatives were explored (Ilie and Welch, 2005):

- linear RGB transfer ("RGB 1" in the following), in which the matrix $\mathbf{C}$ was built considering for each of the K colors its R, G and B values, i.e.

$$\mathbf{C} = \begin{bmatrix} r_1 & g_1 & b_1 \\ r_2 & g_2 & b_2 \\ \vdots & & \vdots \\ r_K & g_K & b_K \end{bmatrix} \tag{4.3}$$

  being r_k, g_k and b_k respectively the R, G and B values for the k-th color;
- quadratic RGB transfer ("RGB 2" in the following), in which the matrix $\mathbf{C}$ was built considering for each of the K colors its R, G and B values as well as their squared values, i.e.

$$\mathbf{C} = \begin{bmatrix} r_1 & r_1^2 & g_1 & g_1^2 & b_1 & b_1^2 & 1 \\ r_2 & r_2^2 & g_2 & g_2^2 & b_2 & b_2^2 & 1 \\ \vdots & & & & & & \vdots \\ r_K & r_K^2 & g_K & g_K^2 & b_K & b_K^2 & 1 \end{bmatrix} \tag{4.4}$$

  where the last column allows for a translation.

The use of a quadratic model as in (4.4) has been suggested recently also by Vidal *et al*. (2013) to capture cameras variability in a different context (i.e., matching of the color spaces of a camera and a colorimeter).

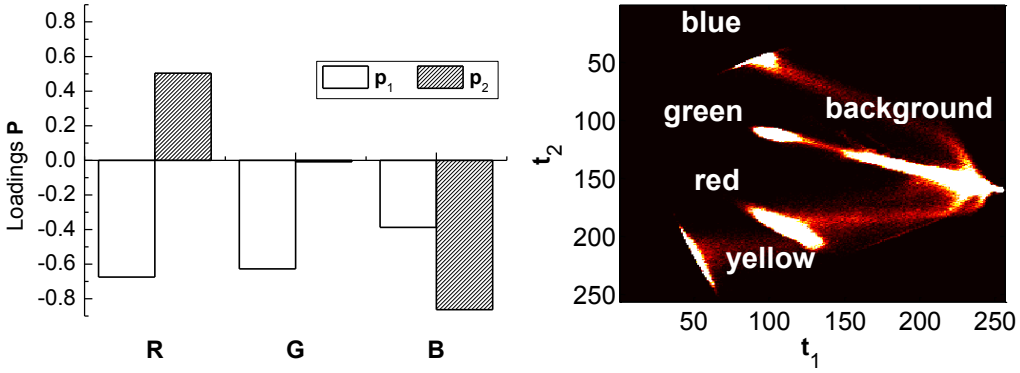
Once the transformation matrix $\mathbf{S}$ has been evaluated, it can be used to correct any new image collected.

4.3 Results and discussion

4.3.1 Monitoring the lighting conditions

The purpose of the light monitoring exercise is defining a sequence of operations to check whether the conditions of the illuminating system are changed with respect to those existing when the MIA model was calibrated. Alterations of the camera sensor can be detected as well.

Images of the color standards (Figure 4.1) collected under normal illuminating conditions (NIC) were used to calibrate the monitoring model. A total of 13 images were taken, with the standards randomly positioned within the image area. An MIA model was built on a couple of (randomly selected) images: its loadings, together with 2D histogram-scatter plot obtained from the projection of the image of Figure 4.1a on it, are shown in Figure 4.5.



(a) *(b)*

Figure 4.5. *MIA model for the color standards under NIC light. (a) Model loadings and (b) 2D histogram-scatter plot for the image of Figure 4.1a.*

The information given by the loadings is similar to that shown in Figure 4.2, i.e. $\mathbf{p}_1$ extracts the average color (which is typical of MIA models), whereas $\mathbf{p}_2$ reveals a contrast between the red and blue channel, with no contribution on the green one. The projection of an image of the four color standards onto the MIA model (Figure 4.5b) shows five clearly separated clusters (the four colors plus the background, as indicated in the figure). The features extracted for each of the 13 calibration images from the 2D histogram-scatter plot were the number of pixel (N_{pix}) and the scores variance-covariance ($\sigma^2_{t_1}, \sigma^2_{t_2}, \sigma_{t_1 t_2}$) of each cluster after the removal of the background, and were collected into a [1×16] $\boldsymbol{\gamma}$ vector:

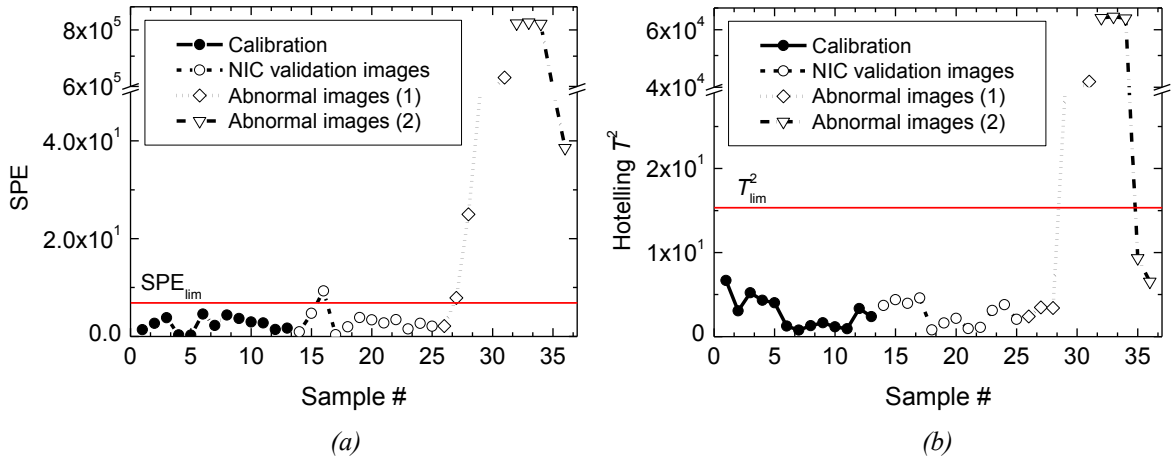
$$\boldsymbol{\gamma} = \left[\left(N_{pix}\ \sigma^2_{t_1}\ \sigma^2_{t_2}\ \sigma_{t_1 t_2} \right)_{\text{yellow}}; \left(N_{pix}\ \sigma^2_{t_1}\ \sigma^2_{t_2}\ \sigma_{t_1 t_2} \right)_{\text{red}}; \left(N_{pix}\ \sigma^2_{t_1}\ \sigma^2_{t_2}\ \sigma_{t_1 t_2} \right)_{\text{green}}; \left(N_{pix}\ \sigma^2_{t_1}\ \sigma^2_{t_2}\ \sigma_{t_1 t_2} \right)_{\text{blue}} \right] \quad (4.5)$$

The background was removed by using a predefined mask, while the masks for each color cluster were obtained by fitting the corresponding scores to a 2-PCs PCA model, with their 99% joint confidence interval representing the desired (elliptical) mask.

The monitoring model was eventually designed by fitting a PCA model on the matrix obtained by stacking on the top of each other the features vectors of the calibration images. Details on the model (R^2 and eigenvalues) are given in Table 4.1; the Hotelling T^2 and SPE monitoring charts with their 99% confidence limits (T^2_{lim} and SPE_{lim}) are shown in Figure 4.6. Note that the normality assumptions beyond the two confidence limits have been verified. The model was built on 3 PCs (adding the fourth PC did not change the results significantly). Both the Hotelling T^2 and the SPE are measures of how well a sample is described by the model, representing the distance from the center of the model hyperplane and the distance from the model hyperplane, respectively (Jackson, 1991).

Table 4.1. *Diagnostics of the PCA model built for light monitoring.*

PC #	Eigenvalues	$R^2\mathbf{X}_{CUM}$
1	6.1	38.3
2	4.9	68.7
3	2.3	83.3
4	1.2	90.7
5	0.92	96.4



(a) *(b)*

Figure 4.6. *(a) SPE and (b) Hotelling T^2 monitoring charts for the light monitoring exercise. 99% confidence limits on the two statistics are indicated. Please note that the scale is logarithmic after axis breaks.*

In addition to the calibration images, Figure 4.6 shows the results of the projection on the monitoring model of the features extracted from three different sets of validation images, which are described in details in Table 4.2. Figure 4.7 shows instead samples of the color standard images taken under different illuminating conditions.

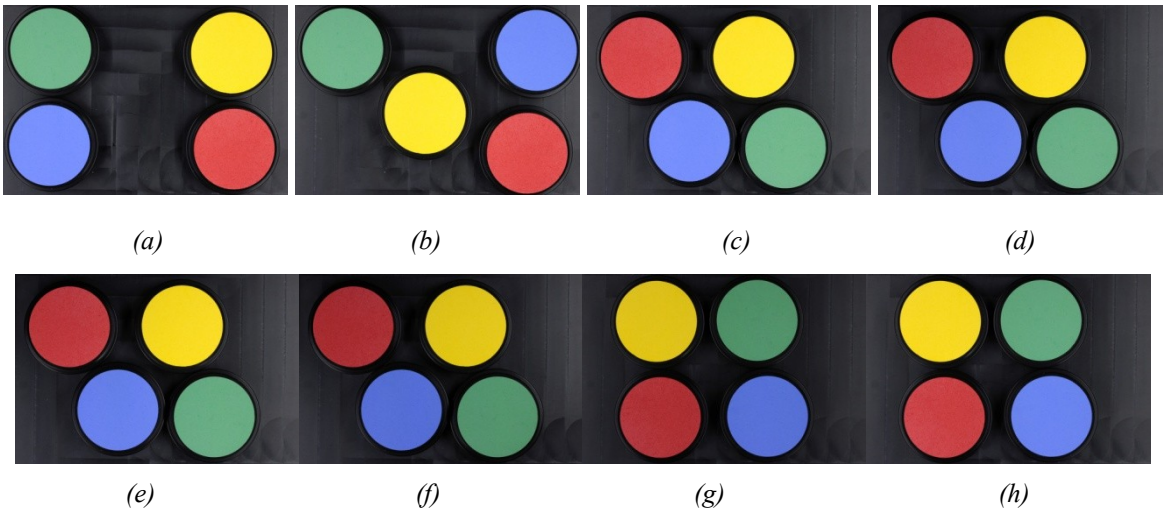


Figure 4.7. *Example of validation images collected (a)-(b) under NIC, (c)-(f) blocking part of the LEDs (respectively 6, 25, 50% on a rack and 50% on both racks) and (g)-(h) regulating light intensity in the machine vision system.*

The condition of the lighting system can be monitored by taking a test image of the color standards and carrying out the following operations:

1. projection of the image onto the available MIA model and evaluation of the scores $\mathbf{T}_{test}$
2. filtering of the background
3. evaluation of the feature vector $\boldsymbol{\gamma}_{test}$
4. projection of the feature vector $\boldsymbol{\gamma}_{test}$ onto the monitoring PCA model
5. evaluation of SPE_{test} and T^2_{test}, and their comparison with the confidence limits
6. definition of the status of the lighting system (NIC vs. altered) by comparing SPE_{test} and T^2_{test} with their (respective) limits.

Table 4.2. *Validation data used to test the performance of the light monitoring model.*

Validation data	Description
Set 1	12 NIC images
Set 2	6 images obtained by blocking part of the LEDs (3, 6, 12, 25 and 50% on a rack and 50% on both racks)
Set 3	5 images of uniformly dimmed light, obtained using the knob regulating lighting intensity in the machine vision system

Figure 4.6 shows the high sensitivity of the proposed monitoring model to changes in the lighting conditions. All images collected under altered light were correctly classified as non-NIC (SPE_{test} much higher than the confidence limits SPE_{lim}) with the exception of those obtained by blocking just two (out of 128) or four (i.e. 3% and 6%, respectively, cf. Figure 4.7c) of the LEDs (the former being within the confidence limits and the latter with

a value of SPE_{test} just above SPE_{lim}). 11 out of the twelve validation NIC images were correctly recognized as such. As for this latter result, it should be reminded that some false positives are indeed expected when monitoring the system status against some confidence limits (per their definition). As with any monitoring exercise, the limits could be adjusted so that an acceptable level of Type I and Type II errors are obtained.
Maintenance of the monitoring model (i.e. its update with new images of the color standards) can be made as an off-line exercise or through a periodic check-up, though no general indication can be given since the specificity of each machine vision installation needs to be considered.

4.3.2 Correcting for light changes

Once a change in the lighting conditions is detected through the proposed monitoring model, an action needs to be taken. Two alternative strategies are proposed to compensate for light alterations, either by changing the camera settings or by adapting the available model to the new lighting conditions.
The strategies for light correction are presented through a case study, which is shortly described in the next Section. The rationale of the two proposed strategies is then explained in detail.

4.3.2.1 Case study

Images of tablets at different percentages of coating were collected under NIC (20 images per percentage of coating, shuffling each time all tablets). The images were manipulated according to the covariance mask method (Yu and MacGregor, 2003) as proposed by García-Muñoz and Gierer (2010) on a similar case. Shortly: an MIA model was designed using a composite image obtained from the concatenation of several images of tablets at different coating levels, and the scores obtained from the projection of all the available images (after filtering out the background) were manipulated according to the covariance mask, which extracts a feature vector for each image (by grouping the pixels in the 2D histogram-scatter plot having similar correlation with the property of interest). The matrix built from this vectors is decomposed through a one-component PCA model, the score $\mathbf{t}_1$ representing an unbiased metric called color signature, whose evolution indicates the change in the tablets color (i.e. the change in the amount of coating applied). Namely, one value of the color signature was evaluated for each image, generating the open boxes of the box plot in Figure 4.8, where the variability shown at a specific coating level is related to the non homogeneous distribution of the coating solution among tablets.
In order to mimic a reduction (dimming) in the light intensity, half of the LEDs of the imaging station were blocked, thus producing an even light disturbance. New images were collected (20 images per percentage of coating applied) and projected onto the previously

calibrated MIA model, hence generating a new distribution of the color signature, which is shown in Figure 4.8 using hashed boxes.

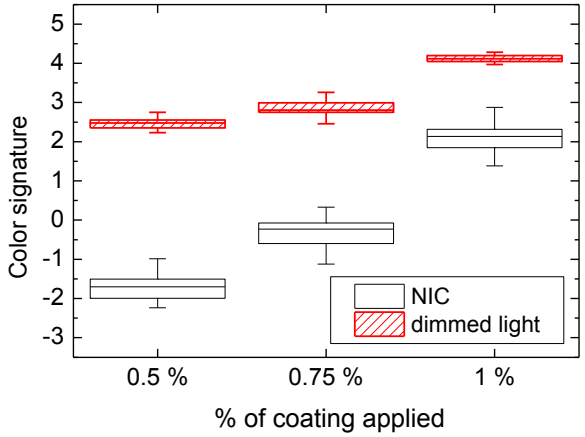


Figure 4.8. *Color signature evolution during the coating process obtained under NIC (black) and projecting the images collected under dimmed light onto the calibrated model (red).*

The reduction in light intensity causes a shift of the color signature towards higher values, and can result in an erroneous analysis of the coating process being monitored. The effect of the reduction can be clearly seen also in the score space. Figure 4.9 reports the 2D histogram-scatter plots for an image of tablets with 0.75% coating applied collected under dimmed light and, for an immediate comparison, the one obtained under NIC and shown in Figure 4.3a.

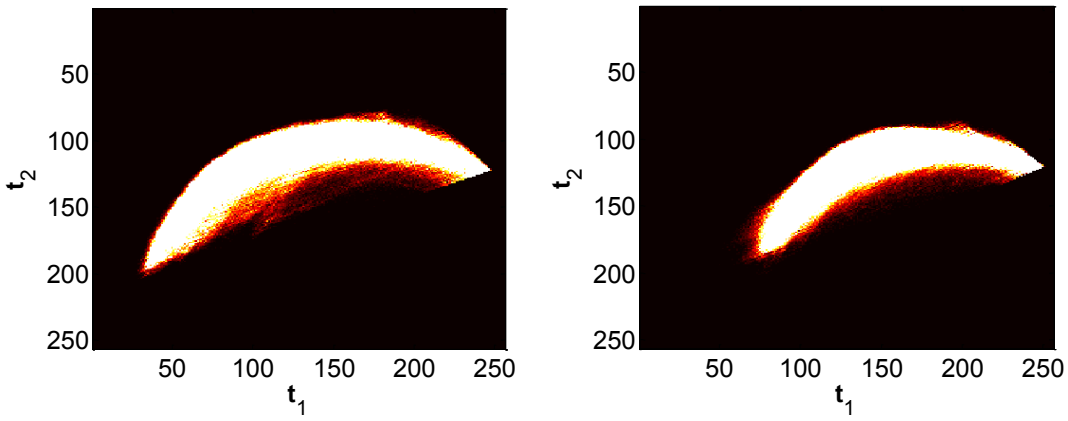


Figure 4.9. *2D histogram-scatter plot of an image of tablets (at 0.75% coating applied) collected under (a) NIC and (b) dimmed light.*

The score space under dimmed light appears shifted toward the background region (right side of the score space) and exhibits a somewhat compressed density distribution of the scores with respect to one under NIC. The shifting is due to the fact that the overall image is darker, and the changes in the density distribution are due (we believe) to the fact that

light is absorbed differently from different colors (Halliday *et al.*, 2012), thus causing a color-dependent shift toward pure black (no light).

The two alternative strategies that are proposed to correct the change in the lighting conditions will be discussed in detail in the following sections. Their rationale is as follows:

- Strategy 1: definition of a new set [t, N, S] of camera settings to obtain an exposure (H_v) of the subject equivalent to the one under which the model was calibrated. Note that the exposure (i.e. the amount of light allowed to hit each area unit of a photographic medium) is related to the settings through:

$$H_v = t \cdot \log_2\left[\frac{100N}{t^2 S}\right] \quad , \tag{4.6}$$

  and is expressed in [lux×s]);
- Strategy 2: definition of a transformation to match the score space obtained under changed light with the one obtained under NIC.

Details on the implementation of the two correction strategies are given in the flowchart of Figure 4.10, which describes the sequence of operations to carry out when a new quality assessment campaign is started (using an image-based model calibrated under a given lighting condition).

Figure 4.10 shows that both correction strategies require some new images to be collected under altered light conditions. As it will be clarified in the following Sections, only images of a single point of the coating progression (or, generally speaking, of a limited portion of the calibration samples) need to be taken. The mid-point (0.75% of coating applied) will be used to illustrate both correction strategies.

4.3.2.2 Strategy 1: correcting camera settings

Camera settings can be manipulated if the way they affect the images (or, equivalently, their score spaces, or the features extracted using the covariance mask method) is known. Therefore, in order to model the relationship between settings (inputs) and images (outputs), the system needs to be excited once the change in the illuminating conditions is detected (step 1a of Figure 4.10).

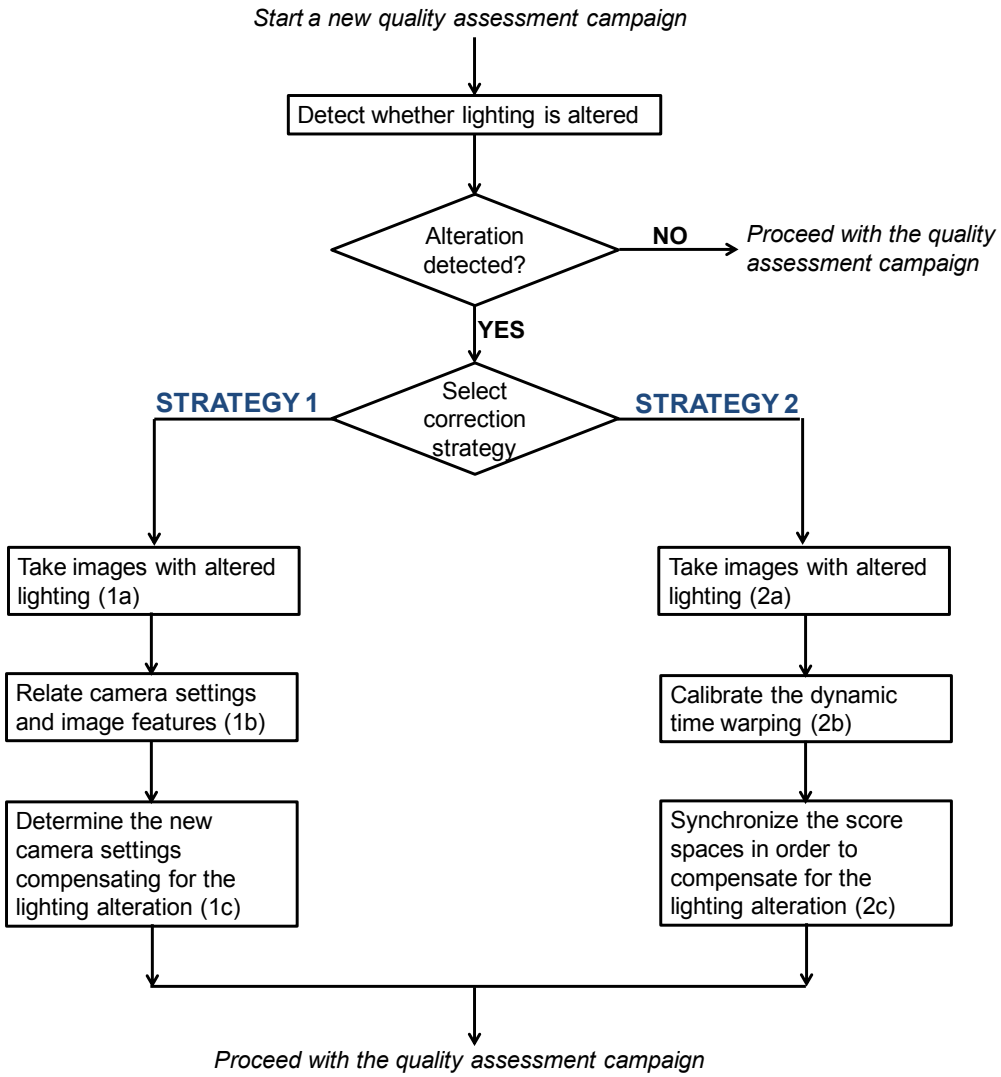


Figure 4.10. *Operations to perform when a new quality assessment campaign is started.*

The required excitation can be provided by changing the camera settings according to a fractional factorial design of experiment (DoE; Montgomery and Runger, 2010). Assuming that the three available degrees of freedom (i.e. the three camera settings) could distribute on three levels each, two different cases were considered (exploring, with respect to the values used in calibration, the ranges of ± ½ and ± 1 stop independently) with the assumption that only mild corrections are required. If a marked light change is detected, the range can be expanded, revising the DoE experiments accordingly. For each case, a total of nine experiments were designed (a D-optimal approach was used). The matrix **X** of the experiments, which will be used as regressor in the modeling step, is the same for both cases considered, and is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \quad . \tag{4.7}$$

$$\underbrace{\quad\quad}_{(t_-;\bar{t};t_+)}\underbrace{\quad\quad}_{(N_-;\bar{N};N_+)}\underbrace{\quad\quad}_{(S_-;\bar{S};S_+)}$$

As already stated, for each camera setting x, the values explored in this study were step variations (x_- and x_+) of ± ½ stop and ± 1 stop around the calibration value ($\bar{x}$). Namely, the values considered were t = (1/10, 1/13, 1/15), N = (f/5, f/5.6, f/6.3) for the ± ½ stop case, and t = (1/8, 1/13, 1/20), N = (f/4.5, f/5.6, f/7.1) for the ± 1 stop case, the values of S being in both cases S = (160, 125, 100). The columns in **X** reflect this order, with the 1's and 0's indicating the setting used or not (respectively) within a certain experiment.

For each experiment, 20 images of tablets at 0.75% coating (the final results, however, are not dependent on the point selected) were collected under dimmed light, features were extracted as described previously, and their median value per experiment was used to build the **Y** matrix (to be regressed against the camera settings **X**). A PLS model (step 1b of Figure 4.10) was calibrated to relate **X** and **Y**, block-scaling the matrix of the settings in order to give equal weights to t, N and S. In order to ensure a proper selection of the number of LVs (i.e., to avoid overfitting the data), the 20 images collected for each experiment were split into two groups of 10 images each: one group was used to calibrate the model, whereas the second one was used to validate its performance. For the PLS model using the ± ½ stop excitation data, the diagnostics (in terms of cumulated coefficient of determination in calibration, R^2, and validation, Q^2) are given in Table 4.3. Similar diagnostics were obtained for the PLS model using the ± 1 stop excitation data. Both the two final models were built on two LVs.

Table 4.3. *Diagnostics of the PLS model relating the settings of the camera and the features extracted from the images, built using the ± ½ stop excitation data.*

LV #	$R^2\mathbf{X}_{CUM}$	$R^2\mathbf{Y}_{CUM}$	$Q^2\mathbf{Y}_{CUM}$
1	14.0	84.1	86.7
2	38.2	90.5	91.9
3	47.3	97.8	95.1

The PLS models were eventually used to determine the camera settings to be used for the new light conditions (step 1c in Figure 4.10), by predicting the features vectors for all possible settings combination within the explored range of either ± ½ or ± 1 stop. The predicted features vector closest (i.e. with the lowest Euclidean distance) to the one obtained during calibration (Section 5.1) indicate the settings to be applied. For the case study under investigation, the new settings to be used were found to be [*t*, *N*, *S*] = [1/13, f/5, 160] for the ± ½ stop model, and [*t*, *N*, *S*] = [1/8, f/5.6, 125] for the ± 1 stop model.
Images of the entire coating progression were collected (20 images per percentage of coating) to test the new settings, and projected onto the MIA model to extract the color signature (once the new settings have been determined, images of the entire coating distribution collected under altered light could be used to test the proposed strategy). The results obtained are shown in Figure 4.11, superimposed to the original data (cf. Figure 4.8).

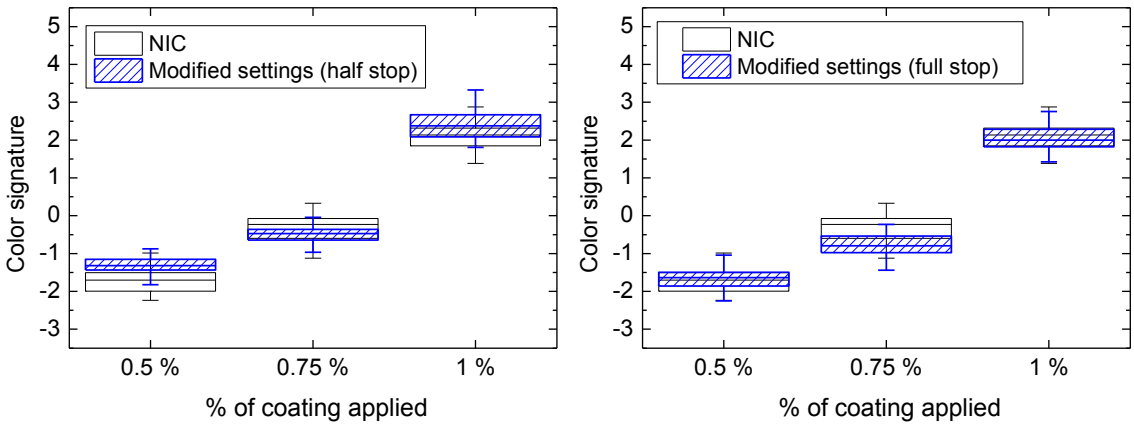


Figure 4.11. *Strategy 1 for light changes compensation: color signature obtained with the new camera settings obtained from the PLS models built on the (a) ±1/2 and (b) ± 1 stop excitation data.*

The color signature obtained with the new camera settings are equivalent to the original one, hence demonstrating the effectiveness of the proposed strategy. It should be stressed at this point that the residual color signature discrepancies in Figure 4.11 are related to the variability of the color signature metric at low percentage of coating applied, an aspect that has been already discussed by García-Muñoz and Gierer (2010). The images collected using the settings suggested by the PLS models, in fact, are different from those used in the calibration step not only in terms of light conditions, but also in terms of tablets being imaged (due to shuffling between two subsequent images).
It is worth to mention that the settings indicated by the PLS model were indeed the best among the 27 (= 3^3) possible combinations, as was checked (results not shown) by collecting images of the tablets at 0.75% of coating applied also for the (27–9) = 18 combinations that were not considered in the DoE matrix **X**.

The proposed strategy allows to use the available MIA model without any change or adaptation, but requires some preliminary images to be collected. Therefore, it may prove helpful to assess its suitability in advance (i.e. to test whether re-calibrating the model would require less images). The main limitation lays in the discrete (i.e., not continuous) correction ability that is provided by the manipulation of the camera settings, as potentially no one of the combinations considered in the DoE matrix **X** can return the desired illuminating conditions.

4.3.2.3 Strategy 2: adapting the score space

Figure 4.9 clearly shows the effects (i.e., shift and compression) of a light alteration on the score space. Since the distortion in the score distribution is beyond that of a simple shift (which is easily corrected with a bias term), the use of the DTW is proposed to align (or synchronize) the two-dimensional distribution of the scores under the new light conditions against that of the two-dimensional distribution observed under NIC, as described in Section 4.2.3.

In the application under study, the average $\mathbf{t}_1$ and $\mathbf{t}_2$ score distributions obtained from the images of the mid-point of the coating progression (40 images in total, the 20 collected under NIC for calibration and additional 20 images collected under dimmed light, step 2a in Figure 4.10) were used to evaluate the $\mathbf{t}_1$- and $\mathbf{t}_2$-warping paths (step 2b in Figure 4.10), which were then applied to all the images collected under reduced light (step 2c in Figure 4.10).

The effectiveness of the correction in terms of color signature evolution is shown in Figure 4.12.

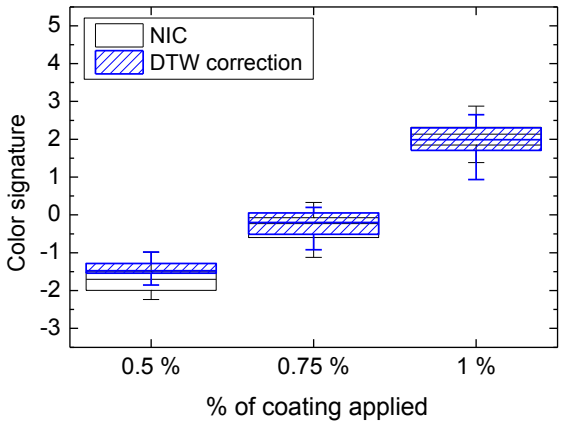


Figure 4.12. *Color signature obtained from the application of the DTW correction on the score space.*

With respect to Strategy 1, the DTW returns a correction of the light tailored on the specific MIA model (whereas the correction of the settings can be applied also to other image-based models). The main advantage is that the DTW correction requires the collection of a limited number of images under altered light.

4.3.3 Camera transfer

4.3.3.1 On the transferability of MIA models across cameras

This Section presents some general considerations regarding the transferability of MIA models across different cameras. A simple case study is illustrated, which refers to images of two subjects (either tablets or calibration standards) of different colors (cf. Figure 4.1a and 4.1e). Two different scenarios are analyzed.

- Scenario 1: transfer between two cameras of the same type and model (camera A1 to camera A2), which may occur in practice when a new camera is bought to replace one of the same type, or when a new machine vision system is installed, replicating an existing one;
- Scenario 2: transfer between different cameras (camera A1 to camera B or camera B to camera A1), which may occur when a new camera is bought to replace one of lower quality (A1 to B) or when a new machine vision system is installed, replicating an existing one but using a camera of lower quality (B to A1).

Prior to the discussion of the two possible scenarios, the extraction of the features from the images and their subsequent manipulation is briefly discussed.

Images manipulation

For each camera used as a reference (i.e. for which the quality assessment model is calibrated), an MIA model was built by concatenating the 20 available images. Each image was then projected onto the MIA model, and for each color of the imaged subject (red, green, blue and yellow for the color standards, and red, green, pink and white for the colored tablets), a feature vector $\boldsymbol{\gamma}_{color}$ was extracted:

$$\boldsymbol{\gamma}_{color} = [\mu_{\mathbf{t}_1}\ \mu_{\mathbf{t}_2}\ \sigma^2_{\mathbf{t}_1}\ \sigma^2_{\mathbf{t}_2}\ \sigma_{\mathbf{t}_1\mathbf{t}_2}] \tag{4.8}$$

collecting the mean of its $\mathbf{t}_1$ and $\mathbf{t}_2$ scores ($\mu_{\mathbf{t}_1}$ and $\mu_{\mathbf{t}_2}$, respectively), their variance ($\sigma^2_{\mathbf{t}_1}$ and $\sigma^2_{\mathbf{t}_2}$) and their covariance ($\sigma_{\mathbf{t}_1\mathbf{t}_2}$). The scores of each color were extracted from those of the whole image using elliptical masks as described in Section 4.3.1. The features vectors of each color (one per image) were averaged and stacked on the top of each other to form a [4×5] matrix $\boldsymbol{\Gamma}$. Eventually, the quality assessment model was obtained by fitting matrix $\boldsymbol{\Gamma}$ to a 2-PC PCA model, the scores being the features of interest.

In order to use this quality assessment model with a different camera, the features were extracted with the following procedure. First, the images were projected onto the MIA model calibrated with the reference camera. Then, the MIA model scores were manipulated using (4.8) to build the $\boldsymbol{\Gamma}$ matrix for the new camera. Finally, the features were extracted by projecting $\boldsymbol{\Gamma}$ onto the PCA model calibrated for the reference camera.

Scenario 1

Let us define as A1/Wb1 the camera/white balance combination used as a reference. The projection of the images collected with other type-A camera/white balances onto the PCA model built on $\mathbf{\Gamma}_{A1/Wb1}$ is shown in Figure 4.13.

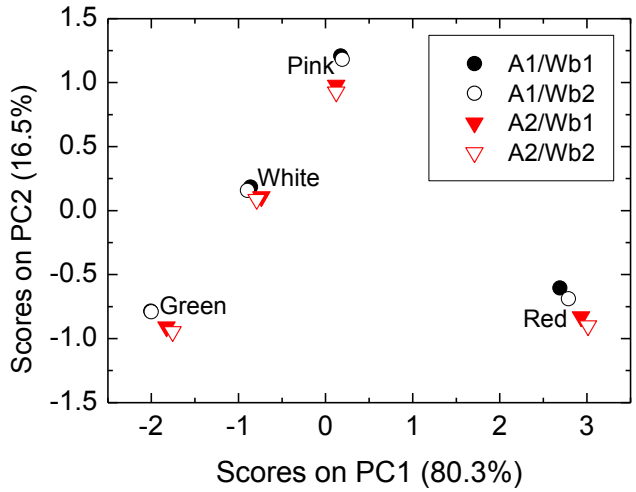


Figure 4.13. *Scenario 1 score space (PCA model built on $\mathbf{\Gamma}_{A1/Wb1}$) obtained from the images of colored tablets.*

Clearly, a bias can be observed in Figure 4.13, i.e. for a given tablet color, the images collected with the second type-A camera (camera A2) are slightly different, although the two type-A cameras are identical (Ilie and Welch, 2005). As a general comment on the need of transferring the quality assessment model between the two cameras, Figure 4.13 suggests that the specificity of each application should be considered. For example, if the purpose is that of recognizing objects of different colors, there is no actual need to transfer the model, since the scores of each color in Figure 4.13 are clearly separated, regardless of the camera/white balance combination considered. However, when considering a specific color (e.g., red), the bias indeed needs to be handled (see Section 4.3.3.2).

Scenario 2

Assuming camera B (lower resolution) as the reference one, the projection of the images collected with camera A1 (higher resolution) onto the PCA model built on $\mathbf{\Gamma}_B$ is shown in Figure 4.14. Figures 4.14a (colored standards images) and 4.14c (color tablets images) suggest that a strong bias between the two different cameras exists and that the bias needs to be tackled, independently from the application considered. To this purpose, Figures 4.14b and 4.14d present the results that can be obtained by applying the two strategies previously proposed for the technology transfer problem. Namely, for the RGB transfer correction, the transformation matrix $\mathbf{S}$ was evaluated using the average RGB values of the four colors of the tablets and of the black background, i.e. $K = 5$.

Figures 4.14b and 4.14d highlight that the DTW correction returns better results. The superior performance of both the DTW and the quadratic RGB transfer over linear RGB can be explained considering that they both enforce a nonlinear correction. With respect to the quadratic RGB transfer, the DTW one has the additional advantage of being calibrated considering, for each tablet color, the entire color distributions and not simply the average R, G and B values.

Both these features explain the different performances of the correction strategies. As a general comment, Figure 4.14 suggests that, for Scenario 2, even in a simple case such as the one considered in this Section, a correction is necessary to compensate for the differences between cameras.

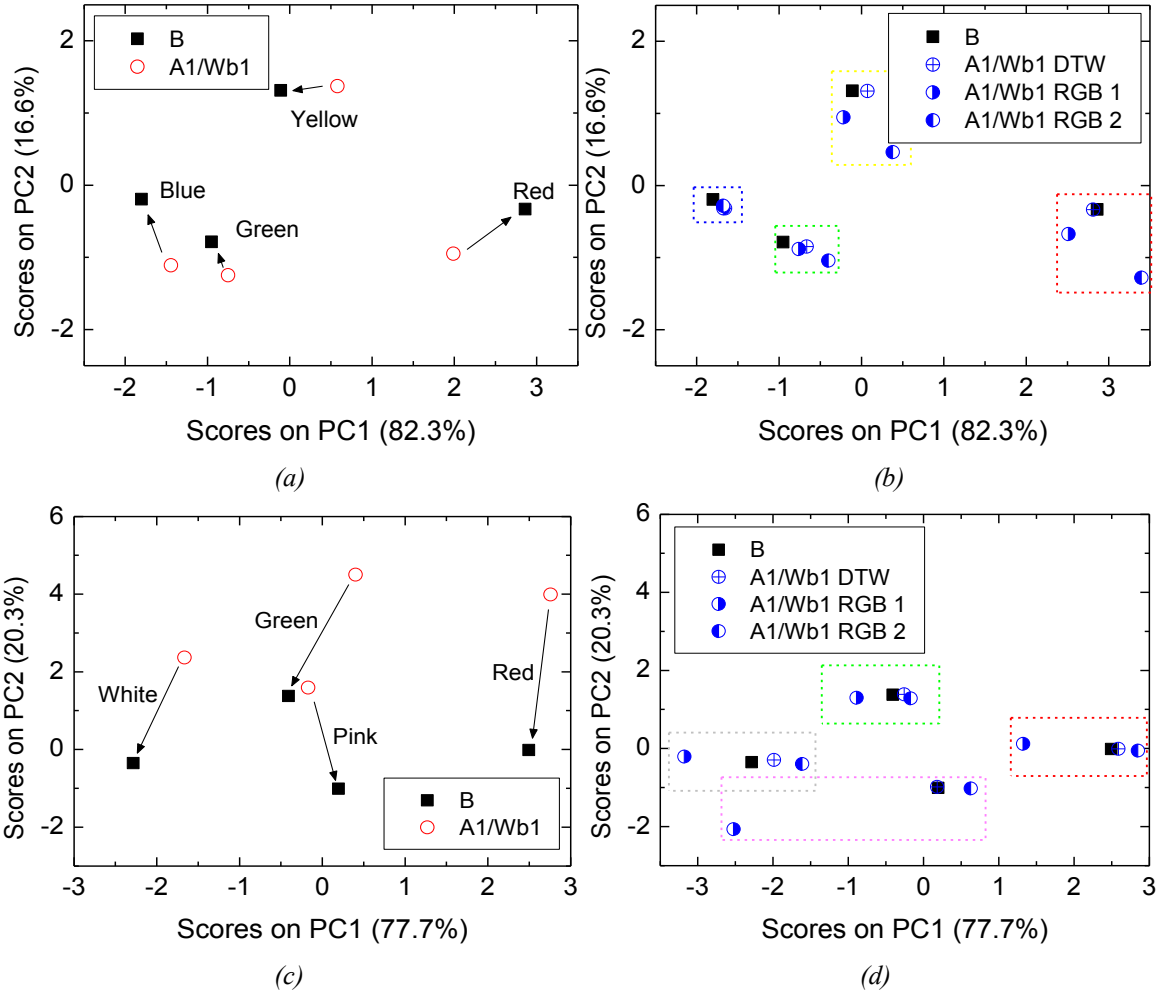


Figure 4.14. *(a; c) Scenario 2 score spaces, with PCA models built on Γ_B using the images of the colored standards (a) and of the color tablets (c). The arrows indicate the points the projections refer to. (b; d) Projection of the features extracted from the scores of the camera A1 images onto the PCA model built on Γ_B after the application of the DTW and RGB transfer correction, respectively for (a) and (c). Boxes are drawn to separate points belonging to different colors.*

4.3.3.2 A technology transfer case study

The transfer of an MIA model is discussed in this Section for the two scenarios previously defined with reference to the case study described in Section 4.3.2.1.

Case study

The box plots in Figure 4.15 reports the color signature for the reference cameras considered for the two scenarios (the A1/Wb1 combination for camera A1 and camera B respectively; each camera was calibrated independently).

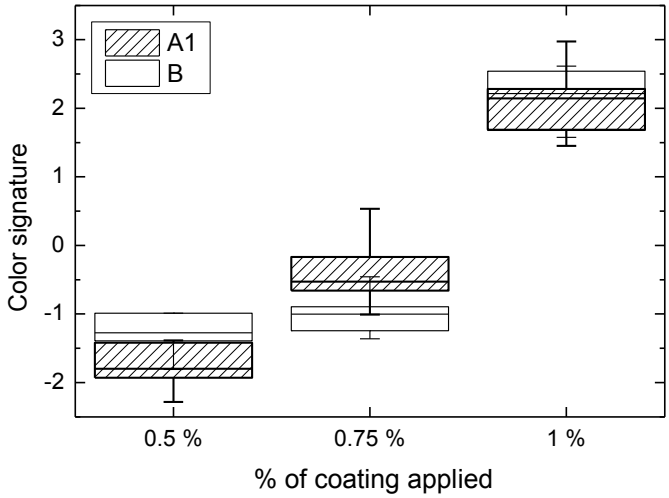


Figure 4.15. *Color signature evolution for camera A1 (hashed box) and camera B (open box).*

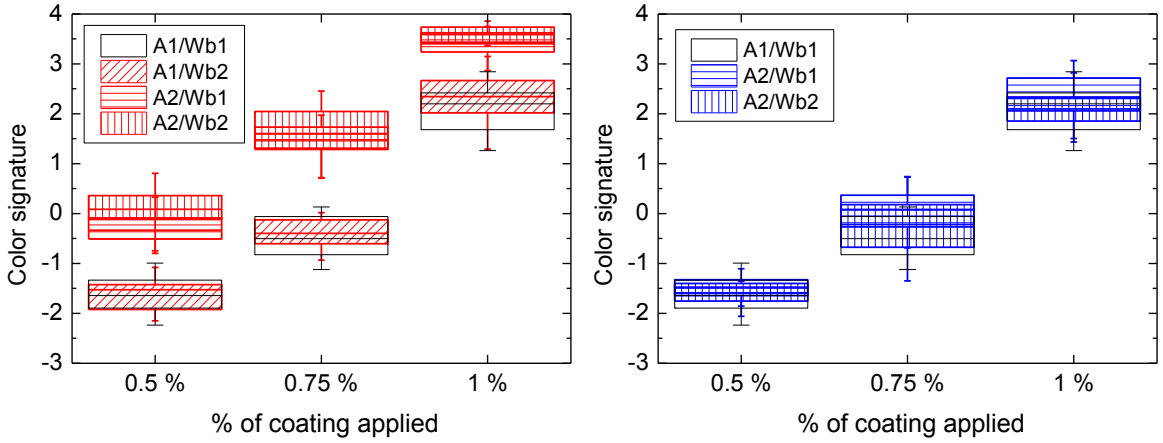
It is useful to note that the important issue is not the absolute value of the color signature, but the fact that it changes as more coating material is applied. Hence, Figure 4.15 shows that the performances of the two cameras are very different: whereas in camera A1 the variation of the color signature changes almost linearly with the % of coating applied, with the distributions at two subsequent percentages of coating applied being statistically different, in camera B the variation is strongly nonlinear, and the distributions at 0.5 % and at 0.75 % are almost indistinguishable.

In the following, the two scenarios previously described are considered. The DTW correction was calibrated as proposed in Section 4.3.2.3, i.e. by evaluating the warping path considering images of tablets of a given percentage of coating material applied (i.e. 0.75 %) collected with the two cameras. For the RGB transfer correction, instead, the same transformation matrix **S** described in Section 4.2.4 was used.

Scenario 1

Figure 4.16a presents the color signatures extracted by projecting onto the reference model the images collected with the camera/white balance combinations not used in the

calibration step (i.e. A1/Wb1), whereas Figure 4.16b shows the effect of applying the DTW correction to the score spaces.



(a) *(b)*

Figure 4.16. *Scenario 1 color signatures extracted by projecting onto the reference model the images collected with the camera/white balance combinations not used in the calibration step (i.e. A1/Wb1) (a) without applying any correction and (b) with the DTW correction.*

As expected (see Figure 4.16a), the color signature extracted from camera A2 turns out to be shifted with respect to the one of the reference camera A1. The DTW correction (Figure 4.16b) effectively matches the color distributions of the two cameras. Also the RGB transfer approaches were found to be effective (results not shown).

Scenario 2

Scenario 2 results are shown in Figure 4.17, considering both the lower resolution camera (camera B; Figures 4.17a-b) and the higher resolution camera (camera A1; Figures 4.17c-d) as a reference.

In both cases, the projection of the images collected with the other camera returnes anomalous color signatures (particularly in Figure 4.17c). The DTW correction does not return a perfect match of the statistics of interest (especially in Figure 4.17b), and the reason for that is the way it was applied. In fact, DTW was calibrated at a given point of the coating distribution (0.75 %) and then linearly applied to the entire distribution. The fact that the color signatures are not perfectly overlapped suggests that the model calibrated on the "old" camera should not be used in a quantitative way with the "new" one. Nevertheless, considering the purpose of the correction strategy (i.e., reducing the downtime until the model is recalibrated), for the case study under investigation the quality assessment model can be used considering that:

- for the case of Figure 4.17a-b (i.e. transferring from a low resolution camera to a higher resolution one), the camera A1 warped color signature allows one to correctly rank the

tablets according to the percentage of coating applied (notice that this was not possible with the reference camera);

- for the case of Figure 4.17c-d (i.e. transferring from a high resolution camera to a lower resolution one), the camera B warped color signature well resembles that of Figure 4.15, for which the previously mentioned limitations still hold.

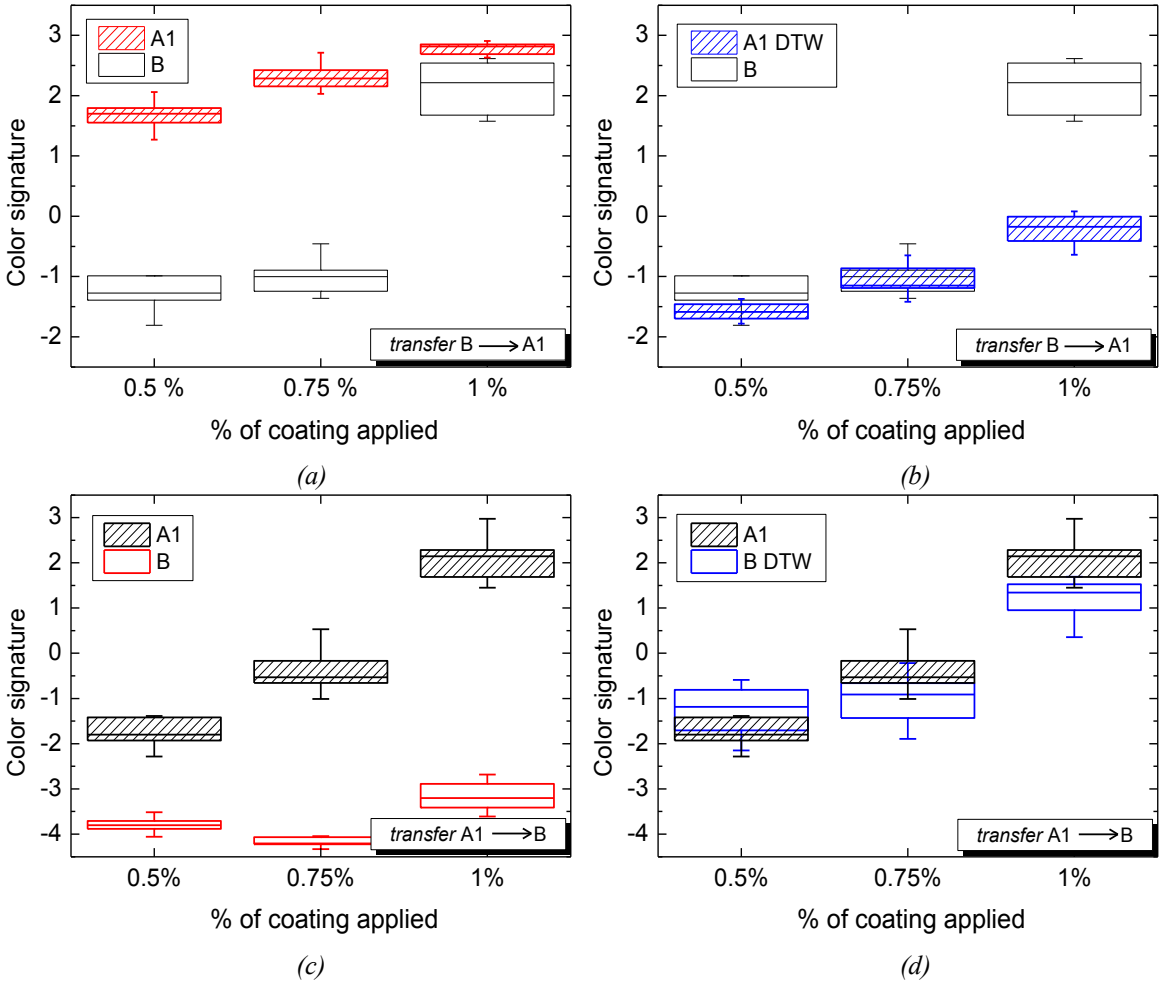
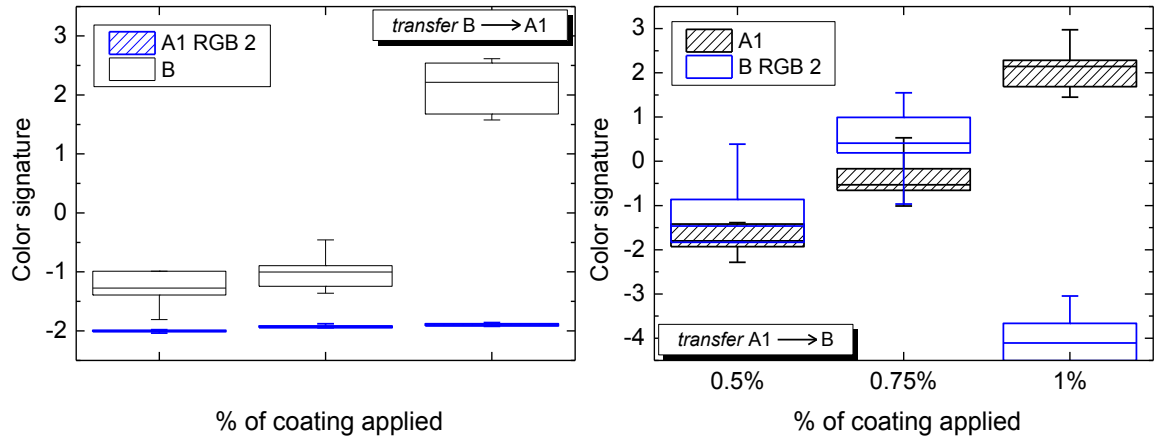


Figure 4.17. *Scenario 2. Color signature extracted by projecting the images collected with (a)-(b) camera A1 and (c)-(d) camera B onto the model calibrated on camera B and camera A1, respectively. In (a)-(c) the color signatures are extracted without applying any correction strategy, while in (b)-(d) the DTW correction is applied. The direction of the transfer is indicated in all Figures.*

With respect to the DTW correction, instead, the RGB transfer was found not be effective in the correction, returning anomalous color signatures, as shown in Figure 4.18 for the RGB 2 strategy (results for the RGB 1 strategy were found to be similar).



(a) *(b)*

Figure 4.18. *Scenario 2. Color signature extracted by projecting the images collected with (a) camera A1 and (b) camera B onto the model calibrated on camera B and camera A1 (respectively) with the RGB 2 correction applied. The direction of the transfer is indicated in all Figures.*

4.4. Conclusions

Despite a change in the lighting conditions or in the camera can dramatically affect the results that can be obtained from an image-based product quality monitoring model, the problem of the automatic maintenance of machine vision systems has never been addressed in the open literature. To fill this gap, this Chapter proposes first a strategy to monitor the status of a machine vision system has been proposed. The monitoring model is built on the features extracted from the 2D histogram-scatter plot obtained from the projection onto an MIA model of the images of four color standards.

Two strategies have then been proposed to cope with a detected light alteration, and they have been tested through a pharmaceutical process case study involving the multivariate image analysis of tablets at different percentage of coating applied. The first strategy aims at determining a new combination of camera settings that returns an exposure of the subject as close as possible to the one considered during the model calibration step, with the selection of the settings guided by a PLS model that relates the camera settings with features extracted from the images. The second strategy, instead, adapts the model to the new light conditions by stretching the score space through the application of a dynamic time warping approach. Both strategies have been shown to effectively correct the model diagnostics of interest (i.e. the color signature evolution as more coating is applied onto the tablets).

The dynamic time warping approach has been shown to be effective also in transferring an advanced technology for product quality assessment across different cameras. Two case studies have been considered. The first one consisted in images of subjects of different

colors. Results showed that the effect of collecting the images with a camera, which is different with respect to the one used in the model calibration step, should be analyzed on a case-by-case basis. In fact, despite differences between cameras have been always observed, their importance needs to be related to the specific application under investigation. The second case study, involving (again) the multivariate image analysis of tablets at different percentage of applied coating, showed that the differences between cameras can be effectively handled by applying the dynamic time warping to stretch the images score spaces. Other correction strategies, such as linear or quadratic transformations between the color spaces of different images, have been shown to return worse results.

Finally, it should be stressed that a novel method to scale the scores obtained from the images has been developed. The "color scaling" has been shown to be a more robust scaling with respect to the one proposed by Yu and MacGregor (2004), as it ensured all the scores values being bounded in the [0-255] range even when images were collected under different light conditions or with a different camera.

Chapter 5

Multispectral data classification using similarity factors*

In this Chapter, the use of similarity factors based on principal component analysis is proposed for the classification of different types of spectral datasets. The proposed classification technique has a very intuitive graphical interpretation and works through an assigned sequence of pretreatment steps on the raw signals, which avoids the trial-and-error selection of the most appropriate preprocessing method usually necessary when dealing with spectral data (independently from their origin). The proposed strategy is successfully validated through two food engineering case studies. A third case study illustrates how the similarity factor-based strategy can be extended to the classification of multi/hyper spectral images.

The Chapter is organized as follows. In the first section, existing classification strategies are briefly reviewed. The second section presents the similarity factors and their use in the classification of spectral data. In the third section, the dataset onto which the proposed classifier is tested are described. The fourth section shows how the similarity-based classification can achieve the same classification performance as other established techniques, though avoiding the trial-and-error selection for spectra preprocessing. Finally, conclusions and further issues are discussed.

5.1 Introduction

The development of innovative analytical technologies based on spectral data has been drawing much attention by academic and industrial researchers in diverse fields of application (e.g. remote sensing, food engineering, fine chemistry, quality characterization, process control, environment protection, archeology, etc…). Different types of spectral data can be (sometimes easily) collected through widespread analytical technologies, such as nuclear magnetic resonance, ultraviolet spectroscopy, visible spectroscopy, near-infrared (NIR) spectroscopy, Raman and mass spectroscopy, and hyperspectral imaging.

*Ottavian, M., P. Facco, L. Fasolato and M. Barolo (2012). Multispectral data classification using similarity factors. *Chemom. Intell. Lab. Syst*., **118**, 13-23.

Commonly, spectral data are utilized for estimation and classification purposes (Martens and Naes, 1989). In particular, several mathematical techniques have been applied to classify spectral data, ranging from traditional methods, such as k-nearest neighbor (k-NN; Sharaf, 1986) or linear and quadratic discriminant analysis (LDA and QDA; Seber, 1984), to multivariate ones such as partial least-squares discriminant analysis (PLS-DA; Chapter 2, section 2.1.2.2) or soft independent modeling of class analogy (SIMCA; Wold, 1977). More sophisticated classifiers, such as support vector machines discriminant analysis (Luts *et al*., 2010) or artificial neural networks (Braspenning, 1995) have been proposed to handle nonlinear problems. Wavelet-based algorithm have been proposed as well (Cocchi *et al*., 2001, Aujol *et al*., 2003) to take advantage of multi-resolution approaches (Addison, 2002).

Typically, the abovementioned prediction/classification methods are implemented after a convenient (and in most cases tailored) pretreatment of the spectral data has been carried out. In fact, preprocessing of the available spectra is usually necessary to remove the variability of the predictors that is not related to the response. Despite the existence of many guidelines and studies (Candolfi *et al*., 1999; Afseth *et al*.; 2006, Burger and Geladi, 2006; de la Roza-Delgado *et al*., 2007), the selection of the best preprocessing strategy is typically a trial-and-error procedure (Broad *et al*., 2006). Therefore, a method that could achieve reliable classification results avoiding the trial-and-error selection of the optimal pretreatment(s) for the spectral data would be most welcome.

The classification algorithm for spectral data that we present in this paper is based on the similarity factors proposed by Krazanowski (1976) and later modified by Gunther *et al*. (2009). In the proposed approach, a matrix is associated to a multi-resolution representation of a signal, and a principal component analysis (PCA) model is then used for each sample (i.e. matrix) created from the available dataset. Each PCA model identifies the directions of maximum variance of the data matrices, thus extracting the classification-relevant information stored into the multiple resolutions of the available spectral data. This information is compared to that pertaining to the reference samples, using similarity indices in order to discriminate between different classes of signals for an effective classification. The proposed algorithm is coupled with a forward variable selection strategy (Jennrich, 1977) applied on window instead of single variables (Andersen and Bro, 2010), with the aim of both assisting in the assessment of the classification performance and allowing for a better separation between different classes.

The main intent of this Chapter is not to compare the performance of different classification techniques. Rather, the purpose is to show that the proposed classification strategy can achieve the same classification performance as other established techniques, but avoids the use of trial-and-error procedures for spectra preprocessing.

The proposed method is validated through two case studies in the field of food authentication using NIR spectroscopy data. One additional case study shows how the proposed procedure can be extended to the classification of multi/hyper spectral images.

5.2 Similarity factors for spectral data

The proposed algorithm for the classification of spectral data based on similarity factors calculated from PCA models built on the available spectra is presented in this section. The rationale behind the proposed approach (which also serves as a limitation for its effective utilization) is that samples can be classified according to the shape of their spectra. Accordingly, it will be shown that similarity factors can quantify the similarity between two spectra shapes. The key issue is finding a proper representation of the spectral data in such a way as to emphasize the most important characteristics of the signals and to extract them at different scales of resolution. Wavelet decomposition (Mallat, 1989; Addison, 2002) can be used to this purpose, providing also for a de-noising action on the available spectra. From a very general perspective, the proposed classification method works as follows: a reference sample is chosen for each class, and the similarity of each other sample with the references is assessed. Samples are then associated to the class for which they show the greatest similarity.

In the following subsections the proposed strategy is presented in detail. The workflow will be as follows. First, each sample (signal) is decomposed using wavelet transform, and the resulting decomposition is used to create a sample matrix. Then, a PCA model is calibrated for each sample (i.e., matrix), and similarity factors are used to quantify how similar/dissimilar each sample is with respect to the reference of each class (i.e. to classify the samples). At the same time, the wavelengths intervals resulting most relevant to the classification are extracted.

5.2.1 Building the sample matrix

Discrete wavelet transform (DWT; Addison, 2002) is used to extract the multiresolution features from the available spectra. In this study, Daubachies 4 is used as a wavelet function. In order to extract the spectral characteristics at different scales of resolution, a matrix is associated to each spectrum. A multi-detail approach is employed in this study, i.e. the reconstructed details at different scales are used, as shown in Figure 5.1. The original spectrum (row vector) of dimension [1 sample × N wavelengths] is first decomposed via DWT into details at different scales of resolution. Then, after discarding the detail at the first (highest) resolution scale for denoising purposes, the details at different resolution scales are concatenated vertically in a sample matrix $\mathbf{X}$ of dimension [$I \times N$], where (I+1) is the last (lowest) resolution scale retained in the DWT decomposition.

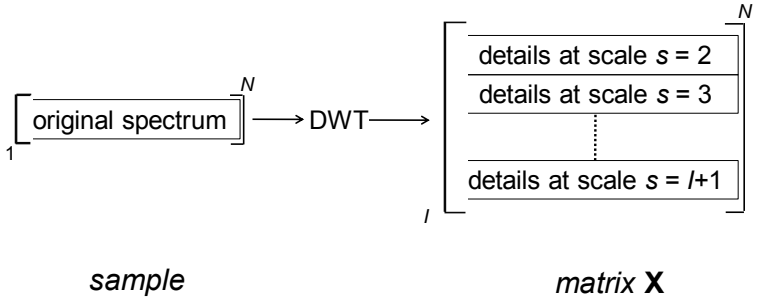


Figure 5.1. *Schematic of the association of matrix **X** to a sample spectrum.*

Note that only the reconstructed details (Addison, 2002) are employed to build matrix **X**. This ensures that all signals can have the same length (N) and that the details at different resolution scales for a certain wavelength can be concatenated correctly.

A PCA model is then designed on **X** to study the correlation between the wavelengths at different resolution scales, and to extract the information embedded in the data from the redundancy generated by the collinearity between wavelengths. Since the correlation structure highlighted by the PCA model is expected to be similar for samples belonging to the same class, the similarity factors introduced by Krzanowski (1979), and later modified by Gunther *et al*. (2009), are used to quantify how similar two PCA models (i.e. two samples) are.

With respect to typical multi-resolution approaches employing the details at all resolution scales plus the approximation at the lowest one (Bharati *et al*., 2004), approximations are not used here because it was found that the information needed for sample classification is entirely stored in the details.

5.2.2 PCA-based similarity factors

Consider two PCA models (both using A principal components) for data matrices $\mathbf{X}_i$ [$M_i \times N$] and $\mathbf{X}_j$ [$M_j \times N$]. The similarity factor S_{ij} between the models is defined as (Krzanowski, 1979):

$$S_{ij} = \frac{1}{A}\sum_{a_i=1}^{A}\sum_{a_j=1}^{A}\cos^2\theta_{a_i a_j} = \frac{\mathrm{trace}(\mathbf{P}_i^{\mathrm{T}}\mathbf{P}_j\mathbf{P}_j^{\mathrm{T}}\mathbf{P}_i)}{A}, \tag{5.1}$$

where $\theta_{a_i a_j}$ is the angle between the a_i-th loading of the first model and the a_j-th loading of the second one, and $\mathbf{P}_i$ and $\mathbf{P}_j$ are the corresponding loadings matrices. The similarity factor compares the direction of maximum variability of two datasets by exploring the collinearity of the principal components of the two PCA models built on those datasets. An improvement of the similarity factors has been proposed by Gunther *et al*. (2009), who modified (1) to explicitly consider that each loading captures a different amount of the total

variance. In practice, the eigenvalues of the eigenvector decomposition are introduced as weights to obtain the modified similarity factor S_{ij}^{λ}:

$$S_{ij}^{\lambda} = \frac{\sum_{a_i=1}^{A}\sum_{a_j=1}^{A}(\lambda_{ia_i}\lambda_{ja_j})\cos^2\theta_{a_ia_j}}{\sum_{a=1}^{A}\lambda_{ia}\lambda_{ja}} = \frac{\mathrm{trace}[(\mathbf{P}_i^w)^{\mathrm{T}}(\mathbf{P}_j^w)(\mathbf{P}_j^w)^{\mathrm{T}}(\mathbf{P}_i^w)]}{\sum_{a=1}^{A}\lambda_{ia}\lambda_{ja}} \tag{5.2}$$

where $\mathbf{P}^w$ is the weighted loading matrix:

$$\mathbf{P}_i^w = \mathbf{P}_i\mathbf{\Lambda}_i\,, \tag{5.3}$$

with $\mathbf{\Lambda}_i$ being an $[A\times A]$ diagonal matrix of the square roots of the eigenvalues:

$$\mathbf{\Lambda}_i = \begin{bmatrix} \sqrt{\lambda_{i1}} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_{i2}} & 0 & \dots \\ \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \sqrt{\lambda_{iA}} \end{bmatrix}. \tag{5.4}$$

The introduction of the eigenvalues in (2) allows weighing more heavily the first PCs, which can lead to better classification results. Both similarity factors (1) and (2) have the useful property of being bounded between 0 (no similarity at all between the samples, i.e. orthogonal loadings between the two PCA models) and 1 (complete similarity between the samples, i.e. models with identical principal components because of complete collinearity between loadings: $\mathbf{P}_i = \mathbf{P}_j$).

5.2.3 Sample classification

The proposed classification algorithm is a supervised method. In fact, it requires a calibration dataset whose class is known *a priori*, in such a way as to define the references for all the classes. Without loss of generality, in this study binary classification problems only are addressed; the extension to multi-class problems is straightforward, and therefore will not be considered.

When the problem is assigning sample *i* to either Class 1 or Class 2 (whose reference matrices have already been identified), the similarity factors S_{i1}^{λ} and S_{i2}^{λ} of the sample with the two references can be calculated using (2), and the sample attributed to the class for which it shows the greatest similarity. Therefore, sample *i* is assigned to Class 1 if $S_{i1}^{\lambda} > S_{i2}^{\lambda}$, i.e. if the sample is more similar to the reference of Class 1 than to the reference of Class 2. Conversely, the sample is assigned to Class 2 if $S_{i1}^{\lambda} < S_{i2}^{\lambda}$.

In the region of the (S_{i1}^{λ}; S_{i2}^{λ}) plane characterized by $0 \le S_{i1}^{\lambda} \le 1$ and $0 \le S_{i2}^{\lambda} \le 1$ (which will be called the (S_{i1}^{λ}; S_{i2}^{λ}) region in the following), samples belonging to Class 1 are expected to be located below the main diagonal ($S_{i1}^{\lambda} = S_{i2}^{\lambda}$), i.e. close to their own reference (1,0). Samples belonging to Class 2 are expected to be located above the diagonal, i.e. close to their own reference (0,1). Samples lying on (or very close to) the diagonal are equally similar to both references, and their classification is uncertain or impossible using similarity factors. Note that the location of a point in the (S_{i1}^{λ}; S_{i2}^{λ}) region (i.e. close to, or far from, the diagonal) also depends on how the two references are similar to one another, i.e. it depends on the value of S_{12}^{λ}.

When a dataset of several samples is available for model calibration, defining the reference is a matter of choice. The ideal reference for a class is the one maximizing the similarity with the spectra of the same class, while minimizing that with the spectra of the other class. To assess whether the reference selection can impact on the classification results, two alternative types of class reference were tested in this study:

- an average spectrum of the class. This is a spectrum obtained by averaging all the available calibration samples along the wavelength direction; the reference matrix for the class is then built from this average spectrum;
- the whole set of samples of the calibration dataset for the class. In this case, the reference matrix is obtained by simply stacking the matrices of the single calibration samples on top of each other.

5.2.3.1 Wavelength selection

When dealing with large and highly correlated **X** matrices (which typically result from spectral data), coupling the classification algorithm with a variable (i.e. wavelength) selection strategy can be beneficial. To this purpose, a forward variable selection strategy is used, and it is applied on windows of variables rather than on single variables, as suggested also for interval PLS by Andersen and Bro (2010). A moving window of assigned width W is used to span all the wavelengths (without overlapping between the windows). At the first iteration, similarity factors are used to classify the calibration spectra within each of the $N_w = N/W$ windows, and the wavelength window ensuring the best classification is selected. During each of the following iterations, spectra are classified using the previously selected window(s) plus one of those not yet selected, scanning the left-out ones once at a time. Iterations stop when adding any new window to the N_w^* windows already selected does not improve the classification result any more.

In order to evaluate the classification performance at each iteration of the wavelength selection procedure, the similarity factors S_{i1}^{λ} and S_{i2}^{λ} between spectrum i and the selected reference for Class 1 and Class 2 are calculated using the wavelength windows selected until that iteration. Then, distance d_{tot} is evaluated, where d_{tot} is the cumulative Euclidean

distance in the (S_{i1}^{λ}; S_{i2}^{λ}) region between the couple of similarity factors of sample *i* and the ideal location of the sample if it exactly matched one of the two references (i.e., either (1,0) if *i* is assumed to belong to Class 1, or (0,1) if *i* is assumed to belong to Class 2):

$$d_{tot} = \sum_{i=1}^{K} \left[(1-\alpha)\, d_{i \to ref} + \alpha \left(C\, d_{i \to ref} \right) \right], \tag{5.5}$$

where:

$$\alpha = \begin{cases} 0, & \text{if } i \text{ is classified correctly} \\ 1, & \text{if } i \text{ is not classified correctly} \end{cases}, \tag{5.6}$$

$$d_{i \to ref} = \begin{cases} \sqrt{(1-S_{i1}^{\lambda})^2 + (S_{i2}^{\lambda})^2} & \text{if } i \in \text{Class 1} \\ \sqrt{(S_{i1}^{\lambda})^2 + (1-S_{i2}^{\lambda})^2} & \text{if } i \in \text{Class 2} \end{cases}, \tag{5.7}$$

K is the total number of available calibration spectra, $d_{i \to ref}$ is the Euclidean distance of sample *i* from the reference, and *C* is a constant term introduced to penalize wrong classifications. By "classified correctly" in (6), the event is considered where point (S_{i1}^{λ}; S_{i2}^{λ}) lays in the correct portion of the (S_{i1}^{λ}; S_{i2}^{λ}) region. A schematic of the distance calculation strategy is shown in Figure 5.2.

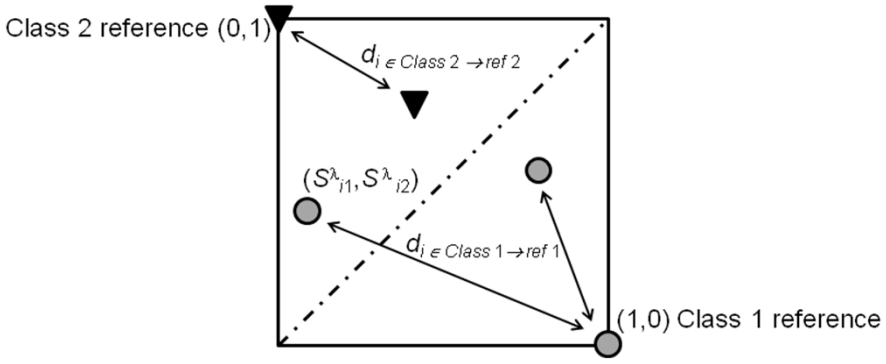
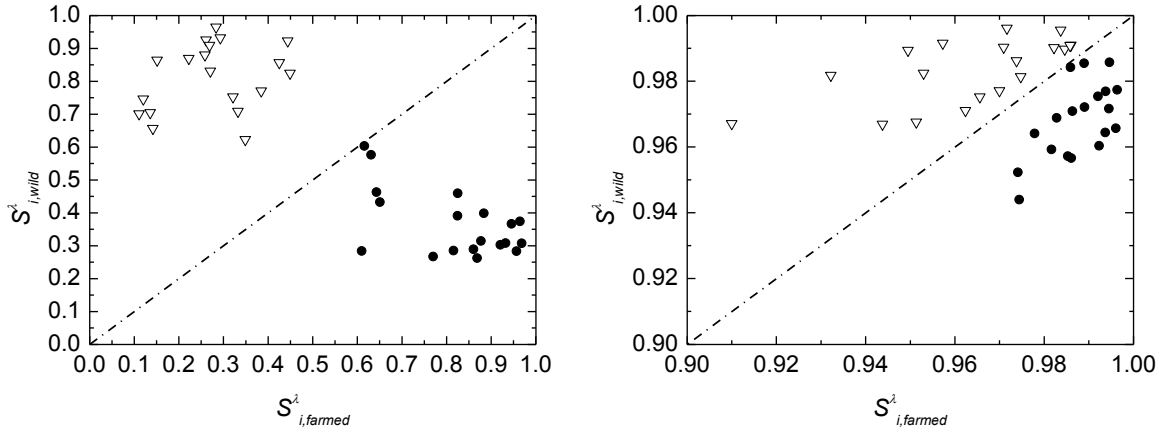


Figure 5.2. *Evaluation of the cumulative Euclidean distance between samples and their own reference.*

The value taken by d_{tot} is a quantitative measure of the ability of a given subset of wavelengths to improve the separation between the two class clusters in the (S_{i1}^{λ}; S_{i2}^{λ}) region: small values of d_{tot} correspond to a small distance between the cluster of points belonging to a class and the reference of that class, and therefore imply a better separation between classes. Note however that the actual value of d_{tot} also depends on how similar the two class references are to each other, i.e. it depends on the value of S_{12}^{λ}. If the two references are very similar to each other ($S_{12}^{\lambda} \cong 1$), then the (S_{i1}^{λ}; S_{i2}^{λ}) points will be located close to the diagonal, the cluster separation will be weak, and eventually d_{tot} will be large.

The reverse is true if the two references are very different from one another ($S_{12}^{\lambda} << 1$). Therefore, d_{tot} serves as a measure of cluster separation only during the wavelength selection procedure: the most appropriate subset of wavelengths is the one providing the smaller value of d_{tot}. However, larger values of d_{tot} (obtained for example by specifying different references for the classes or using different alternatives to scale matrix **X**) do not necessarily imply a larger number of misclassified samples.

To clarify this issue, Figure 5.3 presents two examples of very good classification results (namely, zero misclassifications) for the European sea bass calibration dataset that will be considered later in more detail. These results refer to two different ways to scale the **X** matrix. In one case (Figure 3a), the two class references result quite dissimilar ($S_{12}^{\lambda} = 0.68$), and the wavelength selection procedure returns d_{tot} = 14.7; in the other case (Figure 3b), the two references are much more similar to each other ($S_{12}^{\lambda} = 0.98$), and a larger cumulative Euclidean distance (d_{tot} = 36.7) is obtained. However, in both cases the classification performance of the model in terms of number of misclassifications is exactly the same.



(a) *(b)*

Figure 5.3 *Calibration classification results for the European sea bass dataset (farmed/wild) for two different sets of* ***X*** *scalings, resulting in (a) d_{tot} = 14.7 (model #11 of Table 3) and (b) d_{tot} = 36.7 (model #1 of Table 3). In (b) the x and y axis scales have been magnified to improve readability. Closed symbols: farmed samples (Class 1); open symbols: wild samples (Class 2).*

5.2.4 Interpretation of the similarity factors and discussion on the scaling of the **X** matrix

Before analyzing the results of the proposed classification procedure, it is worth providing some additional comments on the meaning of the similarity factors in the context of spectral data classification. Let us consider the case where the DWT decomposition is carried out up to resolution scale $s = I+1$, and let $(W \cdot N_w^*)$ be the total number of wavelengths eventually selected for the design of the PCA model, N_w^* being the number of

wavelength windows retained according to Section 2.3.1. Therefore, an **X** matrix of dimension $[I\times(W\cdot N_w^*)]$ is associated to each spectrum. The main diagonal of the covariance matrix $\mathbf{X}^T\mathbf{X}$ contains the variance between details at different resolution scales for a certain wavelength, whereas the (i, n) off-diagonal element ($i \neq n$) is the covariance between details at different scales for wavelengths i and n. Therefore, the PCA model built on **X** explains how scale patterns (i.e., the multi-resolution representation of a certain wavelength) correlate among different wavelengths. This can be thought as an indirect way to compare spectra shapes, because it is reasonable to expect that different resolution scales correlate in a likewise manner for a given set of wavelengths only if at those wavelengths the underlying signals (hence samples) are similar. Stated differently, looking for the similarity between PCA models is the same as looking for the similarity between spectra shapes.

The calibration of a PCA model requires **X** to be properly scaled, and the selection of the most appropriate scaling is a matter of choice. With reference to the monitoring of batch chemical processes, Gunther *et al*. (2009) recommend that data matrices be internally autoscaled (i.e., data should be first mean-centered and then divided by their own standard deviation) when similarity factors are evaluated. Due to the very different application considered in this study with respect to that considered by Gunther *et al.* (2009), three different scaling strategies for the **X** matrix were tested to understand the impact of **X** scaling on the classification performance. Namely, the following alternatives were considered: *i*) no scaling at all, *ii*) internal autoscaling, and *iii*) external autoscaling, i.e. autoscaling of **X** with respect to the mean and standard deviation of the reference matrix against which the similarity is assessed (an approach that could potentially magnify similarities and dissimilarities with respect to the reference).

5.2.5 Selection of parameters for the proposed classification strategy

The implementation of the proposed classification strategy requires the values of several parameters to be assigned. The following guidelines for parameter selection, which avoid trial-and-error procedures, can be used:

- (I+1) can be selected as the scale at which the correlation between the original spectrum (i.e., the raw signal) and its approximations at different resolution scales drops below an assigned threshold (e.g. 0.97). A similar procedure was suggested by Facco *et al.* (2010) for a different application. It was found that typically 4 to 6 resolution scales are sufficient when dealing with spectral data;
- the number A of PCs to be retained in the PCA models of each sample and reference matrix can be set as:

$$A = \max\{2, I-1\}, \tag{5.8}$$

which ensures that at least 2 PCs are available for the calculation of the similarity factors. Note that the data stored in the sample matrix are typically highly correlated, which results in a reduction of the actual matrix rank; for this reason, values of A from 2 to 4 are reasonable considering the typical matrix dimension encountered;

- when applying the forward variable selection strategy, a value of 15 can be used for W, this being a reasonable trade-off in terms of window size to avoid using windows that are too large or too small (Andersen and Bro, 2010). As for the penalty parameter C, a small value (C = 3) is sufficient to ensure good classification, since misclassification are already penalized by higher values of the distance from the reference (see Section 5.2.3).

With reference to the selection of the number A of PCs to be retained in each PCA model, it should be stressed that similarity factors cannot be calculated if, given any two samples i and j, $A_i \neq A_j$. This is the reason why the number of PCs must be the same for all samples and should be assigned *a priori*. It was found that using Eq. (8) as a guideline, the PCA models are able to extract almost the entire variance of the sample matrices. Note that commonly used criteria, such as maximization of explained variance or minimization of residuals in cross-validation (Jackson, 1991), are not used in this study to determine the appropriate number of PCs. This is reasonable, both because the models need to have the same structure for all samples and because the models are not used for additional sample projections. Finally, note that the way similarity factors are defined assigns smaller weights to the higher PCs, thus making the selection of the number of PCs less critical than in other applications.

5.3 Available spectral dataset

In this section, the datasets available for this study are described. Two datasets include NIR spectra for the authentication of foodstuff (fish and cheese), while one dataset is concerned with multispectral images of fruits and vegetables.

5.3.1 NIR spectra

NIR spectra were collected in reflectance mode (and then saved as absorbances, i.e. the logarithm of the reciprocal value of the reflectance) with a FOSS NIRSystem model 5000 scanning NIR spectrometer (FOSS NIRSystem, Silver Spring, MD, USA). Data were recorded in a range of wavelengths between 1100 and 2500 nm with at 2 nm interval. Two datasets of NIR spectra were analyzed:

1. European sea bass (*Dicentrarchus labrax*). A typical fraud in seafood is the substitution of wild samples (which can be marketed at a higher price) with farmed

ones; therefore, quick sample classification using NIR spectra is highly desirable (see Chapter 6, Section 6.1.1.2). The available calibration dataset consists of 38 spectra (19 farmed and 19 wild samples), whereas the validation one consists of 66 spectra (32 declared wild and 34 declared farmed samples);

2. Asiago d'allevo cheese. The authentication of the production area (alpine farms vs. lowland factories) for the Asiago cheese is of paramount importance for correct product labeling (see Chapter 6, Section 6.1.2.2). The available dataset consists of 121 spectra, of which 107 from alpine farms (AF) and 14 from lowland factories (LF). Spectra were randomly split into a calibration set (47 AF and 12 LF) and a validation set (60 AF and 2 LF).

The mean spectra for the two investigated datasets are shown in Figure 5.4.

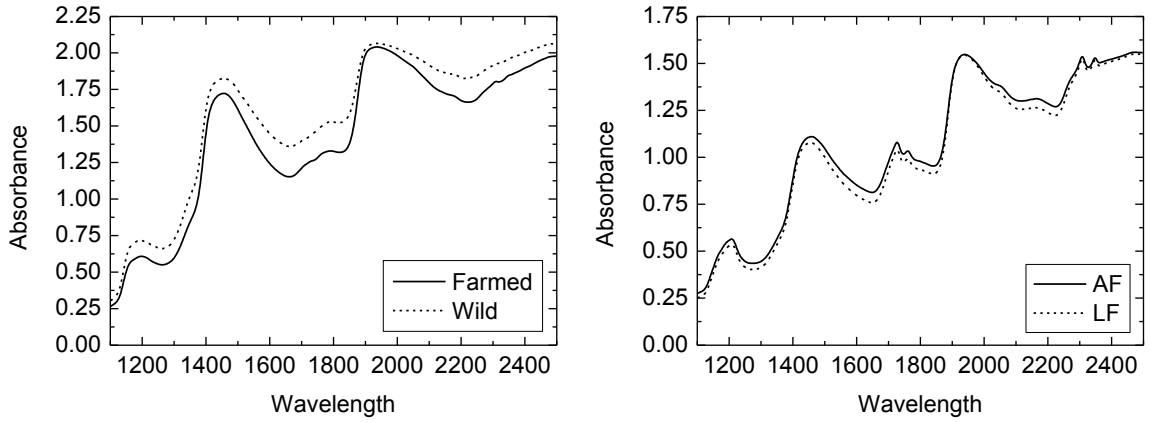


Figure 5.4. *Mean NIR spectra for the two datasets analyzed: (a) European sea bass and (b) Asiago d'allevo cheese.*

5.3.2 Multispectral images

The multispectral images used in this study refer to authentic and fake fruits and vegetables, and have been studied also by Yasuma *et al.* (2008); the problem is discriminating between fake and real products. Each multispectral image of a set of strawberries or peppers has the dimension of [512×512] pixels, and each pixel is characterized by 31 spectral channels corresponding to wavelengths ranging from 400 to 700 nm with a 10 nm interval, giving a total actual size of [512×512×31] per image. The corresponding RGB versions (3 spectral channels per pixel) of the available two image sets are shown in Figure 5.

For both images, 6 sub-matrices were obtained manually by framing a single strawberry or pepper into one matrix. Then, for each image obtained in this way, a subset of 100 randomly selected pixels (hence, a [100×31] sub-image) was used as a calibration dataset, while the remaining pixels were used as an independent validation set.

(a) *(b)*

Figure 5.5. *Multispectral images analyzed in this study: (a) real and fake strawberries (strawberries on the right-hand side are real), and (b) real and fake peppers (upper red, upper yellow and lower green peppers are real).*

5.4 Results and discussion

In the following subsections, the results for the proposed strategy are presented for each dataset. As a primary index of the classification performance, the number of misclassifications is used.

For all datasets considered in this study, the resolution scale selection procedure described earlier showed that four DWT decomposition scales ($I = 3$) were sufficient to extract the spectral features of interest from the available spectra. Note that, as mentioned earlier, the details at the first scale ($s = 1$) were not used, in order to denoise the available signals.

Following Eq. (8), all PCA models were built on $A = 2$ PCs; they extracted more than 90% of the variance from each $\mathbf{X}$ matrix. Note that in the case of images, wavelength selection was not performed due to the limited number of wavelengths available.

For the European sea bass and Asiago d'allevo cheese datasets, classification results are also compared to those from traditional classification algorithms such as *k*-NN, LDA and QDA. Traditional classifiers were applied using two alternative input matrices:

- the matrix of the (preprocessed) spectra;
- the matrix obtained by stacking on the top of each other the vector obtained for each spectrum by concatenating the approximation at the lowest resolution scale plus all the detail coefficients derived from a DWT decomposition (details at the highest resolution scale were not considered, for consistency with the method proposed in this paper). The use of this input matrix allows for a comparison with the typical implementation of DWT in spectra classification problems.

Furthermore, for the matrix of the (preprocessed) spectra, an additional comparison with PLS-DA is provided. *k*-NN, LDA, QDA and PLS-DA were all coupled to a forward

wavelength selection strategy applied on 30 nm windows, thus allowing for a better comparison with the proposed classification strategy.

5.4.1 European sea bass dataset

5.4.1.1 Performance of traditional classification methods

When using *k*-NN (with three values of *k*, namely *k* = 1, 3 and 5), LDA and QDA to classify the sea bass dataset, the very high correlation characterizing the data matrix prevented its direct use as an input to the classification algorithms. Therefore, a PCA model was built on the spectra matrix (with the number of PCs to be retained determined by cross-validation; Chen *et al*., 2009; Balabin *et al*., 2010) and the scores were then used as inputs to the classifiers. For all methods used as a reference (*k*-NN, LDA, QDA and PLS-DA), the wavelength windows to be used were selected by optimizing the classification performance in cross-validation.

The impact of the spectra preprocessing method on the classification results was investigated for four typical spectra preprocessing operations: no preprocessing at all, standard normal variate (SNV) preprocessing, and standard normal variate in combination with first- (SNV-1) or second-order (SNV-2) derivatives. Classification results are reported in Tables 1 and 2, respectively for the (preprocessed) spectra matrix and the matrix derived from the traditional application of DWT. For the calibration dataset, the number of classification errors is reported; for the validation dataset, results are reported as follows. The 66 validation samples are divided into two groups, namely samples declared wild and samples declared farmed; the number of classification errors for each group is determined using the classification reported by Fasolato *et al*. (2010) (who used using chemical variables and morphometric traits for classification) as a reference. Their study showed that, due to 21 substitution frauds, only 11 out of 32 declared wild samples should be actually considered as wild, whereas no declared farmed samples should be classified as wild. For further details, see Chapter 6, Section 6.1.1.3.

Table 5.1 shows that the classification results are highly dependent on the spectra pretreatment. However, whether SNV, SNV-1 or SNV-2 are the most effective preprocessing methods depends on the classification model. Therefore, as expected, the most effective spectra preprocessing method should be determined by trial-and-error. The same considerations hold true also when a standard DWT approach is used to build the input data matrix (Table 5.2), although in this case SNV-1 seems to provide overall better results.

Table 5.1. *European sea bass dataset: effect of different spectra preprocessing on the classification results of k-NN, LDA, QDA and PLS-DA applied on the (preprocessed) spectra matrix.*

Classifier	Spectra preprocessing method	Errors (calibration)	Errors (validation)	
			Declared wild	Declared farmed
k-NN	no preprocessing	1	0	11
	SNV	1	2	2
	SNV-1	0	1	1
	SNV-2	0	1	1
LDA	no preprocessing	0	1	0
	SNV	1	1	1
	SNV-1	0	2	1
	SNV-2	1	2	0
QDA	no preprocessing	1	2	0
	SNV	0	8	9
	SNV-1	0	2	0
	SNV-2	0	2	1
PLS-DA	no preprocessing	2	1	3
	SNV	0	1	1
	SNV-1	0	1	3
	SNV-2	2	1	4

Table 5.2. *European sea bass dataset: effect of different spectra preprocessing on the classification results of k-NN, LDA and QDA applied on the matrix derived from DWT.*

Classifier	Spectra preprocessing method	Errors (calibration)	Errors (validation)	
			Declared wild	Declared farmed
k-NN	no preprocessing	5	6	5
	SNV	0	1	3
	SNV-1	1	1	2
	SNV-2	4	1	5
LDA	no preprocessing	7	1	4
	SNV	3	1	6
	SNV-1	1	0	2
	SNV-2	8	4	5
QDA	no preprocessing	1	0	1
	SNV	3	1	5
	SNV-1	0	1	0
	SNV-2	9	11	6

5.4.1.2 Classification using similarity factors

Table 5.3 reports the classification results in calibration obtained using the proposed similarity factor-based method. The impact of the spectra preprocessing method (SNV, derivatives or their combinations) on the classification results was assessed considering the same four preprocessing methods as in Tables 5.1 and 5.2; however, note that these preprocessing options were explored with the only aim to show that there is actually no need to assign one of them for similarity factor-based sample classification. Only the results obtained for the best preprocessing and with no preprocessing at all are reported in Table 5.3 for the sake of conciseness. Furthermore, since in principle the classification results might depend on the way matrix **X** is scaled (see Section 5.2.4), three **X** matrix scaling procedures were compared: no autoscaling at all, internal autoscaling, and external autoscaling. Additionally, two different class references (indicated as "mean spectrum" and "whole class" in Table 5.3) were tested to assess whether the way the reference is defined may affect the capability to separate the classes.

Table 5.3. *European sea bass calibration dataset: effect of different spectra preprocessing, input matrix scaling and class references on the classification results obtained with the similarity factors.*

Spectra preprocessing	X scaling	Class reference	d_{tot}	S^2_{12}	Errors (calibration)	Errors (validation)		Model #
						Declared Wild	Declared Farmed	
no preprocessing	no autoscaling	mean spectrum	36.7	0.98	0	0	1	1
		whole class	36.7	0.98	0	0	1	2
best preprocessing (SNV-2)		mean spectrum	26.9	0.77	0	0	2	3
		whole class	23.3	0.64	0	0	2	4
no preprocessing	internal autoscaling	mean spectrum	30.6	0.80	1	0	2	5
		whole class	30.3	0.88	1	0	0	6
best preprocessing (SNV-2)		mean spectrum	17.1	0.58	0	1	3	7
		whole class	18.6	0.60	0	0	2	8
no preprocessing	external autoscaling	mean spectrum	27.6	0.87	1	0	2	9
		whole class	21.2	0.73	1	0	2	10
best preprocessing (SNV-2)		mean spectrum	14.7	0.68	0	0	2	11
		whole class	14.7	0.61	0	0	2	12

Several considerations can be drawn from the analysis of Table 5.3. First, very good classification results are always obtained in calibration, regardless of the spectra preprocessing method. Therefore, spectra preprocessing impacts mainly on the value of d_{tot}, rather than on the number of misclassifications. Secondly, the use of autoscaled data improves the separation between the two classes (smaller d_{tot} values), with autoscaling with respect to the mean and standard deviation of the class reference (external autoscaling) slightly outperforming internal autoscaling. This result (i.e., that performing

sample autoscaling before PCA decomposition returns better separation) holds also for the other NIR dataset considered (cf. Table 5.7), leading to the conclusion that data should be autoscaled and the recommendation given by Gunther *et al.* (2009) in an entirely different context is valid also for the application of the similarity factors proposed in this study. Whether the mean spectrum or the whole class should be used to define the reference seems unimportant for class separation.

For graphical analysis, the results for models #11 and #1 of Table 5.3 have been reported in Figure 5.3. The two farmed samples that are close to the diagonal ($S^{\lambda}_{i,wild} \cong S^{\lambda}_{i,farmed}$) in Figure 5.3a are difficult to classify. This was not unexpected, because both samples refer to sea bass reared in semi-intensive farms, and their flesh composition was indeed close to the wild fish one. As a confirmation of this, note that a correct classification for one of these calibration samples was hard even when chemical variables instead of NIR spectra were used (Fasolato *et al.*, 2010).

Validation results are shown in Figure 5.6 for model #11 of Table 5.3. Among the samples declared wild (Figure 5.6a), 11 out of 32 are correctly recognized as such ($S^{\lambda}_{i,wild} > S^{\lambda}_{i,farmed}$), thus highlighting all the 21 expected substitution frauds. Among the samples declared farmed (Figure 5.6b), two are completely misclassified (samples #22 and #27). On the other hand, the six samples in the lower-left corner of Figure 5.6b (samples from #29 to #34) are very close to the diagonal and, as such, of uncertain attribution. As a confirmation, their projections onto a PCA model calibrated on the 38 known samples exhibit very high model residuals; this indicates that these samples are anomalous with respect to the calibration data, and hence conclusions drawn about them should be taken with caution.

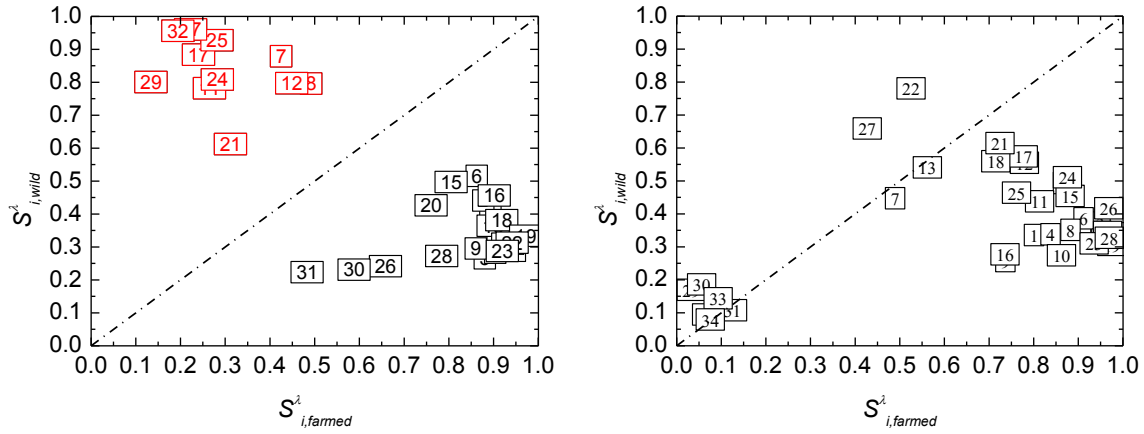


Figure 5.6. *Validation classification results for the European sea bass dataset (model #11 of Table 5.3): (a) declared wild fishes and (b) declared farmed fishes. True wild samples are reported in red; true farmed samples are reported in black.*

Although the algorithm for wavelength selection highlights different wavelength combinations for the 12 models of Table 5.3, there are spectral regions which are

consistently selected throughout all models. These regions comprise the wavelengths between 1640 and 1760 nm and the wavelengths between 2122 and 2270 nm. Since the absorbance of groups -CH, $-CH_2$ and $-CH_3$ is typical of these regions (McClure and Stanfield, 2002), this result indicates that fat and fatty acids are mainly responsible for discriminating between wild and farmed samples, which agrees with previous independent findings (Bell *et al*., 2007; Ottavian *et al*., 2012a).

5.4.1.3 Robustness of the similarity factor-based classification

To examine the robustness of the proposed similarity factor-based classification method to the quality of the available signals, some spectra were artificially altered in such a way as to mimic situations that may be encountered in practice when dealing with near-infrared spectroscopy. It is well know that spectra can be subject to baseline offsets, for example resulting from difficulties in the analysis of solid samples (e.g. due to particle size and non-homogeneity), or from aging of the NIR apparatus (e.g. aging of the light source). Spectra can be also affected by noise, particularly in spectral regions characterized by higher absorbance or for which the integration time of the instrument is not optimized (for example, when the visible and near-infrared regions are scanned simultaneously, the spectra collected may be more noisy within one of the regions).

Two spectra alterations were generated artificially on the sea bass database, corresponding to two different scenarios to be tested, and a classification model based on similarity factors was designed for each scenario:

- Scenario 1: 10 (out of the available 38) randomly selected calibration spectra were shifted of 0.4 absorbance units, resulting in a calibration dataset with the two classes completely overlapped;
- Scenario 2: all the available 106 spectra were corrupted with white noise, where the noise sequence (one for each spectrum) was randomly generated from a normal distribution with zero mean and standard deviation equal to 0.01 absorbance units.

As for Scenario 1, note that a derivative operation in spectra preprocessing could effectively remove by itself the effect of baseline shift. However, to test the robustness of the proposed similarity factor-based classification procedure, spectra were not preprocessed at all. Matrix **X** was subject to external autoscaling, but it was verified that the results obtained with internal autoscaling were basically the same. Validation results for Scenario 1 are shown in Figure 5.7a-b.

Also in Scenario 2 spectra were not preprocessed. However, to avoid amplifying noise, autoscaling of **X** was not performed, and this is recommended any time the available spectra are corrupted by high-level noise. Note a denoising action is always obtained thanks to the removal of the details at the first resolution scale in the DWT of the spectral signals. Validation results for Scenario 2 are shown in Figure 5.7c-d.

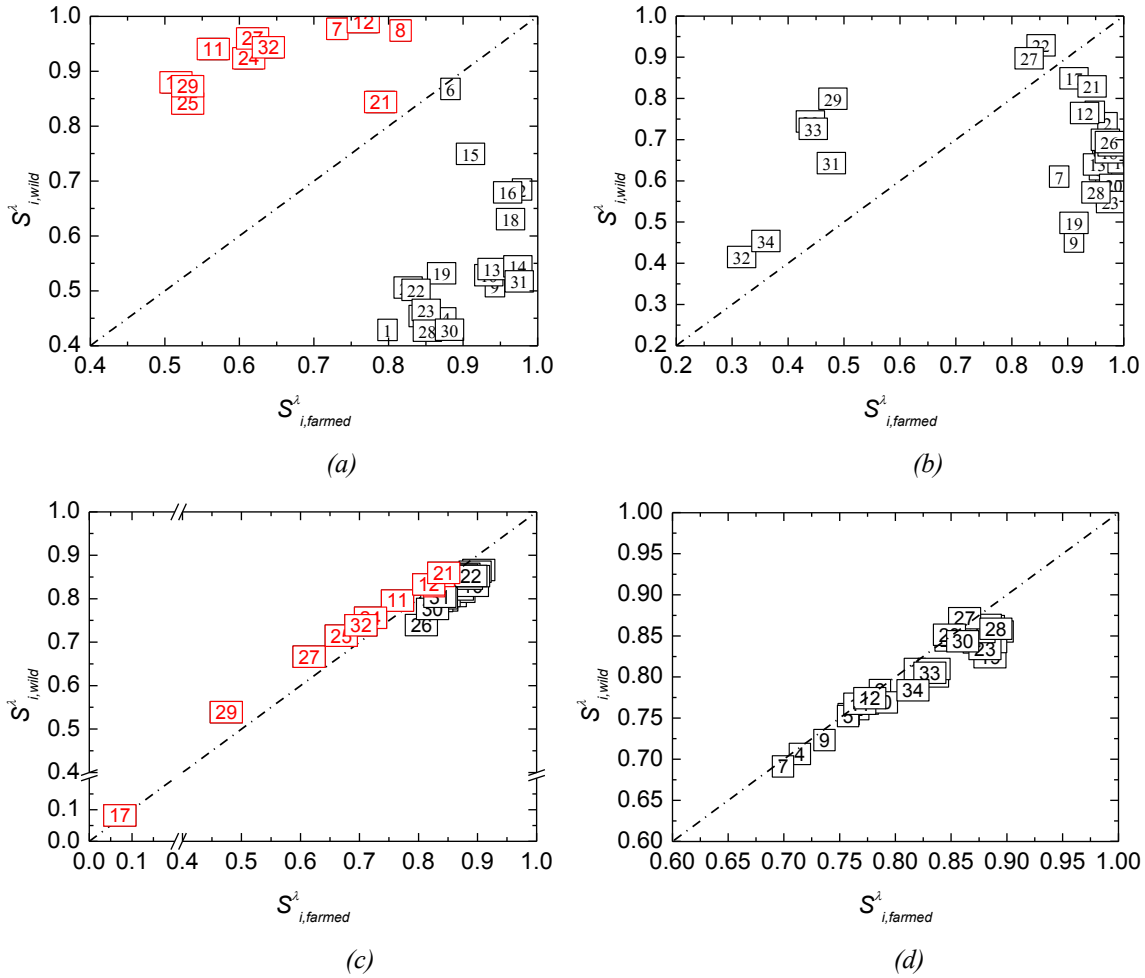


Figure 5.7. *Validation classification results for the European sea bass dataset when spectra are artificially altered with (a) and (b) shift, and (c) and (d) noise. (a) and (c) refer to declared wild fishes, and (b) and (d) refer to declared farmed fishes. True wild samples are reported in red; true farmed samples are reported in black. Note that axis scales are modified to improve readability.*

In all cases presented in Figure 5.7, the results are consistent to those found when the spectra were not altered (Figure 5.6), confirming that the proposed classification method is robust to spectra shift and noise, even without preliminary preprocessing of the available spectra with standard operations. The only major impact of the absence of spectra preprocessing is a shift of the clusters toward the (1,1) point of the (S_{i1}^{λ}; S_{i2}^{λ}) region, which makes the separation between classes somewhat less marked. In Figure 5.7a-c (declared wild samples) all the expected 21 substitution frauds are detected, with the 11 samples classified as wild being exactly the same as in Figure 5.6a. As for the declared farmed (Figure 5.7b-d), samples #22 and #27 are misclassified in both the scenarios considered, whereas for samples numbered from #29 to #34 the consideration drawn earlier still holds.

For comparison, Table 5.4 presents the classification results obtained for the noisy dataset with *k*-NN, LDA, QDA and PLS-DA. Savitzky-Golay smoothing and differentiation (with a gap of 15 and a 2nd-order fitting curve; Saviztky and Golay, 1964) were used to smoothen the effect of noise on the spectra before the classification model was designed.

Table 5.4. *European sea bass noisy dataset: effect of different spectra preprocessing on the classification results of k-NN, LDA, QDA and PLS-DA.*

Classifier	Spectra preprocessing	Errors (calibration)	Errors (validation)	
			Declared wild	Declared farmed
k-NN	only smoothing	0	7	12
	1st order derivative	0	7	12
	2nd order derivative	1	4	11
LDA	only smoothing	0	4	13
	1st order derivative	1	6	12
	2nd order derivative	0	4	11
QDA	only smoothing	0	6	13
	1st order derivative	0	8	12
	2nd order derivative	0	4	6
PLS-DA	only smoothing	0	8	12
	1st order derivative	1	6	2
	2nd order derivative	0	3	12

5.4.2 Asiago d'allevo cheese dataset

5.4.2.1 Performance of traditional classification methods

Classification results obtained with *k*-NN, LDA,QDA and PLS-DA are given in Tables 5.5 and 5.6, respectively for the (preprocessed) spectra matrix and the matrix derived from the traditional application of DWT. Similarly to the European sea bass case study, four different alternatives were tested for spectra preprocessing (i.e.: no preprocessing, SNV, SNV-1 and SNV-2). Since in this dataset the classes are unbalanced, validation results have been expressed also in terms of sensitivity and specificity of the LF class. It can be seen that, as for the sea bass dataset, the classification performance is generally good (with the exception of QDA that returns poor specificity values), but the most appropriate preprocessing method cannot be set *a priori*.

Table 5.5. *Asiago d'allevo cheese dataset: effect of different spectra preprocessing on the classification results of k-NN, LDA, QDA and PLS-DA classifiers applied on the (preprocessed) spectra matrix.*

Classifier	Spectra preprocessing	Errors (calibration)	Errors (validation)	Sensitivity	Specificity
k-NN	no preprocessing	0	2	0.97	1.0
	SNV	0	2	0.98	0.50
	SNV-1	0	9	0.85	1.0
	SNV-2	1	4	0.93	1.0
LDA	no preprocessing	0	4	0.93	1.0
	SNV	0	7	0.88	1.0
	SNV-1	1	5	0.92	1.0
	SNV-2	0	3	0.95	1.0
QDA	no preprocessing	0	3	0.97	0.50
	SNV	1	3	0.97	0.50
	SNV-1	0	2	0.98	0.50
	SNV-2	0	7	0.88	1.0
PLS-DA	no preprocessing	2	6	0.90	1.0
	SNV	1	6	0.90	0.50
	SNV-1	0	3	0.95	1.0
	SNV-2	1	2	0.97	1.0

Table 5.6. *Asiago d'allevo cheese dataset: effect of different spectra preprocessing on the classification results of k-NN, LDA and QDA applied on matrix derived from DWT.*

Classifier	Spectra preprocessing	Errors (calibration)	Errors (validation)	Sensitivity	Specificity
k-NN	no preprocessing	2	0	1.0	1.0
	SNV	0	2	0.98	0.50
	SNV-1	1	1	0.98	1.0
	SNV-2	2	1	0.98	1.0
LDA	no preprocessing	0	5	0.92	1.0
	SNV	0	7	0.88	1.0
	SNV-1	0	2	0.97	1.0
	SNV-2	1	3	0.95	1.0
QDA	no preprocessing	0	2	0.98	0.50
	SNV	1	5	0.92	1.0
	SNV-1	2	11	0.85	0.0
	SNV-2	1	2	0.97	1.0

5.4.2.2 Classification through similarity factors

Table 5.7 presents the classification results for the calibration dataset using similarity factors.

Table 5.7. *Asiago d'Allevo cheese calibration dataset: effect of different spectra preprocessing, input matrix scaling and class references on the classification results obtained with the similarity factors.*

Spectra preprocessing	X scaling	Class reference	d_{tot}	S^{λ}_{12}	Errors (calibration)	Errors (validation)	Sensitiv.	Specif.	Model #
no preprocessing	no autoscaling	mean spectrum	67.8	0.98	3	7	0.88	1.0	1
		whole class	82.6	0.99	9	14	0.77	1.0	2
best preprocessing (SNV & 1st order derivative)		mean spectrum	66.5	0.97	3	4	0.93	1.0	3
		whole class	92.5	0.99	12	11	0.82	1.0	4
no preprocessing	internal autoscaling	mean spectrum	53.7	0.98	0	3	0.95	1.0	5
		whole class	57.5	0.99	0	4	0.93	1.0	6
best preprocessing (SNV)		mean spectrum	53.7	0.99	0	2	0.97	1.0	7
		whole class	58.3	0.99	0	3	0.95	1.0	8
no preprocessing	external autoscaling	mean spectrum	52.8	0.97	0	3	0.95	1.0	9
		whole class	56.7	0.98	1	1	1.0	0.50	10
best preprocessing (SNV)		mean spectrum	49.1	0.97	0	2	0.98	0.50	11
		whole class	57.2	0.99	3	2	0.97	1.0	12

As a general comment to Table 5.7, note that the similarity factor S^{λ}_{12} is close to 1 for any combination of the explored preprocessing techniques, which indicates that the two classes under discrimination (AF vs. LF) are much more similar to each other than those considered in the European sea bass case study. Therefore, generally speaking, sample classification using similarity factors is harder in this case than in the previous one.

Table 5.7 confirms that autoscaling the **X** matrix improves the classification results, with external autoscaling being marginally better than internal autoscaling. Remarkably, in none of the combinations considered in Table 5.7 does preliminary spectra preprocessing improve the classification results.

For calibration model #9 of Table 5.7, results are presented in Figure 5.8. Only 3 out of 62 samples of the validation set (Figure 5.8b, closed symbols) are misclassified. The overall classification performance of the similarity-factor based method is not inferior to that of the traditional methods.

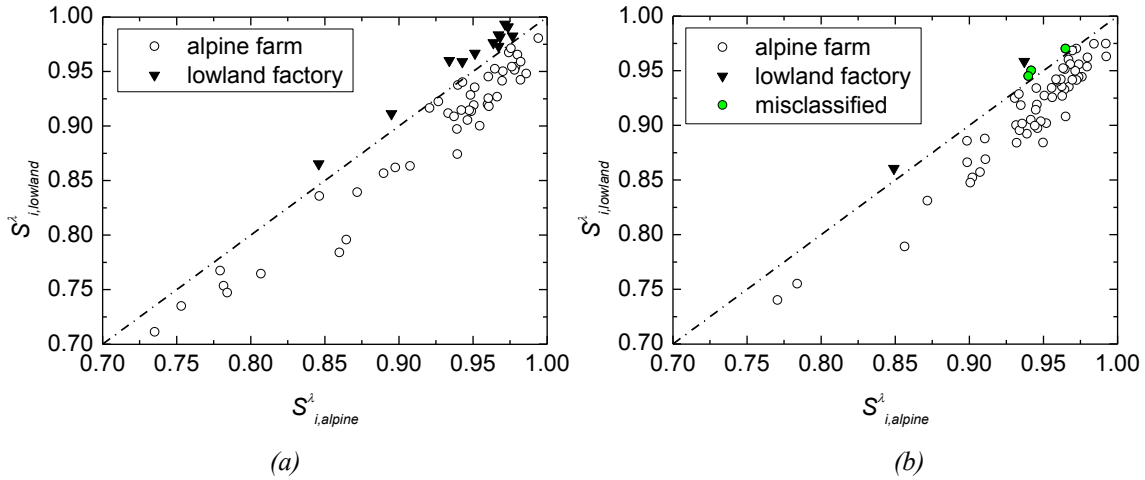


Figure 5.8. *Classification results for the Asiago d'allevo cheese dataset (model #9 of Table 5.7). (a) Calibration and (b) validation spectra. The closed green circles in (b) indicate the three misclassified validation samples.*

The proposed wavelength selection procedure returns spectral regions related to the absorbances of -CH, -CH_2 and -CH_3 groups (wavelengths between 1622 and 1680 nm, and 2152 and 2362 nm; McClure and Stanfield, 2002), indicating that the amount of fat and fatty acids discriminates between the two types of production. Furthermore, the wavelength regions between 1502 nm (N-H stretch; Lucas *et al*., 2008) and 1530 nm were consistently selected throughout all the models. This agrees with the fact that some chemical properties (like protein content and amount of water-soluble nitrogenous compounds) that are related to the selected wavelengths can be shown to be statistically different between the two classes, most likely due to a different effect of proteolysis in the two cheese production systems.

5.4.3 Multispectral images

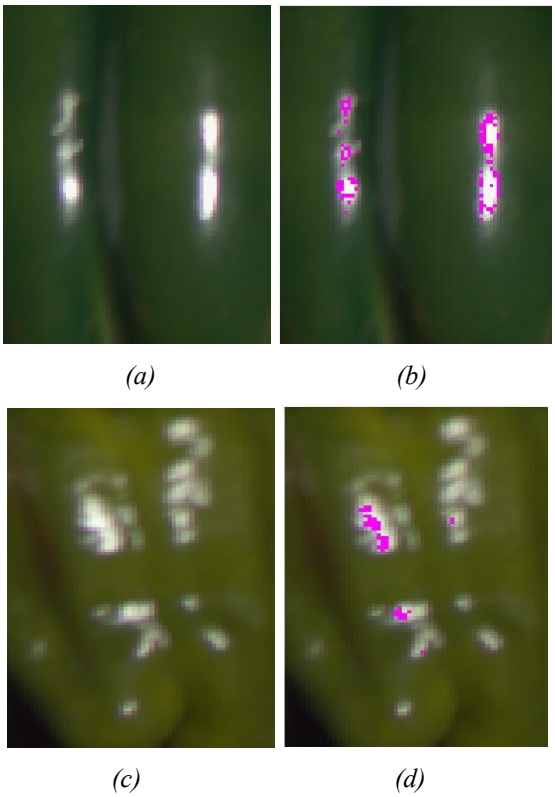
In this section, the proposed similarity factor-based method is extended to the classification of images. The objective is showing that the proposed approach can be easily adapted to also tackle the problem of multispectral (and possibly hyperspectral) data classification.

Using the available sets of strawberry/pepper multispectral benchmark images, four classification models were designed: one model was intended to separate fake strawberries from real ones, while the other three models were designed for the same classification exercise for each pepper color. No spectra pretreatment was carried out in any of the models developed, and the sample matrix **X** was autoscaled externally. For any given model, the fake and real reference matrices were obtained by averaging the calibration spectra along the wavelength direction. The similarity between a validation sample and the reference of each class (fake or real) was checked on a pixel-by-pixel basis; for example, a pixel was deemed to be fake if $S^{\lambda}_{i,fake} > S^{\lambda}_{i,real}$ resulted, i.e. if it showed a greater similarity toward the fake reference than toward the real one. By repeating this operation for all the

image pixels, the total fraction of pixels that is deemed fake was determined: if this fraction largely exceeded 0.50 (say, more than 70% of the pixels are deemed fake), the product (fruit of vegetable) was classified as fake, otherwise it was classified as real.
The results obtained for each product are summarized in Table 5.8. The proposed approach does an excellent classifications job: although limited pixel areas may be misclassified, all the strawberry and pepper validation images are classified correctly. Figure 5.9 clarifies that, in the RGB version of the images, pixel misclassifications are mostly concentrated in the image areas where strong reflection is present (see Figure 5 for further comparison).

Table 5.8. *Classification results for the multispectral image validation dataset.*

Product	Actual class	% of pixels classified as "fake"	Classification
Strawberry	Fake 1	94.6	correct
	Fake 2	90.9	correct
	Fake 3	88.9	correct
	Real 1	2.3	correct
	Real 2	4.5	correct
	Real 3	4.6	correct
Pepper	Fake red	99.1	correct
	Fake green	98.2	correct
	Fake yellow	100.0	correct
	Real red	6.1	correct
	Real green	0.6	correct
	Real yellow	0.0	correct



(a) *(b)*

(c) *(d)*

Figure 5.9. *Location of the misclassified pixels for the green pepper case study (magnified RGB sub-images). (a) and (c) fake and real pepper. (b) and (d) fake and real pepper with misclassified pixels colored in magenta.*

5.5 Conclusions

In this Chapter, a novel algorithm for classification of spectral data based on similarity factors derived from PCA models has been presented. A multi-resolution approach based on discrete wavelet transform was used to associate a matrix (to be summarized with PCA) to each spectrum. In particular, the reconstructed details were employed to build the sample matrix, neglecting those at the first resolution scale in order to accomplish a simultaneous signal denoising action. The classification of a sample was obtained by measuring the similarity of its PCA model to the PCA model of each class reference, and by then assigning the sample to the class for which it shows the greatest similarity. Since the way similarity factors were used intends to quantify the similarity of spectra shapes, the proposed approach is effective if the classes that need to be separated are mainly characterized by spectra of different shapes.

Besides from the graphical interpretation of the classification results, which is very intuitive, the main advantage of the proposed strategy lays in the fact that the design of the classification model can be made automatic, with no need of trial-and-error steps to select the most effective pretreatment to be applied to the raw spectral data. The parameters used in the classification strategy, in fact, can be assigned *a priori* through a series of simple

guidelines. Coupling the proposed classification strategy to a forward-based wavelength selection procedure can return automatically the subset of wavelengths that result most meaningful for sample classification, which facilitates the interpretation of results.

The effectiveness of the proposed approach was demonstrated by means of three case studies. Two of them relate to the use of NIR spectral data for food authentication, the problems having been the discrimination between farmed and wild European sea bass, and the discrimination between Asiago d'allevo cheese produced in alpine farms or lowland factories. The third case study relates to the extension of the proposed method to the classification of multispectral images. In all cases, the classification performance was shown to be very satisfactory, and not inferior to that obtained with conventional methods (*k*-NN, LDA, QDA and PLS-DA).

Chapter 6

Fast fraud detection using spectral data*

In this Chapter, the use of near-infrared spectroscopy coupled with LVM as a fast, non destructive, reliable and cost-effective tool to assist product quality characterization is discussed, with particular attention to the fraud detection problem in the food sector.
The Chapter is made by three Sections. In Section 1, two applications are presented. In the first one, NIRS is used to detect a common commercial fraud for European sea bass, i.e. the substitution of wild fish with farmed fish. This is a fraud that damages the consumers from an economical point of view. In the second application, NIRS is used to assist the authentication and correct labeling of Asiago d'allevo, a typical Italian cheese. In Section 2 of the Chapter, a strategy is proposed to overcome an issue frequently encountered in the food technology literature, i.e. the need of tailoring an anti-fraud system on the specific case investigated. Stemming from the problem of detecting the fraud of substitution of fresh fish with frozen-thawed one, several LVM-based strategies are tested and critically compared with the aim of developing a multi-species/species-independent authentication model. In Section 3 of the Chapter some general conclusions are drawn.

6.1 NIRS for product quality characterization

6.1.1 Problem 1: authentication of wild European sea bass

6.1.1.1 Problem statement

Assessment of seafood origin is a security measure to protect consumers and avoid fraud. Mandatory information required for a full characterization of the marketed fish (species membership, whether wild or farmed, geographic origin) are regulated by stringent laws in the European Union (Commission Regulation No. 2065/2001). Regulatory interventions

*Ottavian, M., P. Facco, L. Fasolato, E. Novelli, M. Mirisola, M. Perini and M. Barolo (2012). Use of Near-Infrared Spectroscopy for Fast Fraud Detection in Seafood: Application to the Authentication of Wild European Sea Bass (*Dicentrarchus labrax*). *J. Agric. Food Chem.*, **60**, 639-648.

Ottavian, M., P. Facco, M. Barolo, P. Berzaghi, S. Segato, E. Novelli and S. Balzan (2012). Near-Infrared Spectroscopy to Assist Authentication and Labeling of Asiago d'Allevo Cheese. *J. Food Eng.*, **113**, 289-298.

Ottavian, M., L. Fasolato, P. Facco and M. Barolo (2013). Foodstuff Authentication from Spectral Data: Toward a Species-Independent Discrimination Between Fresh and Frozen-Thawed Fish Samples. *J. Food Eng.*, **119**, 765-775.

aim at avoiding mislabeling or substituting wild fish with farmed fish, and mitigating risks for the consumers' confidence and health. Therefore, the development of novel analytical technologies, as well as the improvement of the existing ones, can be very helpful to detect fraud in the seafood industry. In particular, the discrimination between wild seafood and farmed seafood is of paramount importance in order to achieve satisfactory quality standards.

Several techniques have been proposed in the last decade to detect the wild/farmed substitution fraud in seafood (Martinez *et al*., 2009). A macroscopic examination of fish is of limited value due to the lack of specific targets in terms of body integrity and loss of morphologic traits at sale time. Other types of analysis, such as genomics and proteomics patterns, present limited application because the selection of reliable markers is very difficult for different populations. Currently, the most informative methodology for discriminating between wild and farmed fish is the determination of fatty acids (FAs) fingerprinting and the ratios isotopes of carbon C and nitrogen N ($^{13}C/^{12}C$ and $^{15}N/^{14}N$, expressed as $\delta^{13}C$ and $\delta^{15}N$) (Martinez, 2006; Bell *et al.*, 2007; Morrison *et al.*, 2007). Both fingerprints vary in the muscles according to the season, the feeding status and the species, but some specific target could be adopted as markers of the production system (Bell *et al.*, 2007). For example, Alasalvar and coworkers (Alasalvar *et al.*, 2002) suggested that a high arachidonic amount could be a marker for wild fish. In farmed fish, the plant oil intake leads to an increase in C18 FA in muscle lipids, particularly 18:2n-6 (linoleic acid), 18:3n-3 (α-linolenic acid), and 18:1n-9 (oleic acid), with the flesh of marine fish retaining these FAs even after a re-feeding period with fish oil (Montero *et al.*, 2005). However, all the methods used to characterize FAs require sample preparation for lipid extraction and gas chromatography analysis, which are expensive and time-consuming compared to the shelf-life of the fish product.

Several emerging technologies have been proposed for the rapid and non-destructive analysis of fish traceability and authentication (Xiccato *et al.*, 2004; Arvanitoyannis *et al.*, 2005; Masoum *et al.*, 2007; Rezzi *et al.*, 2007), such as nuclear magnetic resonance (NMR), front-face fluorescence spectroscopy, and near-infrared spectroscopy (NIRS). Among the most promising techniques, high resolution nuclear magnetic resonance (HR-NMR) was successfully applied to obtain spectral information on the classification of wild/farmed case especially on fish lipids (Masoum *et al.*, 2007; Rezzi *et al.*, 2007). This technique provides a fingerprint of FA profiles linked to other characteristics, such as the positional distribution of polyunsaturated FA on triglycerides (Aursand *et al.*, 1995). HR-NMR spectra combined with different chemometric strategies were used for classification purposes on different species, but this technique is not widely utilized for seafood authentication due to problems in the standardization of the procedures (Martinez et al., 2009; Rezzi *et al.*, 2007; Forshed *et al.*, 2003).

NIRS is particularly favorable because it is simpler, more economical, environmentally safer and faster than many other techniques. Recent studies have highlighted the potential of NIRS to differentiate sea bass (*Dicentrarchus labrax*) from different rearing systems (Xiccato *et al.*, 2004; Majolini *et al*., 2009; Costa *et al*., 2010). In these applications, the classification ability among rearing systems seems to decrease according to the storage time, and typically some additional treatments such as freeze-drying are required to improve the accuracy of origin prediction.

In this Section, the possibility to use NIRS to discriminate between wild and farmed sea bass samples was investigated. Three different chemometric approaches were developed to process the available NIR spectra. In the first approach, a cascaded arrangement was proposed where chemical information was first estimated from spectra using latent variables regression, and this estimated information was then used to build the discrimination model in a latent variable space. The second approach used spectral information to directly build the discrimination model in a latent variable space. The third approach first used wavelets to transform the spectral information, and subsequently developed the discrimination model using the transformed spectra. The classification results were compared to a reference classification obtained using only chemical and morphometric information.

6.1.1.2 Materials and methods

European sea bass collection

Farmed and wild fish was collected in different sales centers and different cities in 2008 (Fasolato *et al.*, 2010). Samples were transported to the laboratory within 24 h from the collection time at refrigerated temperature (4 ± 1 °C, constantly monitored by a data logger Testo 174-T, Testo AG, Germany), and were immediately processed for the analysis upon their arrival at the laboratory. The dataset comprised 38 calibration samples with determined attribution of production method and 66 validation samples with declared methods of production (32 declared wild and 34 declared farmed). Compared to the dataset presented by Fasolato and coworkers (Fasolato *et al.*, 2010), two samples were removed since the corresponding spectra had not been collected. The same study (which was used as a classification reference in the present work) showed that the number of samples classified as farmed among those declared wild was 22 (corresponding to 69% of substitution fraud) and, at the same time, 6 samples among those declared farmed were formally ascribed to the wild group (i.e. misclassified).

A total of 35 chemical properties (fatty acids, bromatological and isotopes) and morphometric traits were measured for each available sample. These variables are listed in Table 6.1, and will be collectively identified as “chemistry” variables in the remaining.

Table 6.1. *List of measured chemical properties and morphometric traits.*

Property #	Property Name
1	fat
2	protein
3	ash
4	moisture
5	C14:0
6	C16:0
7	C18: 0
8	Σsaturated
9*	C16:1 n-7
10	C18:1 n-9
11	C18:1 n-7
12	C20:1 n-9
13*	C22:1 n-11
14*	C22:1 n-9
15	Σmonounsaturated
16	C18:2 n-6
17*	C18:3 n-6
18	C18:3 n-3
19	C20:2 n-6
20*	C20:3 n-6
21	C20:3 n-3
22	C20:4 n-6
23	C20:5 n-3
24	C22:5 n-3
25	C22:6 n-3
26*	Σpolyunsaturated
27	Σn-3
28	Σn-6
29	n-3/n-6
30	EPA + DHA
31†	$\delta^{13}C$
32	$\delta^{15}N$
33¥	KI
34¥	HSI
35¥	CFI

* = not significant according to the VIP index (VIP < 0.5)
† = measured from fat-free extract
¥ = morphometric traits (KI = condition index, HSI = hepatosomatic index and CFI = celomatic fat index, cf. Fasolato *et al.*, 2010)

NIR analysis

After dissection, the epiaxial white muscle portion of the fillet was ground with a Retsch Grindomix (Retsch GmbH, Hann, Germany) at 4000 rpm for 10 s. Two aliquots per sample (approximately 10 g each) were placed in a 50 mm diameter ring cup and scanned in reflectance mode at 2 nm intervals from 1100 to 2500 nm using a scanning monochromator NIRSystem 5000 (FOSS, Silver Spring, MD, USA).

Spectra pretreatment

For each aliquot of a sample, a mean spectrum was obtained by averaging from 32 multiple scans; then, the spectrum of the sample was obtained by averaging those of the two aliquots. Reflectance (R) values were converted into absorbance (A) values through $A = \log(1/R)$. Mathematical pretreatment reduced the light scattering caused by the sample particles and removed the additional variation in baseline shift typically present in diffused reflectance spectra. Standard normal variate and first- and second-order derivates were used to this purpose.

Data analysis

PCA was applied as a preliminary exploratory tool to reveal the internal structure of the available data and to check whether the validation samples could be described by the model developed for the calibration samples. After this preliminary analysis, four different strategies for sea bass classification were developed and tested. These strategies are sketched in Figure 6.1, where the acronyms used to identify the proposed models are also indicated. Figure 6.1 clarifies that the strategies differed for the input information as well as for the type of classification model.

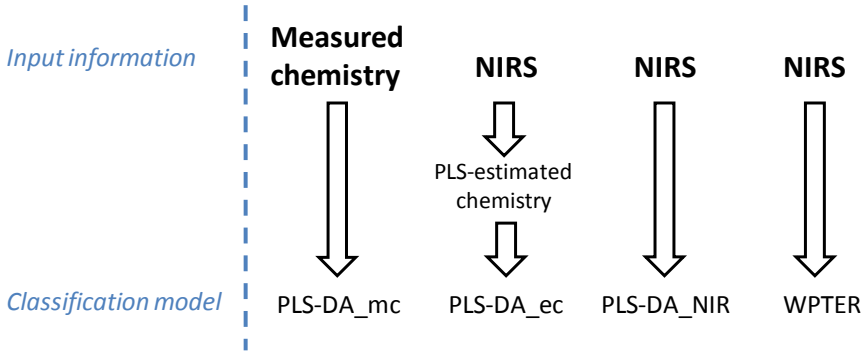


Figure 6.1. *Schematic of the classification strategies considered in this study.*

As far as the model inputs are concerned, either chemistry variables or NIR spectra were used. Note that, with respect to the chemistry inputs, not only the *measured* properties, but also the properties *estimated* from NIRS were used as inputs to the classification model. When NIR spectra were used as inputs directly, two alternative modeling approaches were investigated: PLS-DA (Chapter 2, Section 2.1.2.2) and the wavelet-based WPTER method (wavelet packet transform for efficient pattern recognition; Cocchi *et al.*, 2001).

6.1.1.3 Results and discussion

Exploratory PCA model

The scores plot of the first 2 PCs (principal components) of a 4-PC model on the full set of measured chemistry variables of the calibration dataset (Figure 6.2) clearly identified two

clusters (open triangles vs. closed triangles), which include the farmed and wild sea bass samples, respectively. The separation between the clusters occurred along PC1 (PC1 > 0 for wild samples). When projected onto the PCA model, the validation samples (circles) separated into the same clusters. However, quite a number of declared wild samples (closed circles) fell within the farmed sea bass cluster (open symbols), i.e. a number of substitution frauds were highlighted in the scores plot. Although not reported here, it was verified that all validation samples fell well below the 95% confidence limit of the *Q* statistic, which indicated that, from the point of view of the chemistry variables, the validation dataset completely conformed to the calibration one.

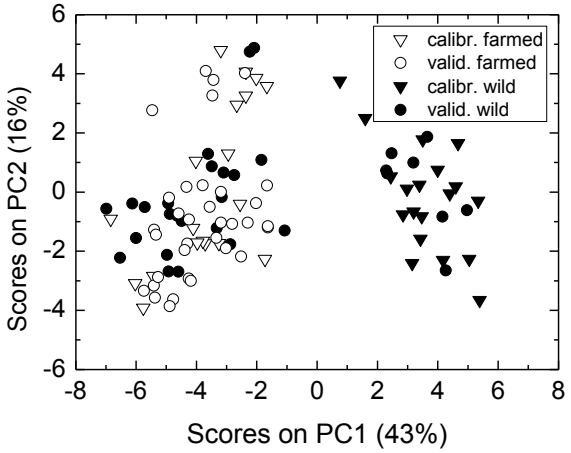


Figure 6.2. *Exploratory PCA analysis on the measured chemistry properties of Table 6.1. The validation samples (circles) are projected onto the model defined by the calibration samples (triangles).*

Although not included here, the combined analysis of the scores and loadings of the PCA model revealed that the farmed samples were characterized by higher levels of fat, protein, ash, celomatic fat index (CFI), hepatosomatic index (HIS) and, among the fatty acids listed in Table 1, C18:3 n-6, C18:2 n-6, C18:1 n-9, Σn-6, C20:3 n-3, Σmonosaturated, C20:2 n-6 and C14:0. The wild samples, instead, were characterized by higher level of moisture, $\delta^{13}C$ and $\delta^{15}N$ and, among fatty acids, Σsaturated, C16:0, C18:1 n-7, C18:0, C20:4 n-6, EPA+ DHA, Σn-3 and C22:5 n-3. These results were in agreement with other studies on sea bass (Bell *et al.*, 2007; Alasalvar *et al*., 2002).

A similar exploratory analysis was carried out on NIR spectra. The average spectra for the farmed and wild classes of the calibration dataset have been shown in Chapter 5, Figure 5.4a. The scores plot and SPE values for the exploratory PCA model (using 5 PCs) built on the NIR spectra of the calibration dataset are shown in Figure 6.3. Figure 6.3a indicated that the first PC is not able to separate the two classes, although it explains a much higher fraction (72% vs. 43%) of the data variability than in the PCA model derived on the measured chemistry data.

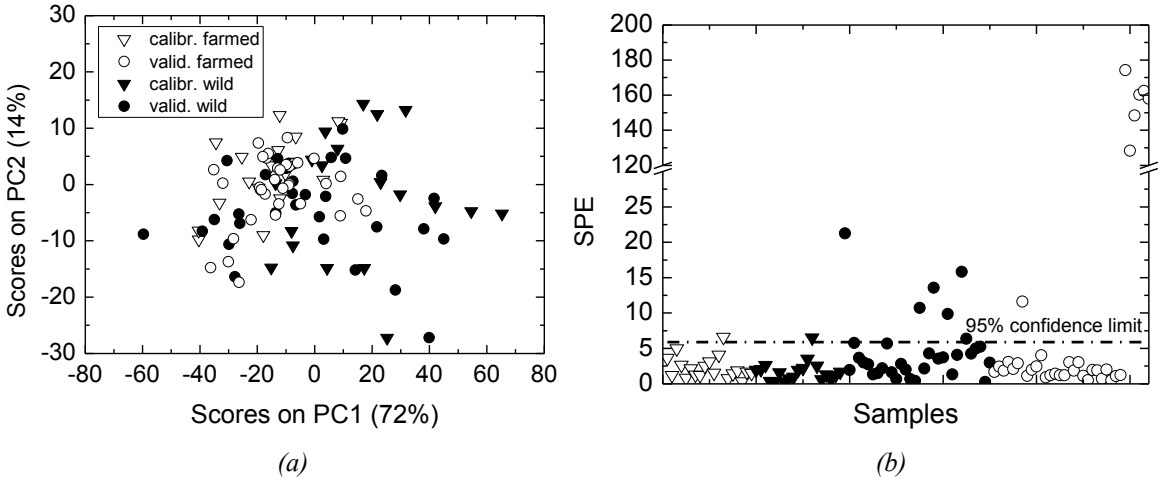


Figure 6.3. *Exploratory PCA analysis on NIR spectra: (a) scores plot and (b) SPE values. The validation samples (circles) are projected onto the model defined by the calibration samples (triangles).*

Furthermore, some of the declared farmed samples exhibited very high model residuals (rightmost open circles in Figure 6.3b), indicating that the corresponding NIR spectra did not conform to those of the calibration dataset. These samples were assigned the IDs #29 to #34. Although a detailed discussion on these samples is beyond the purpose of this work, it was worth noticing that the high model residuals indicate that these samples explored a variability on NIR spectra that is different from that described by the calibration dataset.

Authentication using measured chemistry variables (PLS-DA_mc)

A PLS-DA model using one latent variable was built using the full measured chemistry calibration dataset to discriminate between farmed and wild sea bass samples. The VIP index is plotted for the most influential (VIP > 1) inputs of this model in Figure 6.4.

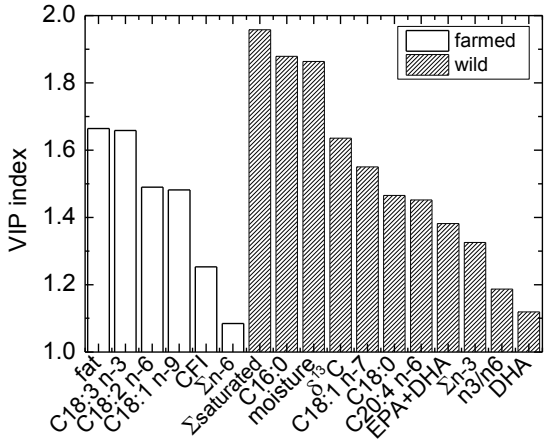


Figure 6.4. *Most influential (VIP > 1) chemistry variables for the PLS-DA_mc classification model according to the VIP index.*

The input variables were classified as belonging to the farmed class or to the wild one according to the indications provided by the exploratory PCA model loadings. In terms of meaningfulness of the input variables for each class, the results of Figure 6.4 agreed with those obtained using a nonparametric permutation test to analyze the same dataset (Fasolato *et al.*, 2010). Furthermore, the variables deemed as totally not significant by the VIP index (VIP < 0.5, see Table 6.1) were the same as those discarded by the nonparametric test.

As mentioned earlier, a more parsimonious model can be developed by using a subset of the chemistry variables having VIP > 1 as inputs to the model. To this purpose, a subset including only 3 (namely: fat, moisture and $\delta^{13}C$) out of the 17 the variables shown in Figure 6.4 was selected. The rationale for this selection was as follows. First, according to VIP, the selected variables were among the most discriminating ones of the two classes; secondly, these variables could be estimated with sufficient accuracy from NIR spectra (to be discussed later). A PLS-DA model (1 latent variable) using only these 3 measured chemistry variables as inputs was therefore developed for sea bass authentication; this model was denoted as PLS-DA_mc. Although this model had the same classification ability of a PLS-DA model using the full dataset of measured chemistry, reducing the number of inputs from 35 to 3 significantly reduced the distance between the clusters in the scores plot of the PCA exploratory model. This means that the observable variability among samples was reduced when the number of model inputs was reduced. However, the variability described by the PCA model still allowed the two classes to be separated.

A comparison between the reference classification results provided by Fasolato and coworkers (Fasolato *et al.*, 2010) and those obtained using the PLS-DA_mc model is shown in Table 6.2 for declared wild samples and in Table 6.3 for declared farmed samples. The two approaches detected the same total number of substitution frauds (Table

6.2), although with one sample classification shift (samples #6 and #12). On the other hand, the PLS-DA_mc model did not misclassify any of the declared farmed samples (Table 6.3).

Authentication using estimated chemistry variables (PLS-DA_ec)

Fat, moisture and $\delta^{13}C$ (i.e., the inputs to the PLS-DA_mc model) were estimated from NIR spectra using three distinct PLS models. The estimation results are shown graphically in Figure 6.5, while the estimation model characteristics and performance metrics are reported in Table 6.4. The models developed for fat and moisture returned accurate estimations even when extrapolated outside the calibration range, whereas the model for stable carbon isotope estimation was less accurate.

Table 6.2. *Declared wild samples (validation dataset): comparison between the reference classification results (indicated as Ref.; Fasolato et al., 2010) and the performance of the classification methods considered in this study. For each classification method, symbol ✓ indicates a sample classified as wild. No symbol is used if a sample is declared wild but is classified as farmed (substitution fraud).*

Sample ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Ref.						✓	✓	✓			✓						✓							✓	✓		✓		✓			✓
PLS-DA_mc							✓	✓			✓	✓					✓							✓	✓		✓		✓			✓
PLS-DA_ec							✓	✓			✓	✓					✓							✓	✓		✓		✓			✓
PLS-DA_NIR							✓	✓			✓	✓					✓							✓	✓		✓		✓			✓
WPTER							✓	✓			✓	✓					✓				✓				✓		✓		✓			✓

Table 3. *Declared farmed samples (validation dataset): comparison between the reference classification results (Fasolato et al., 2010) and the performance of the methods considered in this study. For each classification method, symbol X indicates a misclassified sample. No symbol is used if a sample is declared farmed and is classified as farmed.*

Sample ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Ref.								X									X												X		X		X	X
PLS-DA_mc																																		
PLS-DA_ec																					X						X							
PLS-DA_NIR																											X		X†	X†				
WPTER																													X†	X†		X†		

† = sample with high PCA residual

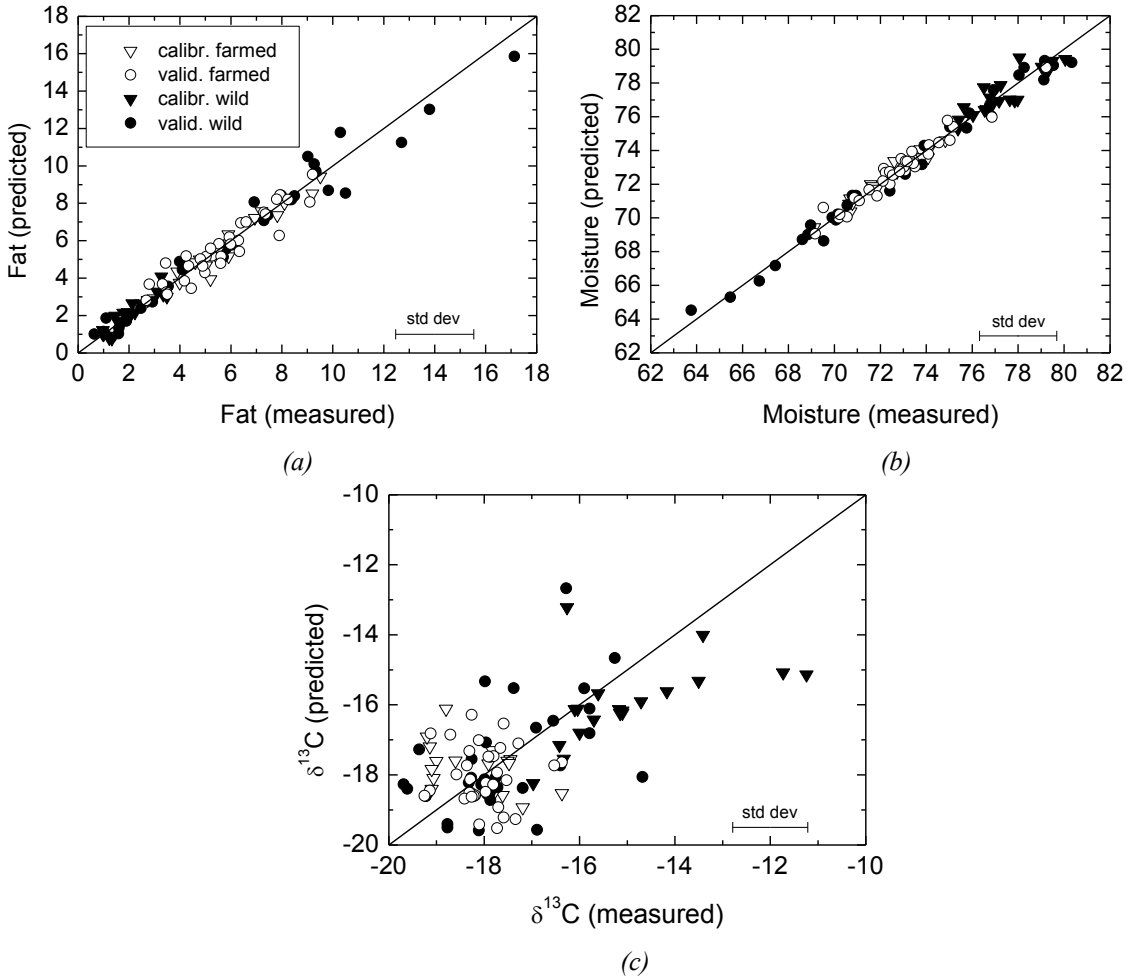


Figure 6.5. *Estimated versus measured values of the three chemistry variables used as inputs to the PLS-DA_ec classification model: (a) fat, (b) moisture and (c) $\delta^{13}C$. The standard deviation (std dev) on the measured values is also indicated.*

A PLS-DA model (PLS-DA_ec) with 1 latent variable was built using the estimated chemistry variables as inputs. Although the estimated value of $\delta^{13}C$ was not very accurate, it was used as an input to the PLS-DA_ec model because the other chemistry variables with VIP greater than that of the $\delta^{13}C$ (namely, C18:3 n-3, Σsaturated and C16:0, cf. Figure 6.4) were estimated even worse (with RPDs of 1.07, 1.20 and 1.05, respectively). The classification results are reported in Table 6.2 and Table 6.3. Despite the fact that one of the model inputs was not estimated as accurately as the other two by the NIR spectra, for the declared wild samples the classification results obtained using the PLS-DA_ec model totally agreed with those obtained using the PLS-DA_mc model, whereas for declared farmed samples only two misclassifications were obtained.

Table 6.4. *Number of latent variables (LVs) retained in each PLS model for the estimation of fat, moisture and $\delta^{13}C$, and mean performance of the estimation models. R^2_p indicates the coefficient of determination in the estimation (i.e. when the samples belongs to the independent validation group).*

	no. of LV	R^2	Q^2	R^2_p	RPD
fat	7	0.98	0.97	0.97	5.69
moisture	9	0.99	0.98	0.98	6.66
$\delta^{13}C$	9	0.67	0.42	0.45	1.25

Authentication with direct use of spectral data (PLS-DA_NIR and WPTER)

NIR spectra were used directly as inputs to a PLS-DA discrimination model (PLD-DA_NIR). Table 6.2 shows that the ability of this model to identify the substitution fraud was very good. Three misclassified farmed samples were present (Table 6.3); however, note that two of them (samples #29 and #30) did not conform to the calibration dataset due to high SPE values, and therefore the use of spectral information for these samples was questionable.

As was noted earlier, the dimension of the calibration input dataset may be extremely large when spectra are used as model inputs. For this reason, an attempt to reduce this dimension by extending the use of the VIP index to spectra was first attempted (Figure 6.6), but this approach did not lead to satisfactory results in term of discrimination ability of the resulting PLS-DA model. We conjecture that this can be ascribed to the much higher correlation existing between the model inputs (i.e. wavelengths) than in the measured chemistry variables case.

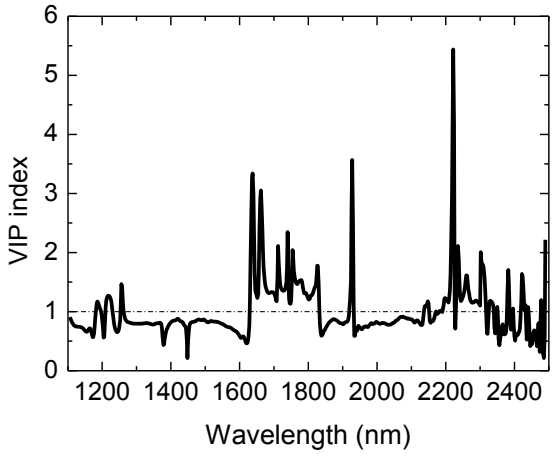


Figure 6.6. *VIP index for the PLS-DA_NIR model.*

As an alternative approach for selective pruning of the input dataset and sample classification, the WPTER algorithm was used (Cocchi *et al*., 2001). Figure 6.7, which was obtained with a WPTER model with Daubechies-2 wavelets, shows the mean reconstructed

signals for both the farmed class and the wild class. WPTER unambiguously showed that the only influential spectral regions for sample classification were those within the range ~ 1600-1750 nm. These regions were somewhat overlapped to those with a VIP index much larger than 1 in Figure 6.6. However note that, due to the reconstruction operation, the actual input signal used by WPTER for sample classification was different from that used by a PLS-DA model using the same intervals of wavelengths as inputs.

The classification results of WPTER were in very good agreement with those provided by the other methods considered in this work. Three misclassified farmed samples were present (Table 6.3), but they referred to validation samples not conforming to the calibration dataset of spectral signals (IDs #29, #30 and #32).

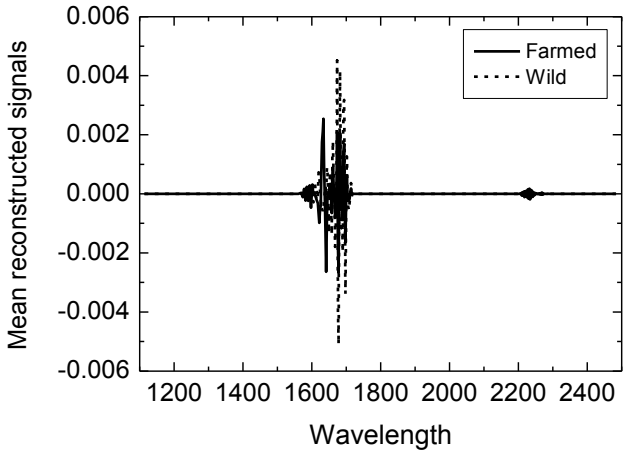


Figure 6.7. *Mean reconstructed signals (in the wavelet domain) according to the WPTER algorithm. Non zero signals indicate the wavelengths selected by the WPTER algorithm.*

A thorough analysis on the meaning of the wavelengths selected as the most informative is beyond the purpose of this Chapter.

6.1.2 Problem 2: authentication of Asiago d'allevo cheese

6.1.2.1 Problem statement

Asiago is a Protected Designation of Origin (PDO) semi-hard cheese of a well-defined geographical area located in the north-eastern regions of Italy, and is manufactured with partially skimmed raw bovine milk according to the Regulation of the European Community as specified in the Official Journal of European Union (1996). It is commonly sold in two forms: "*pressato*", a pressed cheese with a minimum ripening of 20 days, and "*allevo*", the non-pressed variety whose maturation takes from 6 to 18 months or more (with a minimum ripening of 60 days). The height and diameter of allevo cheese blocks vary from 9 to 12 cm and from 30 to 36 cm respectively; the weight typically ranges from

8 to 12 kg at the minimum ripening age. In 2010, 245751 blocks of Asiago were produced from 66 different plants (www.asiagocheese.it). According to the official Italian protocol (Gazzetta Ufficiale, 2006), Asiago d'allevo cheese produced above 600 m a.s.l. is defined and labeled as "*Mountain product*" and as such has a higher economical value.

The cheese inherent quality and perceived characteristics are affected by several parameters, among which the geographical area of manufacturing, milk composition, production environment, and cheese ripening (McSweeney, 2004). It is well known that milk fat from highlands is important from the nutritional point of view, because grazing increases the proportion of healthy compounds in milk (Collomb *et al*., 2003). Moreover, when the cheese is made from raw milk without starter cultures, the contribution of autochthonous fermenting microflora to the flavour is significant and helps to identify a mountain product from a lowland one. The healthiness and the sensory traits of cheese from alpine zones are well appreciated by the consumers, and the higher qualitative characteristics of the mountain cheese determine a higher price of the final marketable product. From this perspective, correct labeling of the product is very important.

A rapid and reliable authentication of the cheese product, including the determination of its geographical origin, its ripening traits as well as the estimation of its chemical properties, would be highly desirable both for the manufacturers and for those involved in the protection of the mark (Consortium) and of consumers' interests (Public Institutions). Note that the possibility of reliably estimating the ripening parameters is of paramount importance both for real-time monitoring of the cheesemaking process and for correct product labeling and traceability. Traditional approaches require sampling and chemical analysis, which are expensive and time consuming. Reducing the cost of the analysis would be important to reduce the overall production costs, that is very stringent specially for smaller manufacturers.

NIRS is a well known analytical technology that satisfies these requirements (Fagan *et al*., 2007). Although NIRS has already been applied successfully to the characterization of different varieties of cheese (Karoui *et al*., 2006; Lucas *et al*., 2008; Gonzáles-Martín *et al*., 2011), few studies have been focused on the Asiago d'allevo cheese, and only one (Cozzi *et al*., 2009) specifically dealt with the use of NIRS. Marchesini *et al*. (2009) used color traits to evaluate the Asiago ripening time. For the authentication of the production chain, Schievano *et al*. (2008) and Cozzi *et al*. (2009) used nuclear magnetic resonance and NIRS, respectively, as an alternative to chemical and color analysis. However, the problem of complete authentication of Asiago d'allevo cheese (i.e., the estimation of chemical composition, ripening age and traits, and production site) using NIRS was not addressed in the above studies.

The aim of the present study was the implementation of NIRS as a nondestructive, fast and relatively low expensive technique for the authentication of Asiago d'allevo cheese. The

performance of the classification models built on spectral data was compared to that of classification models built on chemical composition, which were used as a reference for authentication.

6.1.2.2 Materials and methods

Sampling and cheese analysis

This study considered 121 Asiago d'allevo samples produced in 9 plants of the Vicenza territory (north-eastern Italy). Three plants were industrial factories located in the lowland (60-200 m a.s.l.), whereas the remaining farms were located in the mountain area (> 800 m a.s.l.). The cheese made in the lowland dairies was produced with milk from animals reared intensively and always fed with total mixed ration. For the other cases, as already described by Lignitto *et al*. (2010), at the beginning of the trial cows were kept in barn (May, indoor period) and fed a total mixed ration based on hay and concentrate (maize and soybean). During the second part of the trial, cows were moved to alpine grazing ground (from June to September, outdoor period), kept on the site of the farm and fed on pasture plus a concentrate supplement. During the indoor period, cheese was manufactured in an industrial dairy plant close to the farms, whereas during the outdoor period the cheesemaking activity was carried out in situ in an artisanal dairy plant belonging to the alpine farm.

The sampling campaign and the respective analyses were performed during the ripening time. Asiago d'allevo samples were monitored over a period of 36 months (at 6, 12, 18 and 36 months of ripening) using different whole blocks each time. In addition, three different cheese production periods were monitored, May (indoor period), July (middle of the alpine grazing period) and September (proximity of the end of the summer alpine grazing).

Samples (about 2 kg each) were segregated as a quarter of a whole block and taken to the laboratory under refrigeration (4 °C). Each sample portion was cut across the whole block (height) in order to take into account concentration gradients (Karoui *et al*., 2006). The first 2 cm starting from the rind were discarded while the rest was grated and stored at –80 °C. An exhaustive list of the measured chemical properties is given in Table 6.5.

Table 6.6 shows the sampling schedule adopted for the 107 alpine samples, highlighting three out of the four classification criteria (i.e. production period, ripening age and production height) that will be discussed in Section 6.1.2.3. Production height in Table 6.6 refers to the height of the grazing. Low height ranges between 1100 and 1300 m a.s.l., medium height around 1400 m a.s.l., medium-high height between 1500 and 1750 m a.s.l., and high height between 1700 and 2000 m a.s.l.

Table 6.5. *List of the measured chemical properties.*

Property #	Property Name	Property #	Property Name
1	water activity	28	C14:0
2	pH	29	C14:1
3	dry matter	30	C16:0
4	ash,% (d.m.[§])	31	C16:1
5	protein,% (d.m.)	32	C17:1
6	fat, % (d.m.)	33	C18:0
7	moisture, %	34	C18:1 n-9
8	ash, % (w.w.[¥])	35	C18:1 n-7
9	protein, % (w.w.)	36	C18:2 n-6
10	fat, % (w.w.)	37	C18:3 n-6
11	NaCl, %	38	C18:3 n-3
12	proteolysis index	39	C18:2 c9-t11
13	proteolysis TCA[þ]	40	C18:2 t10-c12
14	Vitamin A (ca[†])	41	C20:0
15	Vitamin E (ca)	42	C20:1 n-9
16	Cholesterol (ca)	43	C20:2
17	Vitamin A (ed[#])	44	C20:3 n-6
18	Vitamin E (ed)	45	C20:4 n-6
19	Cholesterol (ed)	46	C20:3 n-3
20	Vitamin A (fat[ϒ])	47	EPA
21	Vitamin E (fat)	48	Σtot id fame %
22	Cholesterol (fat)	49	Σsaturated
23	C4:0	50	Σmonounsaturated
24	C6:0	51	Σpolyunsaturated
25	C8:0	52	Σn-3
26	C10:0	53	Σn-6
27	C12:0	54	n-3/n-6

[§]d.m. = dry matter; [¥]w.w. = wet weight; [þ]TCA = trichloro acetic acid 12%; [†]ca = per g/sample; [#]ed = per 100 g/sample; [ϒ]fat = per g/fat

NIR analysis

A FOSS Nirsystem 5000 (FOSS Nirsystems, Inc., Laurel MD, USA) instrument was used. The absorbance data in the IR zone 1100-2500 nm were recorded at 2-nm intervals. To minimize the sampling error, all samples were analyzed in duplicate. Thirty-two scans were performed for both the reference and the samples. All of the duplicated spectra were averaged prior to statistical analysis.

Table 6.6. *Sampling schedule for the 107 alpine samples in terms of number of samples collected for three out of the four classification problems considered (production period, ripening age, and production height).*

	Production period	**May**				**July**				**September**				
	Ripening age	**6**	**12**	**18**	**36**	**6**	**12**	**18**	**36**	**6**	**12**	**18**	**36**	**Total**
Height	**Low**	2	2	2	2	4	6	4	2	4	6	4	2	40
	Medium	1	1	1	1	2	3	2	1	2	3	2	2	21
	Medium-high	-	-	-	-	2	3	2	-	2	3	2	-	14
	High	-	-	-	-	4	6	4	2	4	6	4	2	32
	Total	3	3	3	3	12	18	12	5	12	18	12	6	107
		12				47				48				

Data analysis

Two PCA models, one on the measured chemical properties and one on the spectral data, were developed to assess the performance of PCA as an unsupervised classification technique. Standard Normal Variate (SNV) and first- and second-order derivatives were used to preprocess the spectral data.

The chemical properties listed in Table 6.5 were estimated from NIR spectra using PLS regression. To design each PLS model, a calibration dataset ($\mathbf{X}_{cal}$, $\mathbf{Y}_{cal}$) containing only 75% of the data available in the original matrices of spectral data ($\mathbf{X}$) and chemical properties ($\mathbf{Y}$) was used. The samples to be included in the calibration dataset were selected randomly; the number of latent variables (LVs) to be retained in the model was determined by minimizing the root-mean squared error in cross-validation (Wold, 1978). The model estimates were then validated using the remaining 25% of data, which were organized as an independent validation dataset ($\mathbf{X}_{val}$, $\mathbf{Y}_{val}$). Since the number of available samples for each class was not large, depending on the random sample selection the validation dataset might have included only few samples per class (cf. Table 6.6). Therefore, to allow using all the available samples as independent validation data, the procedure described above was repeated in four steps. At each step, the random selection of the calibration samples was made among those samples that had not been selected as calibration ones in the previous steps. The prediction model performance was eventually evaluated by averaging the model performance obtained at each step. The procedure was intended to ensure model robustness.

PLS-DA using either the measured chemical properties or the NIR spectra was used to classify the cheese samples. The classification was carried out to discriminate alternatively the farm management (lowland vs. mountain), the ripening age, the period of production, and the height of milk and cheese production. A procedure similar to the one of the PLS

estimation models was used to design and validate the PLS-DA models. To create an independent validation dataset, the same percentage of samples (25%) was removed from each class in order to take into account that the number of samples varied from class to class. The number of LVs to retain within each model was determined by maximizing the sensitivity (S_e) and specificity (S_p) indices in cross-validation.

6.1.2.3 Results and discussion

Exploratory PCA models

Although it appeared that for this study PCA was not an effective unsupervised classification technique, the PCA models were nevertheless useful to investigate the correlation structure within the available datasets (measured chemical properties and spectra). For both models, results related to only two (ripening age and production period) out of the four possible classifications are presented for the sake of conciseness.

Figure 6.8a-b presents the score plot of the first two PCs for a twelve-PC model calibrated on the measured chemical data for the 107 alpine-farmed samples (the 14 lowland samples were not used for model building). The model explains ~85% of the total variance; the high number of PCs retained in the model was a consequence of the low correlation existing among the measured chemical variables. This aspect was highlighted also by the loading plot of Figure 6.8c, where few clusters of variables were detected (note that variables with both PC1 and PC2 smaller, in absolute value, than 0.1 were not included to improve readability).

Although PC1 and PC2 explained a limited amount of the total variability (38%), the score plot revealed a certain degree of separation among classes. For instance, the ripening age tended to increase along PC1 (Figure 6.8a), whereas samples produced in the first period (May) clustered in the lower-right part of the score plot (Figure 6.8b). With respect to the classification by the height of the production site (not shown in Figure 6.8), it was found that the production height increased along PC2, with samples belonging to the low and medium-height classes having PC2 < 0. Projecting the 14 samples produced in lowland factories onto the PCA model of Figure 6.8, very high SPE values were observed, i.e. these samples represented a clearly distinct class, spanning a different range of variability.

Some additional considerations arose from the combined analysis of model scores and loadings of Figure 6.8. The increase of ripening age, for instance, was highly correlated to lower values of water activity, moisture, Σmonounsaturated and Σpolyunsaturated and to higher percentages of C6:0, C8:0, C10:0, C12:0, C14:0 and Σsaturated (Malacarne *et al.*, 2006). Samples produced in May (first period) were characterized by a higher amount of C16:0 (and in general of saturated fatty acids) and by lower contents of C18:1 n-7 and vitamin E. On the other hand, the discrimination between samples produced in July and

September (second and third period) was more difficult to analyze, as suggested also by Figure 6.8b.

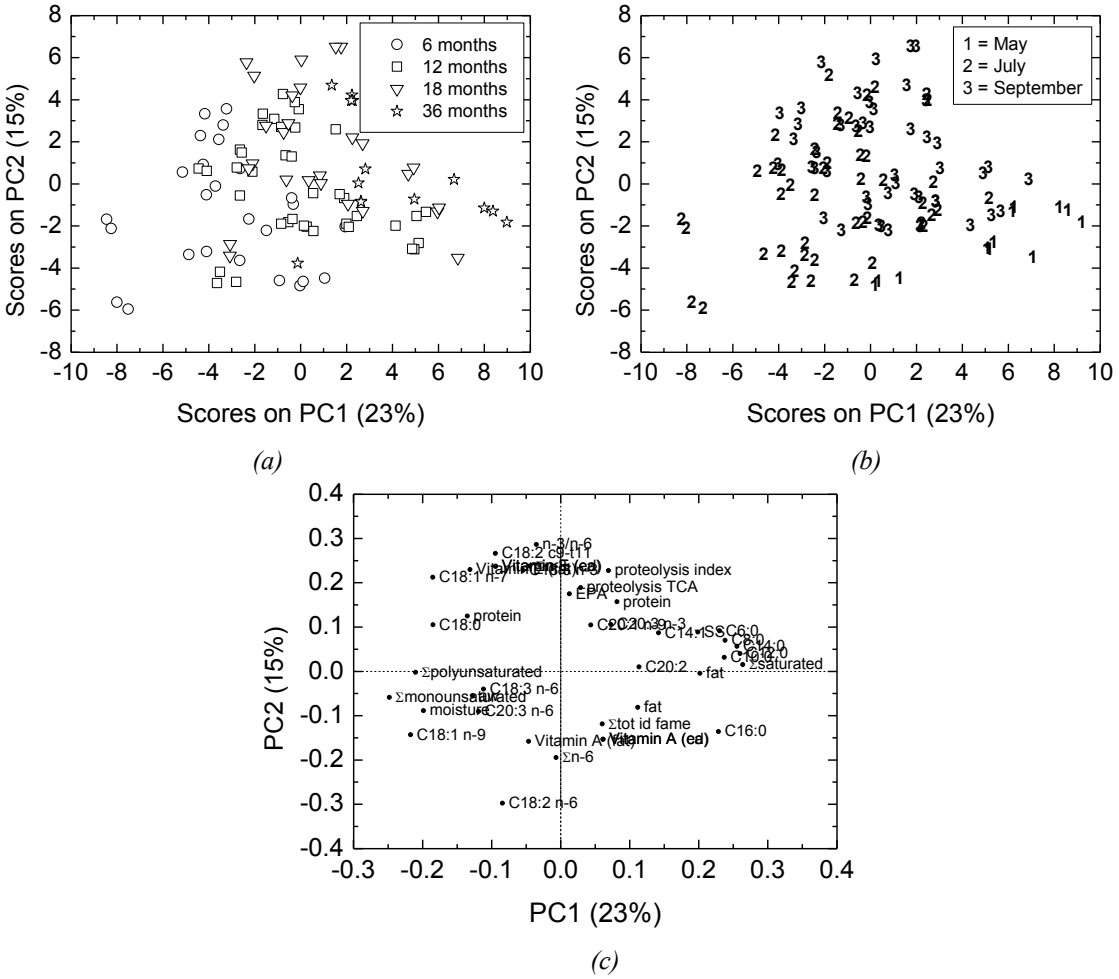
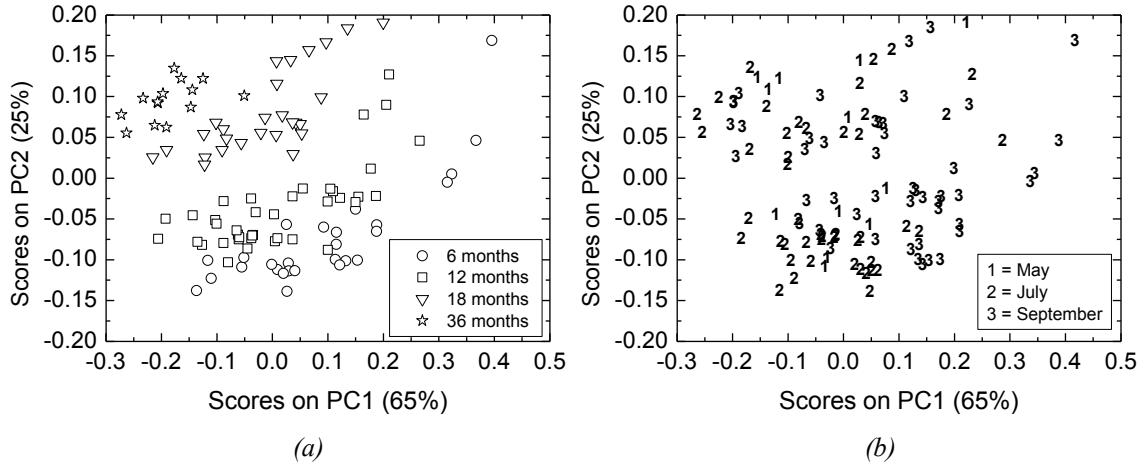


Figure 6.8. *Exploratory PCA analysis on the measured chemical properties. (a) and (b) Model scores: samples have been separated by (a) ripening age and (b) production period. (c) Model loadings.*

Figure 6.9 presents the score plot of a three-PCs model (explaining ~93% of the total variance) calibrated on the spectral data for the same samples as in the previous subsection. To enhance class separation, spectra were pretreated with SNV and first-order derivative.



(a) *(b)*

Figure 6.9. *Exploratory PCA analysis on the NIR spectra pretreated with SNV and first-order derivative. Model scores: samples are separated by (a) ripening age and (b) production period.*

The score plot highlighted quite clearly the separation by ripening age, which occurred along the minor diagonal. Due to the very large number of variables, the simultaneous analysis of model scores and loadings was not carried out for the spectral data.

NIR estimation of chemical properties

Table 6.7 presents the estimation results obtained for a subset of the properties reported in Table 6.5, together with those related to the estimation of the atherogenic (AI) and thrombogenic (TI) indices, which were calculated on the basis of the acidic composition using the equations proposed by Ulbricht and Southgate (1991). Both AI and TI are dietary factors related to the incidence of coronary heart disease. Results have been expressed in terms of the coefficients of determination in calibration (R^2), and cross (Q^2) and independent validation (R^2_p), together with the RPD (ratio of prediction to deviation).

Table 6.7 highlights that NIRS can return a quick and accurate estimation of several chemical properties, which generally are expensive and time consuming to analyze with traditional methods. As an example, water activity and proteolysis parameters (variables # 1, 12 and 13) could be used to assess the ripening age, while fatty acids profile (variables # 24 to 51) could assist in the traceability of the Asiago d'allevo (Cozzi *et al*., 2009). Eventually, TI and AI (which are calculated from fatty acids) could be related to the nutritional characteristics of the cheese.

It was interesting to note that NIRS can return an accurate estimation of the sodium chloride content of the cheese (variable # 11). A similar result was obtained by Lucas *et al*. (2008), and was justified by claiming that, although no absorbance occurs in the NIR region, sodium chloride can change the spectrum of water in the infrared overtone region

and consequently can be indirectly estimated by NIRS. The VIP index for the NaCl PLS model supported this explanation.

Table 6.7. *Number of latent variables (LVs) retained in each PLS model for the estimation of a subset of the chemical properties, and mean performance of the estimation models. The chemical properties have been indicated with their corresponding number (see Table 6.7).*

Estimated property #	**1**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**TI**	**AI**
LV	9	8	11	9	4	6	10	10	3	9	3	5	12	13
R^2	0.96	0.98	0.95	0.95	0.73	0.97	0.95	0.97	0.81	0.90	0.81	0.85	0.94	0.99
Q^2	0.93	0.96	0.87	0.89	0.58	0.96	0.89	0.94	0.76	0.73	0.79	0.80	0.80	0.92
R^2_p	0.92	0.96	0.87	0.90	0.56	0.96	0.89	0.93	0.77	0.76	0.78	0.79	0.82	0.92
RPD	3.8	5.2	2.7	3.0	1.6	5.1	3.1	4.0	2.1	2.0	2.2	2.3	2.2	3.6
Estimated property #	**24**	**25**	**26**	**27**	**28**	**30**	**33**	**34**	**35**	**36**	**39**	**49**	**50**	**51**
LV	10	9	11	10	9	7	8	9	13	10	11	10	10	7
R^2	0.89	0.95	0.95	0.97	0.93	0.88	0.89	0.94	0.96	0.9	0.94	0.95	0.97	0.85
Q^2	0.78	0.81	0.85	0.88	0.85	0.77	0.70	0.85	0.65	0.71	0.65	0.88	0.88	0.61
R^2_p	0.72	0.78	0.82	0.88	0.82	0.75	0.73	0.85	0.66	0.68	0.66	0.87	0.88	0.6
RPD	2.1	2.3	2.7	2.9	2.7	2.7	1.8	2.6	1.6	1.8	1.7	3.0	3.0	1.6

Classification results

PLS-DA classification results obtained employing the chemical measures were used as a reference to assess the results output from cheese classification based on the use of NIR spectra. Results (for the independent validation set) are summarized in Tables 6.8 and 6.9.

Table 6.8. *PLS-DA classification performance using the measured chemical properties listed in Table 6.5 (independent validation dataset).*

Classification	**Class sensitivity and specificity**				
		Lowland	Alpine		
Lowland/alpine	S_e (%)	100	100		
	S_p (%)	100	100		
		6 months	12 months	24 months	36 months
Ripening age	S_e (%)	88.9	92.3	96.3	100
	S_p (%)	98.7	97.1	95.0	100
		May	July	September	
Production period	S_e (%)	100	87.2	87.5	
	S_p (%)	100	90.0	89.8	
		Low	Medium	Medium-high	High
Production height	S_e (%)	75.0	57.1	100	100
	S_p (%)	86.6	91.9	98.9	97.3

Table 6.9. *PLS-DA classification performance using the NIR spectra (independent validation dataset).*

Classification		Class sensitivity and specificity			
		Lowland	Alpine		
Lowland/alpine	S_e (%)	100	99.1		
	S_p (%)	99.1	100		
		6 months	12 months	24 months	36 months
Ripening age	S_e (%)	100	100	100	100
	S_p (%)	100	100	100	100
		May	July	September	
Production period	S_e (%)	91.7	87.2	87.5	
	S_p (%)	98.9	91.7	88.1	
		Low	Medium	Medium-high	High
Production height	S_e (%)	82.5	57.1	92.9	87.5
	S_p (%)	88.1	95.3	98.9	89.3

6.2 Seafood authentication: toward a species-independent discrimination between fresh and frozen-thawed samples

6.2.1 Problem statement

The substitution of fresh fish with frozen-thawed fish is a typical fraud that not only damages consumers from an economical point of view, but can also cause safety issues (Pavlov, 2007). In fact, although freezing is one of the most widely used methods to extend the shelf life of seafood, it can affect the overall organoleptic properties of the product, and thawed meat is characterized by a higher susceptibility to microbial growth. Furthermore, fish authentication is important for correct product labeling (Martinez *et al*., 2003), as promoted by recent regulatory actions (Uddin, 2010; European Parliament Legislative Resolution, 2011).

Several methods have been proposed for the identification of the fresh/frozen-thawed substitution fraud (e.g., eye lens evaluation, measurements of dielectric properties, erythrocytes lysis, hematocrit evaluation, muscles histology, enzymatic methods, etc.; Uddin, 2010). The classification ability of the majority of these systems is strongly affected by the species under investigation, the integrity of the product (whole fish or fillet) or by its shelf life (Uddin, 2010). For example, the use of methods based on changes in dielectric properties, while being accurate on intact fish, provides poor results when applied on fillets (Duflos *et al*., 2002). Enzymatic assays were found to be useful in fillets, but not applicable to all species (Duflos *et al*., 2002). Recently, Bozzetta *et al*. (2012) proposed muscles histology as a simple method for the evaluation of the fresh/frozen-

thawed status. Despite the good classification results obtained on a wide range of species (more than 35 different species), the method requires time for sample processing (e.g., fixation, coloration) and the use of several reagents.

As an alternative to abovementioned techniques, more rapid analytical technologies have been developed. Among them (Nott *et al*., 1999; Karoui *et al*., 2006; Vidaček *et al*., 2008; Fernández-Segovia *et al*., 2012; Leduc *et al*., 2012; see also Chapter 7, Section 7.1), near-infrared spectroscopy (NIRS) has been suggested by the promising results obtained on some species (Uddin, 2010; Sivertsen *et al*., 2011; Fasolato *et al*., 2012; Zhu *et al*., 2012; Kimiya *et al*., 2013; Chapter 7, Section 7.1). NIRS is a well consolidated analytical technology and plenty of applications can be found in the field of seafood authentication (Cozzolino and Murray, 2012). To the author's knowledge, there are currently no NIRS application to multi-species databases, i.e. so far the fresh/frozen-thawed authentication problem has been solved only analyzing single species separately.

In this section, three alternative strategies based on LVM techniques are proposed and compared in order to develop a multi-species classifier of the fresh/frozen-thawed status of fish samples. While the first two strategies model the information on the species and on the fish fresh/frozen-thawed status together (either jointly or sequentially), the third strategy aims at explicitly separating them to improve the classification performance. A thorough validation of the proposed strategies is carried out using two NIR instruments exploring different spectral regions, on a total of more than 1200 samples.

6.2.2. Materials and methods

6.2.2.1 Available dataset

The number of samples available for model calibration and model validation (per species, class and instrument; see also section 6.2.2.2) is given in Table 6.10 and Table 6.11. The fresh/frozen-thawed classification models were built considering only the samples of the four species indicated in Table 6.10 (independently on the strategy used; see section 6.2.2.3). For model validation, instead, two datasets were considered, namely V1 and V2 (Table 6.11). The V1 spectra were collected at the same time of the calibration samples, whereas the V2 spectra were collected at a different time.

For each species, the I spectra considered were collected into an $\mathbf{X}_{sub}$ [$I \times N$] matrix, where N is the number of wavelengths (N = 401 for FOSS spectra and 421 for UNITY spectra) and subscript $_{sub}$ refers to the initial letter of the species Latin name. As an example, swordfish (*Xiphias gladius L*) FOSS samples were collected into $\mathbf{X}_{xg}$ [260×401]. Superscript * is used to identify the species which were used only for model validation (Table 6.11).

Table 6.10. *Available dataset in terms of number of samples per species, per class and per NIR instrument: calibration samples.*

Species	Symbol	Number of samples		FOSS	UNITY
Gilthead sea bream (*Sparus aurata*) (Fasolato et al., 2010b)	$\mathbf{X}_{sa}$	Fresh	53	✓	✓
		Frozen-thawed	53		
Red mullet (*Mullus barbatus*) (Fasolato et al., 2010a)	$\mathbf{X}_{mb}$	Fresh	53	✓	✓
		Frozen-thawed	53		
Sole (*Solea vulgaris*) (Fasolato et al., 2008)	$\mathbf{X}_{sv}$	Fresh	71	✓	X
		Frozen-thawed	17		
Swordfish (*Xiphias gladius L*) (Fasolato et al., 2012)	$\mathbf{X}_{xg}$	Fresh	101	✓	✓*
		Frozen-thawed	74		

* only 53 fresh and 53 frozen-thawed samples were available

Table 6.11. *Available dataset in terms of number of samples per species, per class and per NIR instrument: validation sets V1 and V2 samples.*

Validation set	Species	Symbol	Number of samples		FOSS	UNITY
V1	Gilthead sea bream (*Sparus aurata*) (Fasolato et al., 2010b)	$\mathbf{X}_{sa}$	Fresh	27	✓	✓
			Frozen-thawed	27		
	Red mullet (*Mullus barbatus*) (Fasolato et al., 2010a)	$\mathbf{X}_{mb}$	Fresh	27	✓	✓
			Frozen-thawed	27		
	Sole (*Solea vulgaris*) (Fasolato et al., 2008)	$\mathbf{X}_{sv}$	Fresh	35	✓	X
			Frozen-thawed	8		
	Swordfish (*Xiphias gladius L*) (Fasolato et al., 2012)	$\mathbf{X}_{xg}$	Fresh	50	✓	✓*
			Frozen-thawed	35		
V2	Gilthead sea bream (*Sparus aurata*) (Fasolato et al., 2010b)	$\mathbf{X}_{sa}$	Fresh	71	✓	✓
			Frozen-thawed	71		
	Red mullet (*Mullus barbatus*) (Fasolato et al., 2010a)	$\mathbf{X}_{mb}$	Fresh	71	✓	✓
			Frozen-thawed	71		
	Swordfish (*Xiphias gladius L*) (Fasolato et al., 2012)	$\mathbf{X}_{xg}$	Fresh	71	✓	✓
			Frozen-thawed	71		
	European sea bass (*Dicentrarchus labrax*) (Chapter 6, Section 6.1.1)	$\mathbf{X}^{*}_{dl}$	Fresh	-	✓	X
			Frozen-thawed	38		
	Different species‡	$\mathbf{X}^{*}_{mix}$	Fresh	-	✓	X
			Frozen-thawed	66		
	Carp/tench (*Cyprinus carpio*/*Tinca tinca*)	$\mathbf{X}^{*}_{cctt}$	Fresh	15	X	✓
			Frozen-thawed			

* only 27 fresh and 27 frozen-thawed samples were available
‡ among the species included in $\mathbf{X}_{mix}$: *Sarda sarda*, *Pollachius virens*, *Scorpaena scrofa*, *Pangasius spp*., *Scomber scombrus*, *Gadus macrocephalus*, *Hippoglossus hippoglossus*, etc.

6.2.2.2 NIR analysis

For details about the origin, freezing, thawing and storage of the samples, the reader is referred to the references wherein they were originally presented (Fasolato *et al*., 2008, 2010a, 2010b, 2012; Chapter 6, Section 6.1.1). As for the NIR analysis, the epiaxial white muscles of fresh and frozen-thawed samples were minced using a Retsch Grindomix (Retsch GmbH, Hann, Germany) at 4000 rpm for 10 s. Two aliquots per sample were scanned in small ring cups in reflectance mode with two different instruments: a FOSS NIRSystem 5000 (FOSS NIRSystem Inc., Silver Spring, MD, USA) at 2 nm intervals from 1100 to 2500 nm; and a UNITY Scientific SpectraStar 2500TW (Unity Scientific, Columbia, MD, USA) at 1 nm intervals from 680 to 2500 nm. For each aliquot, a mean spectrum was obtained by averaging multiple scans. Then, the spectrum of the sample was obtained by averaging those of the two aliquots. Reflectance (R) values were converted into absorbance (A) values through $A = \log(1/R)$. The analysis of the UNITY spectra was limited to the region between 680 and 1100 nm due to the noise characterizing the wavelength region above 1900 nm and considering that the NIR regions explored by the two instruments partially overlap. This was done in order to explore two different spectral regions with the available instruments.

6.2.2.3 Data analysis

PCA was applied as an exploratory tool to reveal the internal correlation structure of the available datasets. This preliminary analysis was intended to identify the major sources of variability of the data (sample species, sample status, etc.).

After the preliminary analysis, three LVM-based alternative strategies were developed for sample classification, with the aim of developing a classifier for the fresh/frozen-thawed status of the samples independently from their species. A schematic of the three strategies is given in Figure 6.10.

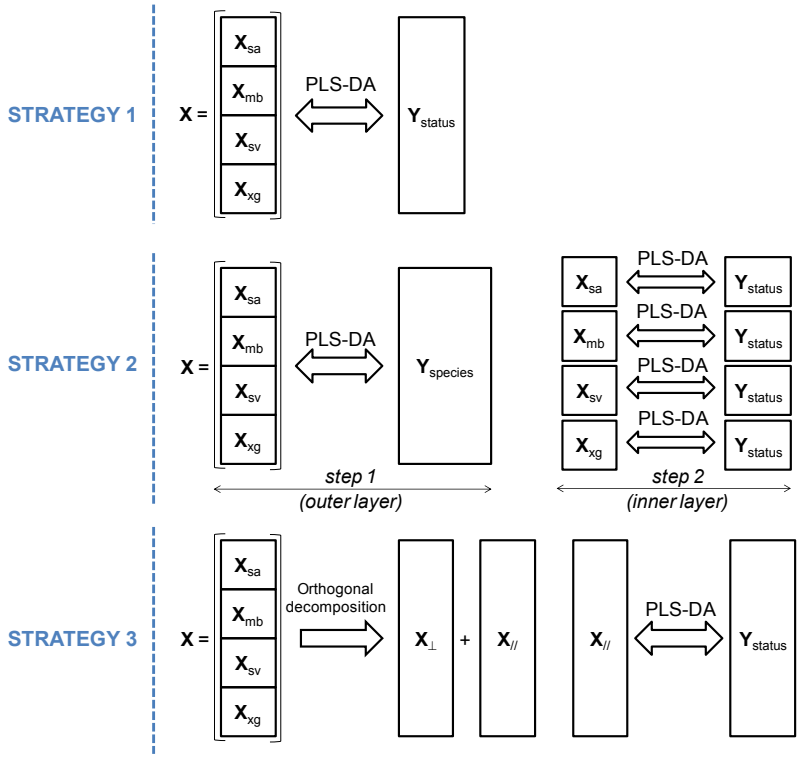


Figure 6.10. *Schematic of three strategies considered to build a species-independent fresh/frozen-thawed classification model. The sample status is either "fresh" or "frozen-thawed".*

In the first strategy, a PLS-DA model to classify the fresh/frozen-thawed status of each sample was built considering the species altogether, i.e. calibrating the model on the matrix **X** obtained by stacking on the top of each other the matrices $\mathbf{X}_{sa}$, $\mathbf{X}_{mb}$, $\mathbf{X}_{sv}$ and $\mathbf{X}_{xg}$ of the calibration set (Table 6.10).

In the second approach, a two-level cascade arrangement of PLS-DA models was proposed: in the first level, a PLS-DA model classified the samples according to their species; in the second level, different PLS-DA models (one for each species considered in the calibration set) discriminated between fresh and frozen-thawed samples.

In the third strategy, orthogonal PLS-DA (OPLS-DA; Trygg and Wold, 2002; Bylesjö *et al.*, 2006; see also Appendix A) was used to remove the information in the spectral data which is not related to the fresh/frozen-thawed status of the samples. OPLS-DA decomposed matrix **X** into matrices $\mathbf{X}_{\perp}$ and $\mathbf{X}_{//}$, containing respectively the orthogonal (i.e. not correlated) and parallel (i.e. correlated) information concerned with the fresh/frozen-thawed status of the fish. Therefore, the information related to the species difference is removed from the spectra matrix before the fresh/frozen-thawed status classification is carried out. This enables the calibration of a species-independent fresh/frozen-thawed classifier.

It should be stressed that both Strategy 1 and Strategy 2 represent a multi-species fresh/frozen-thawed classifier, with the variability related to the species modeled *together*

with the variability of the status (fresh vs. frozen-thawed). However, especially for Strategy 2, when a sample of a new species (i.e. a species not included in the calibration dataset) is analyzed, the classification results are not reliable, since the PLS-DA model in the outer layer cannot correctly assess the sample species. Conversely, when Strategy 3 is applied, only the variability strictly related to the sample status is retained, whereas the variability related to the species is not modeled at all, facilitating a species-independent classification.

In the calibration of the PLS-DA models of Strategy 1 and Strategy 2, four different spectra preprocessing techniques were considered combining standard normal variate (SNV; Barnes *et al*., 1989) and first and second order derivatives (D-1 and D-2; Savitzky and Golay, 1964), namely: (i) no preprocessing at all; (ii) SNV; (iii) SNV and D-1 and (iv) SNV and D-2. For Strategy 3, since the OPLS itself can be considered a preprocessing step, no preprocessing techniques were applied.

6.2.3 Results and discussion

6.2.3.1 Preliminary analysis

A 2-PC PCA model was calibrated on the matrix **X** obtained by stacking on the top of each other the matrices $\mathbf{X}_{sa}$, $\mathbf{X}_{mb}$, $\mathbf{X}_{sv}$ and $\mathbf{X}_{xg}$ of the calibration set (Table 6.10 and Figure 6.10). No preprocessing was applied on the spectral data[†]. The PCA model extracted more than 97% of the total variability (89.3 % on PC1 and 8.1 % on PC2). Model scores are shown in Figure 6.11, which shows that the greatest source of variability of the data is given by the difference between species. Samples of the same species, in fact, cluster in the same zone of the score space: sea bream samples are in the upper left region, whereas swordfish, mullet and sole samples are aligned along PC1 from the left to the right. Also the discrimination of interest (i.e. the fresh/frozen-thawed status) seems to insist mainly along PC1, with fresh and frozen-thawed samples partially overlapping. The overlap for samples of a given species can be attributed to the instrument used for spectra collection (FOSS, see Section 3.5), whereas the overlap for samples of different species is related to the collinearity between the information on the status and on the species. Another way of looking at the partial collinearity highlighted in Figure 6.11 consists in the analysis of the PLS-DA models that classify the **X** samples according to their fresh/frozen-thawed status (Strategy 1) and to the species they belong to (Strategy 2). The angle between the first

[†] Calibrating the PCA model on pretreated spectra did not change the conclusions drawn on Figure 1. Increasing the number of preprocessing steps, in fact, only affected the distance among clusters, suggesting that pretreating the spectra does not improve the classification results.

loadings of each model, in fact, was found to be 4°, i.e. the two models almost share the first latent variable (result not shown for the sake of conciseness). This means that there is a high collinearity between the major source of variability (first latent variable) of the two PLS-DA models, which are intended to capture the differences among samples due either to the status or to the species, but are built in such a way as to model the two sources of information jointly.

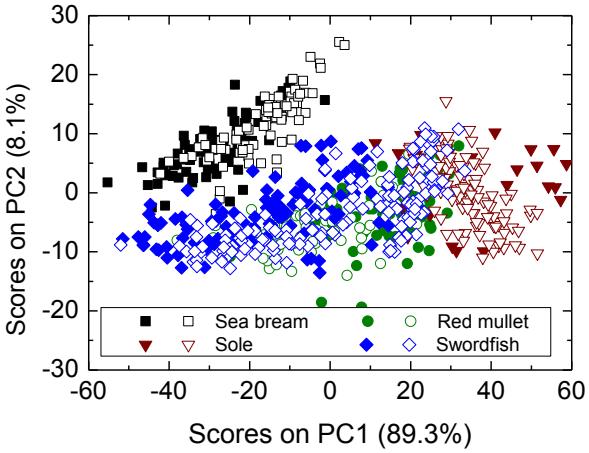


Figure 2. *Scores t_1 and t_2 of the preliminary PCA model. Closed symbols: frozen-thawed samples. Open symbols: fresh samples.*

Generally speaking, Figure 6.11 suggests that it is possible to classify the samples according both to their species and to their status, but model complexity (in terms of number of retained LVs) would be high, due to the multiple sources of variability within the data.

6.2.3.2 Strategy 1

Model calibration

Details on the classification model obtained with Strategy 1 are given in Table 6.12 for all the spectra preprocessing strategies explored. For each preprocessing, the number of LVs, the explained variance on the response matrix **Y** (var**Y**, %), and the sensitivity and specificity (in cross-validation, Se_{cv} and Sp_{cv}) towards the fresh fish class are presented for the best classification model.

Table 6.12. *Strategy 1 PLS-DA model statistics for different preprocessing techniques (calibration data).*

Spectra pretreatment	LVs	varY (%)	Se_{cv} (%)	Sp_{cv}(%)
No preprocessing	22	30.3	82.7	86.7
SNV	22	32.0	82.1	85.7
SNV and D-1	24	36.0	80.6	80.6
SNV and D-2	23	34.7	79.1	80.6

Although the data in Table 6.12 show that the performance of the models obtained with no pretreatments and with SNV alone are similar, the former was preferred. The large number of LVs retained can be justified from the considerations drawn in the previous section, as only a (small) fraction of the overall variability can be attributed to the fresh/frozen-thawed classification. The VIP index for the classification model highlighted mainly wavelengths related to the absorbance of water and lipids (i.e. wavelengths around 1150, 1400, 1700, 1850-1900 and 2250-2400 nm).

Model validation

The classification results for the validation datasets (Table 6.11) are presented in Table 6.13 in terms of misclassifications per class and overall accuracy (i.e., the percentage of correctly classified samples) for each species.

The classification results on samples of species used in the calibration dataset were similar in validation sets V1 and V2, with an overall accuracy comparable with that shown in Table 6.12. An exception was represented by the $\mathbf{X}_{sa}$ (*Sparus aurata*) samples of the V2 dataset, whose sensitivity towards the fresh class was found to be poor (36.6%). It should be noted that similar results were obtained also with other classification strategies (see Tables 6.16, 6.18 and also Fasolato *et al*., 2012b). The fact that different rearing conditions (with respect to those of the calibration samples) may have induced a change in the proximate composition (in terms of fat and water content) is a possible explanation for the poor classification result obtained with Strategy 1. In fact, the fat and water contents can affect the spectral response. Unfortunately though, it was not possible to test this assumption, since proximate composition data were not available.

Table 6.13. *Strategy 1 classification results for the validation datasets.*

Validation dataset	Species	Fresh samples misclassified	Frozen-thawed samples misclassified	Classification accuracy
V1	$\mathbf{X}_{sa}$	4	5	84.2
	$\mathbf{X}_{mb}$	7	11	68.4
	$\mathbf{X}_{sv}$	2	3	88.4
	$\mathbf{X}_{xg}$	6	9	82.6
V2	$\mathbf{X}_{sa}$	45	5	64.8
	$\mathbf{X}_{mb}$	15	17	77.5
	$\mathbf{X}_{xg}$	9	9	87.3
	$\mathbf{X}^*_{dl}$	-	0	100
	$\mathbf{X}^*_{mix}$	-	22	66.7

Although the PLS-DA model was calibrated on $\mathbf{X}_{sa}$, $\mathbf{X}_{mb}$, $\mathbf{X}_{sv}$ and $\mathbf{X}_{xg}$ samples, it was used to classify the $\mathbf{X}^*_{dl}$ and $\mathbf{X}^*_{mix}$ samples of validation set V2. First of all, the capability of the

model to describe these samples was checked by means of the square prediction error (SPE) and the Hotelling's T^2 statistics, where SPE is a measure of the representativeness of the model and T^2 is a measure of the difference between the sample condition and the average spectrum. Both the SPE and the Hotelling T^2 of each new sample were checked against the reference 95% confidence limits (SPE_{lim} and T^2_{lim}) obtained from the calibration data. It was found that both the $\mathbf{X}^*_{dl}$ and the $\mathbf{X}^*_{mix}$ samples showed an SPE value below the reference limits, but the $\mathbf{X}^*_{dl}$ samples had an Hotelling T^2 value much higher than T^2_{lim}. Hence, despite the model was found to be extrapolating (within the model plane), the classification accuracy for sea bass samples was 100%.

6.2.3.3 Strategy 2

Model calibration

Details on the models calibrated for Strategy 2 (Figure 6.10) are given in Tables 6.14 and 6.15, respectively for the classification of the samples according to the species and according to the status (one model for each species). For the PLS-DA model discriminating among species (Table 6.14), average sensitivities and specificities for the four species of the calibration dataset are reported.

With respect to the model discriminating among species (Table 6.14), spectra preprocessing only affected the number of the LVs to retain, whereas the explained variance on **Y** and the sensitivities and specificities did not change across preprocessing. With the aim of minimizing the preprocessing operations, the model obtained with no preprocessing at all was selected. The VIP index of the model pointed to the water and lipid contents as the major factors causing differences among the species. Note that the possibility of classifying fish (as well as meat) samples according to their species using VIS-NIR spectra has already been discussed in the literature (Cozzolino *et al*., 2002; Mamani-Linares *et al*., 2012).

Table 6.14. *Strategy 2 model statistics for the PLS-DA model discriminating among species for different preprocessing techniques (calibration data). Average sensitivities and specificities values are reported.*

Spectra pretreatment	LVs	varY (%)	Se_{cv} (%)	Sp_{cv}(%)
No preprocessing	22	70.0	100	100
SNV	22	70.7	99.9	99.8
SNV and D-1	20	71.0	100	99.8
SNV and D-2	16	70.9	100	100

Table 6.15. *Strategy 2 models statistics for the PLS-DA discriminating among status for different preprocessing techniques (calibration data).*

Species	Spectra pretreatment	LVs	varY (%)	Se_{cv} (%)	Sp_{cv}(%)
$\mathbf{X}_{sa}$	No preprocessing	9	35.3	86.8	84.9
	SNV	11	41.3	92.5	92.5
	SNV and D-1	5	35.4	86.8	86.8
	SNV and D-2	5	40.4	83.0	79.2
$\mathbf{X}_{mb}$	No preprocessing	13	41.6	88.7	90.6
	SNV	5	39.6	81.1	83.0
	SNV and D-1	8	43.4	86.8	92.5
	SNV and D-2	3	39.22	86.8	86.8
$\mathbf{X}_{sv}$	No preprocessing	10	30.6	88.2	98.6
	SNV	8	25.5	88.2	97.2
	SNV and D-1	3	22.3	88.2	91.5
	SNV and D-2	4	25.9	82.4	93.0
$\mathbf{X}_{xg}$	No preprocessing	15	38.9	89.0	91.1
	SNV	11	34.5	89.0	89.1
	SNV and D-1	8	37.2	86.3	89.1
	SNV and D-2	4	28.4	79.2	78.1

As for the species-tailored PLS-DA models that discriminate for the fish status, no preprocessing was used for all species except for $\mathbf{X}_{sa}$, for which SNV was preferred.[‡] A comparison between Table 6.15 and Table 6.12 shows that, as expected, the sensitivities and specificities towards the fresh class are higher than those obtained with Strategy 1.

Model validation

The accuracy of the PLS-DA model discriminating for the fish species (outer layer) was found to be almost 100% on the validation sets V1 and V2, with only four misclassifications on the $\mathbf{X}_{sa}$, $\mathbf{X}_{mb}$, $\mathbf{X}_{sv}$ and $\mathbf{X}_{xg}$ samples. The fish status classification results for the same validation datasets (inner layer) are presented in Table 6.16.

[‡] Again, it was noticed that derivatives did not improve the classification accuracy, as they just affected the model structure.

Table 6.16. *Strategy 2 fish status classification results for the validation datasets.*

Validation dataset	Species	Fresh samples misclassified	Frozen-thawed samples misclassified	Classification accuracy
V1	$\mathbf{X}_{sa}$	5	2	87.0
	$\mathbf{X}_{mb}$	3	6	83.3
	$\mathbf{X}_{sv}$	1	0	97.7
	$\mathbf{X}_{xg}$	5	15	76.7
V2	$\mathbf{X}_{sa}$	24	4	80.3
	$\mathbf{X}_{mb}$	26	7	76.8
	$\mathbf{X}_{xg}$	9	5	90.1
	$\mathbf{X}^*_{dl}$	-	2	94.7
	$\mathbf{X}^*_{mix}$	-	40	39.4

The effect of having a PLS-DA model tailored on each species can be clearly appreciated from Table 6.16, since for the species used also in the calibration step the classification accuracy is greater with respect to that shown in Table 6.13 for Strategy 1, with the exception of $\mathbf{X}_{mb}$ for V2 (only one misclassification). As for the $\mathbf{X}^*_{dl}$ and $\mathbf{X}^*_{mix}$ samples, the species attribution obtained from the species PLS-DA model was necessarily erroneous, as these species were not included in the calibration dataset. The capability of the model to adequately describe the samples was poor, i.e. SPE was higher than the reference limits. Nevertheless, in order to define which fresh/frozen-thawed classification model use, samples were assigned to the class towards which they showed the highest probability of attribution. For several samples, however, the class membership was not clear, since the probability of assigning the sample to a specific class was below 5% for all classes (i.e., species). The majority of the samples for which a class was clearly defined was assigned mainly to the sea bream and swordfish classes.

6.2.3.4 Strategy 3

Model calibration

As mentioned in section 6.2.2.3, the selection of the number of orthogonal components to remove should be derived from the comparison of the results of multiple criteria (Trygg & Wold, 2002). Figure 6.12 presents the effect of the number of OLVs removed on the number of LVs retained for the PLS-DA model, as chosen by cross-validation.

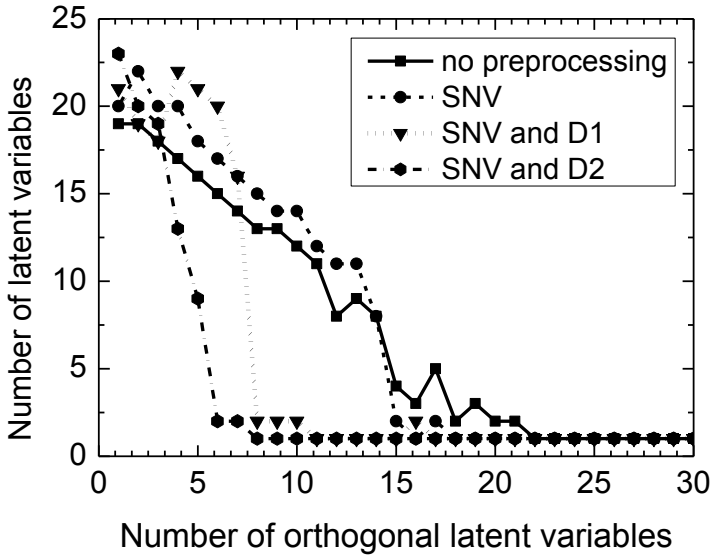


Figure 6.12. *Effect of the number of orthogonal latent variables removed on the final PLS-DA model structure (in terms of number of latent variables retained).*

Figure 6.12 clearly shows the effect of the pretreating the spectra prior to the OPLS analysis: spectra preprocessing removes an undesired variability and therefore reduces the number of OLV to be removed in order to obtain an 1-LV classification model. A model obtained with no preprocessing on the spectra was used[§]. According to Figure 6.12, 22 OLVs need to be removed to obtain a model with 1 LV (note the consistency of this results with the number of LV retained in the PLS-DA model of Strategy 1). The final model was eventually built by removing 28 OLVs, since an improvement in the sensitivity and specificity (in cross-validation) was observed when slightly more than 22 OLVs were removed. Details on the models with 22 and 28 OLVs are given in Table 6.17.

Table 6.17. *Strategy 3 OPLS-DA model details (calibration data).*

Spectra pretreatment	OLVs	OvarX (%)	varY (%)	Se_{cv} (%)	Sp_{cv}(%)
No preprocessing	22	84.1	29.8	90.0	92.1
No preprocessing	28	92.1	31.8	93.9	94.2

Ovar**X** in Table 6.17 represents the percentage of the variability of **X** removed by the OPLS algorithm. It can be concluded that only a very small fraction of the original variability was retained (and was therefore useful) for the fresh/frozen-thawed classification, as anticipated from the preliminary analysis. A comparison between the sensitivities and specificities towards the fresh class (in cross-validation) of the three

[§] The variability removed by SNV and/or D1 and D2 was deemed as "uneffective", since the classification accuracy of the final model was similar with and without preprocessing.

strategies (Tables 6.12, 6.14 and 6.15, and 6.17) shows that the accuracy of Strategy 3 is greater than that of Strategy 1 and 2.

An additional advantage of the OPLS algorithm is the possibility of analyzing separately $\mathbf{X}_{\perp}$ and $\mathbf{X}_{//}$. The VIP index obtained from the PLS-DA model calibrated on $\mathbf{X}_{\perp}$ to separate the samples according to their species (not shown) clearly resembles the one obtained for the outer PLS-DA model of Strategy 2. This is reasonable, because the PLS-DA model calibrated on $\mathbf{X}_{\perp}$ (no preprocessing, 17 LVs, 64.0% of explained variance on **Y**; Table 6.14) resembles the outer one calibrated for Strategy 2, i.e. the variability removed from the OPLS algorithm contains information related to the species. As a further confirmation, by calibrating a PLS-DA model on $\mathbf{X}_{//}$, it was verified that little information about the samples species was retained in $\mathbf{X}_{//}$: the model was found to explain only 3% of the variability on **Y**, hence having poor discriminating capability.

A drawback of pretreating the spectra with the OPLS algorithm was observed in the VIP index of the OPLS-DA model. Since the portion of the original **X** variance retained for the fresh/frozen-thawed classification is small, all wavelengths showed VIP values very close to one. Thus, it was not possible to identify specific spectral regions responsible for the discrimination, since the information retained in $\mathbf{X}_{//}$ for the classification of interest is related to the correlation among all the wavelengths of the spectra.

Model validation

The classification results for the validation datasets are presented in Table 6.18.

Classification results of Strategy 3 are generally good (Tables 6.13 and 6.16). As happened for Strategy 1 and 2, for the V2 dataset the $\mathbf{X}^*_{dl}$ samples were found to be far from the center of the model plane (large value of the Hotelling T^2 statistic), and hence the results should be taken with caution.

Table 6.18. *Strategy 3 classification results for the validation datasets.*

Validation dataset	Species	Fresh samples misclassified	Frozen-thawed samples misclassified	Classification accuracy
V1	$\mathbf{X}_{sa}$	0	4	92.6
	$\mathbf{X}_{mb}$	3	7	81.5
	$\mathbf{X}_{sv}$	0	1	97.7
	$\mathbf{X}_{xg}$	1	2	96.5
V2	$\mathbf{X}_{sa}$	31	2	76.8
	$\mathbf{X}_{mb}$	15	15	78.9
	$\mathbf{X}_{xg}$	6	6	91.5
	$\mathbf{X}^*_{dl}$	-	1	97.4
	$\mathbf{X}^*_{mix}$	-	12	81.8

6.2.3.5 Comparison among the proposed strategies

Table 6.19 provides a compact summary of the classification accuracy in validation for the three proposed strategies.

Table 6.19. *Classification accuracy in validation for the three proposed strategies.*

Strategy	V1		V2		Total	
	Misclassified	Classification accuracy	Misclassified	Classification accuracy	Misclassified	Classification accuracy
1	47	80.2	122	77.0	169	77.8
2	37	84.4	117	77.9	154	79.9
3	18	92.4	89	83.2	107	86.0

Clearly, Strategy 3 outperforms Strategy 1 and 2. As reported by some authors (Trygg and Wold, 2002; Svensson *et al*., 2002), the introduction of the OPLS algorithm is expected to be beneficial in terms of interpretation of the results, but it should not affect the overall accuracy. Stated differently, the performance of Strategy 1 and 3 are expected to be similar. The difference in terms of classification accuracy observed in Table 6.19 may be related to the fact that the greatest part of the variability within the calibration data **X** is not related to the classification of interest (i.e. the fresh/frozen-thawed status).

All the proposed strategies represent a fairly accurate (within the limitation of the FOSS spectral range) multi-species classification approach, and these results can open the route to the development of a species-independent approach to fresh/frozen-thawed fish classification. We stress the fact that all strategies were tested on samples of species not included within the calibration datasets and, although the models were found to be somewhat extrapolating (within the model plane), classification accuracies were generally good.

6.2.3.6 Results for UNITY spectra

All the results presented so far were obtained using a FOSS instrument to obtain NIR spectra. Results for UNITY spectra are not reported for the sake of concisness. With respect to results of FOSS spectra, classification accuracies were found to be higher (82.1% for Strategy 1, 91.0% for Strategy 2 and 88.4% for Strategy 3), thus confirming the conclusion drawn in section 6.2.3.5.

The main difference with the spectral region considered previously is the possibility of highlighting some wavelengths that are responsible for the fresh/frozen-thawed classification, i.e. some markers of the fish status. It was found that the region above 950 nm provided the greatest contribution within the classification model of Strategy 3 (particularly the regions 970-1030 nm and 1060 -1088 nm). The region 950-1100 nm was reported as informative for frozen-thawed status on pork meat (Park *et al*., 2001), and it is mainly related to the modification of the water content and NH_2 compounds. Additionally, changes on the absorbance at 970 nm (second overtone of O-H stretching) were observed also in *Psetta maxima* fillets using NIR imaging (Sivertsen *et al.*, 2011).

6.2.3.7 Comparison with other studies

Since several applications of fish sample classification according to their fresh/frozen-thawed status can be found in the literature, this section compares the classification accuracies obtained using the methods proposed in this study compared to those achieved with other methods reported in the literature.

Spectroscopic techniques have been applied to fish muscles, exuded juice or dry meat extract (Uddin *et al*., 2005; Karoui *et al*., 2006; Karoui *et al*., 2007; Vidaček *et al*., 2008; Uddin, 2010; Sivertsen *et al.*, 2011; Fernández-Segovia *et al*., 2012; Leduc *et al*., 2012; Fasolato *et al*., 2012; Zhu *et al*., 2012; Chapter 7, Section 7.1; Kimiya *et al*., 2013). The reported classification accuracy ranges between 85% and 96-100%, according not only to the instrument considered, but also to the classification strategy employed. The results presented in Table 6.19 for the FOSS instrument are generally worse than those reported by other authors, but the reason for the difference observed lays mainly in the spectral region used for the discrimination (NIR region above 1100 nm vs. VIS/NIR region below 1100 nm), as the results of the UNITY spectra seem to confirm. Additionaly, many of the investigations cited above were meant to be feasibility studies, namely the results were obtained using a limited number of samples (generally less than one hundred) having relatively limited variability, and only cross-validation statistics are reported. The overall classification accuracy obtained in the present study with UNITY spectra (680-1100 nm) is comparable with that of other studies (88.5% and 91.0%, respectively for Strategy 2 and 3, with values up to 97.0% for some species), but a much larger validation dataset was used in the present study.

An interesting comparison can be made with the histology-based classification proposed by Bozzetta *et al*. (2012), both because they used part of the same dataset and because it is one of the very few examples of a multi-species approach reported in the open literature. Generally, the histology-based method achieved higher classification accuracies than those presented here, but it must be emphasized that it relied on the operator's experience and required reagents and sample preparation. Additionally, when new (i.e. different from those used in the calibration step) species were introduced, the resulting classification accuracy was reduced.

6.3 Conclusions

The effectiveness of NIR spectroscopy coupled to LVM was shown in this Chapter to be an effective tool for food product characterization and fraud detection.

With respect to the examples presented in Section 1, it was shown that NIRS can be very effective for the assessment of the authenticity of wild European sea bass and Asiago d'allevo cheese. In both cases, the results obtained were found to be strongly consistent with those derived from chemical analysis, which were used as references to check the reliability of those based on NIRS. NIRS, however, allows a significant reduction in the time and cost of the analysis.

With respect to the problem of developing a multi-species fresh/frozen-thawed classification model from NIR data for fish samples presented in Section 2, three different strategies based on LVM have been proposed. In the first strategy, a PLS-DA model was built by concatenating vertically the spectra of samples of different species. In the second strategy, a cascade arrangement was proposed, where first (outer layer) a PLS-DA model separated the samples according to their species, and then (inner layer) a PLS-DA model tailored for each species classified the samples according to their status. In the third strategy, OPLS-DA was used to remove the variability in the data that is not related to the fresh/frozen-thawed status of the samples. The three strategies were tested on a very large database of spectra of two NIR instruments exploring different spectral regions, respectively from 680 to 1100 nm and from 1100 to 2500 nm, and using also samples of species not included in the calibration data. Strategy 2 and 3 returned the best validation classification accuracies, with values of 91% and 88.4%, and of 80% and 86%, respectively.

Chapter 7

Data fusion to enhance product quality characterization*

In this Chapter, the problem of how to combine the information derived from different analytical instruments in order to enhance the characterization of product quality is discussed through two food authentication case studies. In the first case study, which is described in Section 1, the purpose is that of discriminating between fresh and frozen-thawed fish samples using spectral data and RGB images. In the second case study, which is described in Section 2 of the Chapter, the purpose is that of classifying rainbow trout fillets according to their rearing farm and genetic strain using spectral, mechanical and colorimetric data. The last Section of the Chapter provides some general conclusions.

7.1 Data fusion for the authentication of fresh goatfish fillets

7.1.1 Problem statement

The problem of authenticating fresh West African goatfish fillets (*Pseudupeneus prayensis*) using different analytical technologies, namely a portable visible/near infrared (VIS/NIR) spectrometer, a compact digital RGB camera (red, green, blue color space) and a texture analyzer (whose use in fish characterization has already been reported, though not specifically to identify the fresh/frozen-thawed substitution fraud; Montero *et al.*, 2004; Costa *et al.*, 2011), is discussed in this Section. *Pseudupeneus prayensis* fillets are commonly marketed in Italy and often subject to fraud since the commercialization area is far from the fishing area.

To capture as much as possible the variability that can be encountered in the case of an on-line/at-line application, a large number of samples (more than 200) were collected along

*Ottavian, M., L. Fasolato, L. Serva, P. Facco and M. Barolo. Data Fusion for Food Authentication: Fresh/Frozen-Thawed Discrimination in West African Goatfish (*Pseudupeneus prayensis*) Fillets. *Food Bio. Tech.*, DOI: 10.1007/s11947-013-1157-x.

Dalle Zotte, A., M. Ottavian, A. Concollato, L. Serva, R. Martelli and G. Parisi. Authentication of Raw and Cooked Freeze-Dried Rainbow Trout (*Onchorhyncus mykiss*) by Means of Near-Infrared Spectroscopy and Data Fusion. *Food Res. Int.*, DOI: 10.1016/jfoodres.2013.10.033.

almost one year. All the analytical technologies tested are fast and can work in an on-line/at-line fashion in the production chain. However, while both the VIS/NIR spectrometer and the digital camera have the additional advantage of being non-invasive (requiring no sample preparation at all), sample texture measurement through the texture analyzer is destructive and requires one to remove samples from the production chain. Additionally to the comparison among the abovementioned techniques, information from different sensors were also combined into a unique model through the fusion of different analytical technologies. Data fusion was intended to enhance the classification accuracy (Bruwer *et al*., 2007; Cozzi *et al*., 2009; Casale *et al*., 2010).

7.1.2 Materials and methods

7.1.2.1 Sampling

The $N = 222$ samples of West African goatfish (*Pseudupeneus prayensis*) were collected in the middle-eastern Atlantic Ocean (Senegal, FAO 34).

Samples were immediately filleted (manually) by qualified personnel (in Senegalese facilities certified by the European Union). Freezing was carried out at –35 °C soon after filleting, and frozen fillets were then stored at –18 °C for 48 h prior to the analysis. All samples arrived to the laboratory covered in a plastic bag (to avoid drying) within 2 days after fishing: frozen samples were transported at –18 °C, and fresh samples at 2 °C. The fillets to be used as fresh standards were analyzed upon arrival to the laboratory, whereas the fillets to be used as frozen-thawed standards were analyzed after thawing the samples at 4 °C overnight (standardized, 18 ± 2 h) and after additional 30 min of exposure to the air at room temperature. Before the analysis, all samples were weighed. Samples were all characterized by the same initial quality. In fact, a sensory analysis carried out by a trained panel of experts assigned all samples to the same freshness category ("2 = very good" on a "1 = excellent" to "4 = poor" scale, according to the internal standard of the industrial partner that provided the samples).

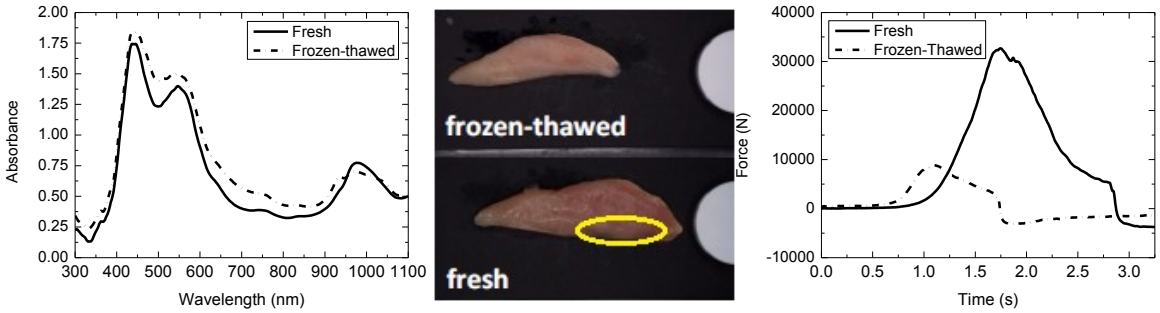
Table 7.1. *Number of samples considered in the study.*

Sampling period	Fresh	Thawed
Dataset 1 (June)	53	50
Dataset 2 (November)	51	-
Dataset 3 (January)[†]	35	33
Total	139	83

[†] 10 samples (5 for each class) were analyzed 24 hours after the other samples.

As reported in Table 7.1, for Dataset 3 ten samples (five per class) were analyzed with a 24 h delay with respect to the other samples (during the 24 h, they were kept at 4 °C and

covered with a plastic bag). These samples were considered only for the purpose of testing the classification model performance. The reason beyond the choice of the 24 h delay is that in Italian legal practices the recommended shelf life of frozen-thawed fish is 24 h. An example of the data collected (NIR spectra, RGB images and shear stress profile, collected on the spine side) is given in Figure 7.1.



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

Figure 7.1. *Example of the available data: (a) NIR spectra, (b) RGB images and (c) shear stress force. In (b) the sampling site for the NIR spectra is indicated.*

7.1.2.2 NIR spectroscopy data

VIS/NIR spectra were collected without sample pretreatment using a portable instrument (VIS/NIR diode array, Hamamatsu S3904) scanning wavelength from 300 to 1100 nm in transflection mode at 2 nm intervals. Spectra were collected from the cranial muscle portion of the fillets avoiding regions with evident residual blood: the area, which is indicated in Figure 7.1b, was approximately one fifth of the whole fillet area (see Figure 7.6 for more details on the fillets area). Spectra were collected in triplicate and then averaged. No preprocessing on the spectral data was applied prior to the statistical analysis[†]. An example of the available data is given in Figure 7.1a. In the following, the VIS/NIR spectra matrix will be identified as $\mathbf{X}_{NIR} = [N \times M_1]$, where M_1=401 is the number of wavelengths considered.

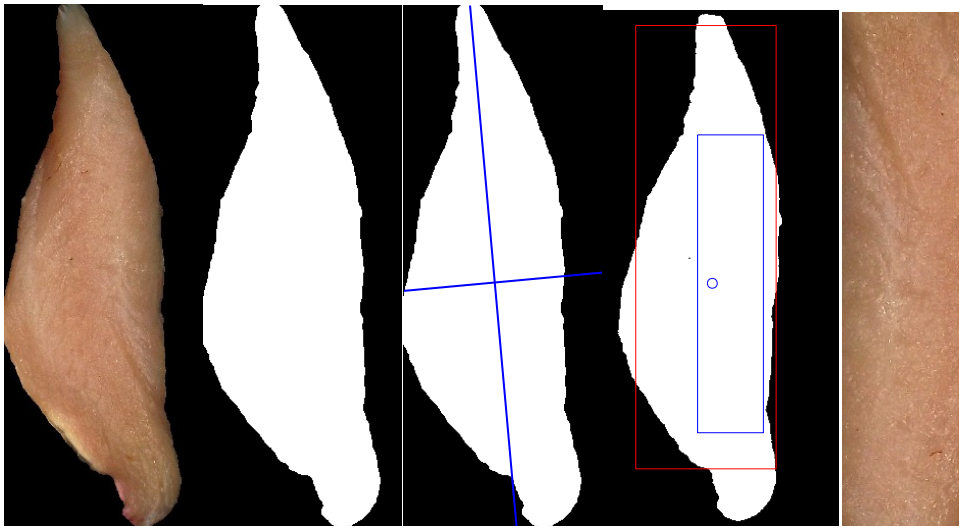
7.1.2.3 RGB images

Images were collected using a compact digital camera Kodak EasyShare M530 (RGB images of size [4000×3000] pixels saved in .jpg format) within the photographic box described in Chapter 1. Figure 7.1b gives an example of the available images. The camera was kept at 60 cm above the fish fillets with a resulting resolution of 1 pixel = $1.8 \cdot 10^{-5}$ cm^2, and images were taken in the automatic mode with a fixed exposure compensation (-

[†] It was verified that pretreating the data does not improve the classification results, and can even worsen them if derivatives are applied.

2, according to the camera user manual). All images were collected using a black background, a white ceramic standard and a ruler (to roughly check resolution).

After conversion to grayscale, images were segmented using two thresholds (a lower one and an upper one to filter out the background and the white ceramic standard, respectively). An example of the mask identified from the thresholding procedure is given in Figure 7.2. Two series of features (namely, color and textural features) were then extracted from the segmented images.



(a) *(b)* *(c)* *(d)* *(e)*

Figure 7.2. *Schematic of the procedure to extract the sub-image used in the texture analysis. (a) Original image, (b) mask obtained from thresholding, (c) detection of the major axes, (d) example of rectangles including mask pixel and (e) example of a sub-image.*

Color-related information within each image was summarized into M_2=12 features, namely the mean, standard deviation, skewness and kurtosis of the light intensity distribution of the red, green and blue channels. Hence, a matrix $[N \times M_2]$ was created that will be indicated as $\mathbf{X}_{color}$ in the following.

Texture-related information was extracted from the grayscale images by means of the wavelet transform (see Chapter 2, Section 2.2.2). For the available dataset, six resolution scales were found to be sufficient (Facco *et al.*, 2010). At each scale, six textural descriptors were extracted from each approximation: mean, standard deviation, skewness, kurtosis, entropy and energy of the light intensity distribution. This procedure generates an $[N \times M_3]$ matrix that will be indicated as $\mathbf{X}_{texture}$[‡], where M_3=36 derives by the extraction of 6 textural indices from the 6 relevant resolution scales of the wavelet decomposition. Since

[‡] $\mathbf{X}_{texture}$ will be used throughout the paper to indicate the texture measurement obtained from the images. The texture-related information obtained from the texture analyzer through the analysis of the shear stress profiles, instead, will be indicated as $\mathbf{X}_{shear}$.

wavelet transform can be applied only to matrices (i.e. images of regular shape), texture analysis using wavelets requires an additional preprocessing step on the images after segmentation, with the aim of framing the greatest possible area of the fillet within a rectangular region. This sub-image extraction was carried out through the automated process shown in Figure 7.2. The procedure goes through the following steps: after extracting the fillet from the background, the image is converted into a binary black-and-white image (where black is background and white the goatfish fillet) and the major axes of the fillet are identified (Pratt, 1991); after rotation around the fillet centroid, the area of all the rectangles including fillet pixels is evaluated, and the biggest one containing only fillet pixels is selected as the relevant sub-image.

7.1.2.4 Shear stress analysis

The shear stress profiles of each sample were evaluated using a TA-HDi texture analyzer (TA.xT2i Stable Micro System, Survey, UK) with a 100 kg load cell, 2 g of resolution, speed between 0.01 and 10 mm/s, equipped with a multiple blade probe (Kramer 10 blade – HDP/KS 10) that allows one to test the entire fillet surface. For each sample, a profile of the force versus time was obtained: positive forces were recorded at the beginning of the trial and until the sample was completely destroyed, whereas negative forces were recorded while the probe was returning to its initial position, the force being negative as a result of the cohesion between the fish meat and the probe itself. The information stored within the force profiles was summarized into six features, namely: the maximum force, the standard deviation of the positive part of the shear profile, the ratio between maximum and mean force (for the positive part of the signal), the work required to destroy the sample, the minimum force and the work of the probe during its way back. The M_4=6 features were normalized on the sample weight with the aim of accounting for size effects, and stored into an $[N \times M_4]$ matrix that will be indicated as $\mathbf{X}_{shear}$. A similar technique, with analogous descriptors of the force profile, was used also by Bruwer *et al.* (2007) for the analysis of snack food.

7.1.2.5 Data analysis

PCA was used as an exploratory tool of the available data, while PLS-DA was used for sample classification. In order to combine and fuse the information derived from different analytical techniques, data were arranged in a multiblock fashion (Westerhuis *et al.*, 1998). Namely, matrices were concatenated horizontally (as sketched in Figure 7.3) and block-scaled, i.e., each variable was scaled according to:

$$x_{n.m_k} = \frac{x_{n.m_k} - \bar{x}_{m_k}}{\sigma_{m_k}\sqrt{M_k}} \tag{7.1}$$

where $\bar{x}_{m_k}$ and σ_{m_k} represent the mean and standard deviation of each variable. The division by the square root of the number of columns (variables) of the block ensured the same representativeness of each block.

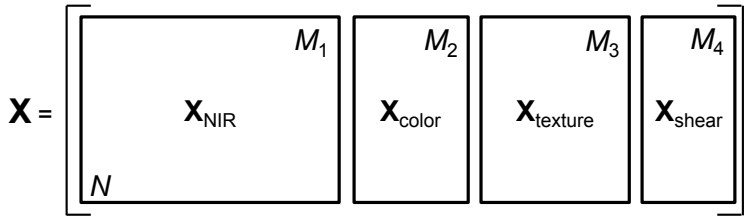


Figure 7.3. *Horizontal concatenation of the matrices of data obtained from different sensors within a multiblock framework.*

The samples to be used for calibration and validation were selected using a 2/3 – 1/3 split of the data and a D-optimal approach (de Aguiar *et al*., 1995) applied to the score obtained from a multiblock PCA model.

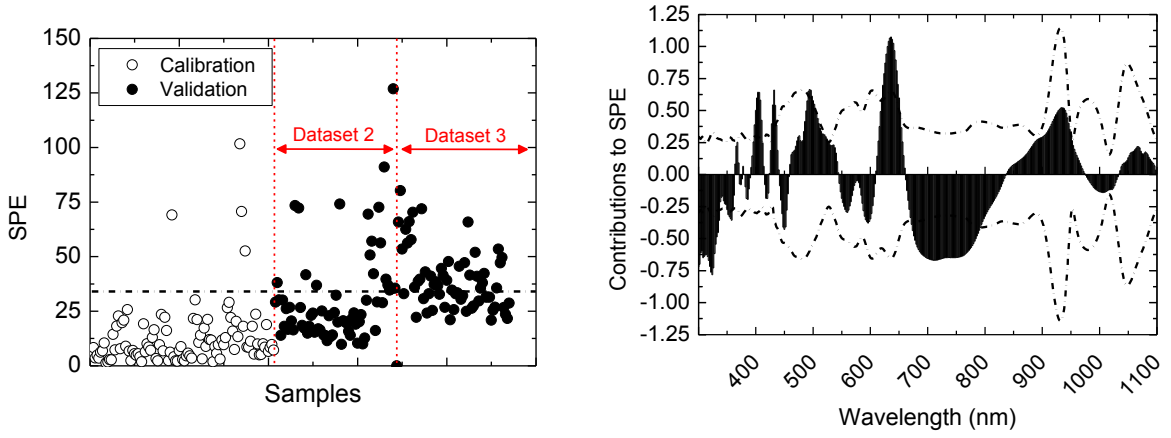
7.1.3 Results and discussion

7.1.3.1 Preliminary considerations

The purpose of this section is to explore the sources of variability within the data. Namely, since the samples were collected in three different periods of the year (Table 7.1) and the sample sizes were different, it was important to assess whether seasonality and size effects were present and, if so, whether they could be recovered from each of the three analytical technologies used throughout the study. This preliminary investigation is a key step to calibrate a robust classification model[§].

First, PCA was used to assess whether or not the samples showed a seasonality effect. A PCA model was calibrated on Dataset 1 for each of the three analytical technologies, and Dataset 2 and 3 samples were projected onto each model. As an example, model residuals are shown in Figure 7.4a for the spectral dataset ($\mathbf{X}_{NIR}$). It can be observed that both Dataset 2 and Dataset 3 samples display different variability with respect to the calibration samples. In fact, a fraction much larger than 1 % of the projected samples have a *SPE* value that exceeds the 99% confidence limit SPE_{lim}. A similar behavior was obtained by calibrating the PCA model on Dataset 2 (or 3) and projecting the samples belonging to the left-out datasets onto it. Overall, this means that the three datasets are different according to their VIS/NIR spectra, i.e. that VIS/NIR spectra can indeed display the seasonality effect across samples.

[§] Please note that the term *robustness* is used here referring to the sources of variability (such as seasonality and sample size) the available data explicitly account for. Other sources of variability, whose effect on the fresh/frozen-thawed discrimination capability has been demonstrated by other studies to be limited, were not considered.



(a) *(b)*

Figure 7.4. *Preliminary PCA analysis on the NIR spectra: (a) model residuals; (b) contributions plot for a sample whose residual SPE exceeds the confidence limit* SPE_{lim}. *The dash-dotted lines in (a) and (b) represent the 99% confidence limits.*

Similar results were obtained when images were used to build the PCA model; however, when shear stress profiles were employed all samples conformed to the PCA model. This leads to the conclusion that the seasonality effect can be detected from the PCA analysis of the NIR spectra and of the images, but not from the shear stress profiles. To further investigate this issue, the contributions plot for one of the Figure 7.4a samples exceeding the SPE_{lim} confidence limit is reported in Figure 7.4b. This figure suggests that the region around 700 nm, which exceeds the confidence limits defined for the contributions (dashed line), is responsible for the anomalous SPE value of the sample. Similar results were obtained also for other samples exceeding the confidence limits in the SPE plot (whether they belonged to Dataset 2 or Dataset 3); the region around 700 nm always provided the greatest contribution to the difference in the variability between Dataset 1 and Datasets 2 and 3. This region is characterized by the absorbance of the red color. The existence of a difference between samples related to the absorbance of the red color was confirmed also by the analysis of the contributions plot obtained with the color features in image analysis (not shown), where anomalous values for Dataset 2 and 3 were observed for the average value of the red channel. Hence, seasonality seemed to produce a color difference across the year. A detailed analysis of the loading plots (not shown) revealed that Dataset 1 samples exhibit the highest redness, and that the redness level decreased across the datasets, with Dataset 3 samples having the lowest redness values. Although there exist several muscles factors (e.g., anatomic location of the samples, fibers density/diameter, blood residual, proximate composition) or environmental or animal conditions (e.g, sexual maturity, feeding, methods of slaughtering) affecting the fillets color (Johnston, 2001; Schubring, 2010), some authors reported changes of fillet redness throughout the year both in reared and caught fish (Waters, 1982; Roth *et al*., 2005; Herland *et al*., 2010). It should be noted that the conclusion drawn about the seasonality effect are not dependent on the

fresh/frozen-thawed status of the samples; hence, at this point of the analysis, it is not relevant that for Dataset 2 there were no frozen-thawed samples available.

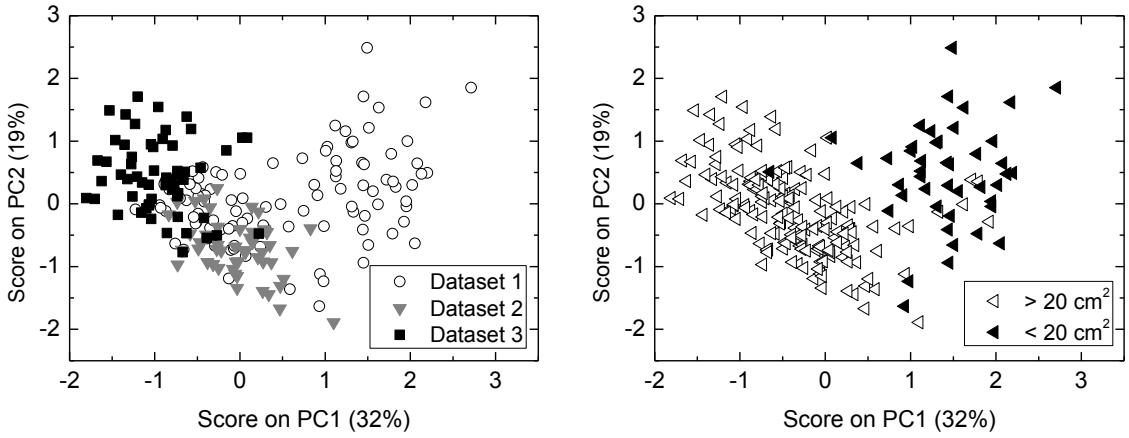
The seasonality highlighted by the PCA analysis must be taken into account in order to calibrate a robust fresh/frozen-thawed classification model. Namely, to compensate for the observed seasonality effect, the samples used for the calibration of the classification model must be chosen from all three datasets. Furthermore, for the comparison of the classification performance using different instruments to be fair, the same samples should be used for model calibration, independently from the analytical technology considered. To this purpose, first a multiblock PCA model was built on the matrix **X** obtained from the horizontal concatenation of $\mathbf{X}_{NIR}$, $\mathbf{X}_{color}$ and $\mathbf{X}_{shear}$, i.e. $\mathbf{X} = [\mathbf{X}_{NIR}\ \mathbf{X}_{color}\ \mathbf{X}_{shear}]$. Four PCs were retained, explaining 75% of the total variance, with 32% on PC1, 19% on PC2 and 13% on PC3. Then, a D-optimal sample selection was carried out on the these scores: in this way, the samples that better described the entire variability were selected, and this selection (which will be used in all classification exercises discussed later) was consistent for the three analytical technologies. Note that the $\mathbf{X}_{texture}$ data were excluded from the final multiblock PCA model because, as it will be clarified in the next Section, they were not predictive of the fresh/frozen-thawed status of the fish fillets.

Details on the partitioning of the samples into the calibration and validation sets are given in Table 7.2. The scores of the resulting multiblock model are shown in Figure 7.5 with two alternative ways of sample highlighting: by dataset (Figure 7.5a) and by size (Figure 7.5b).

Table 7.2. *Number of calibration samples per class and per season according to the D-optimal selection approach.*

Sampling period	Fresh	Thawed
Dataset 1 (June)	25	40
Dataset 2 (November)	35	-
Dataset 3 (January)	25	17
Total	85	57

Figure 7.5a indicates that samples from different datasets tend to separate, confirming the existence of a seasonality effect. Figure 7.5b suggests that an additional source of variability within the available data is given by the sample size. Although this source of variability was observed also on the PCA models calibrated on $\mathbf{X}_{NIR}$, $\mathbf{X}_{color}$ and $\mathbf{X}_{shear}$ separately, the D-optimal selection of the calibration samples through the multiblock approach allowed its inclusion within the classification model, thus ensuring robustness.



(a) *(b)*

Figure 7.5. *PC1 and PC2 scores of a multiblock PCA model built on the matrix obtained from the concatenation of* $\boldsymbol{X}_{NIR}$, $\boldsymbol{X}_{color}$ *and* $\boldsymbol{X}_{shear}$ *(calibration data). Samples are highlighted by (a) season and (b) size.*

The weight and the exposed surface area of all samples are given in Figure 7.6; note that the area was measured from the images, knowing the number of the fillet pixels (Figure 7.2) and the image resolution. Both weight and area were scaled with respect to their own average value (28.1 cm^2 and 22.0 g, respectively) in order to highlight the observed high correlation (the correlation coefficient was found to be equal to 0.96); incidentally, this suggested also the possibility of measuring the sample weight directly from the image (Gümüs and Balaban, 2010). The size effect was mainly explained by PC1 of the multiblock PCA, as clearly shown in Figure 7.5b. An arbitrary cut-off of 20 cm^2 was defined on the area in order to discriminate between "small" and "big" samples. The choice of an arbitrary cut-off is justified from the fact that few data are available on the biology of this species (Azzouz et al., 2010).

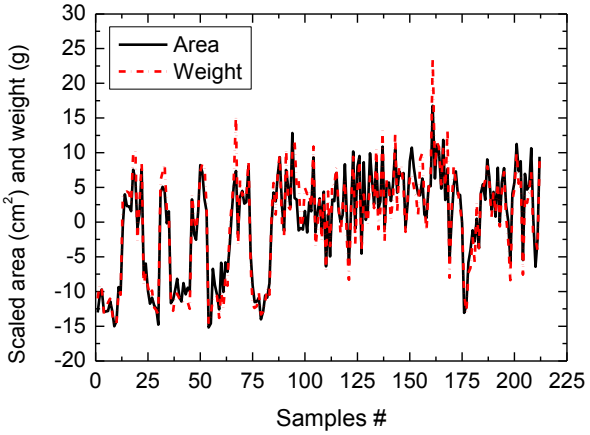


Figure 7.6. *Scaled area (solid line) and weight (dashed line) of the entire set of available samples.*

7.1.3.2 Fresh/frozen-thawed classification from single analytical technologies

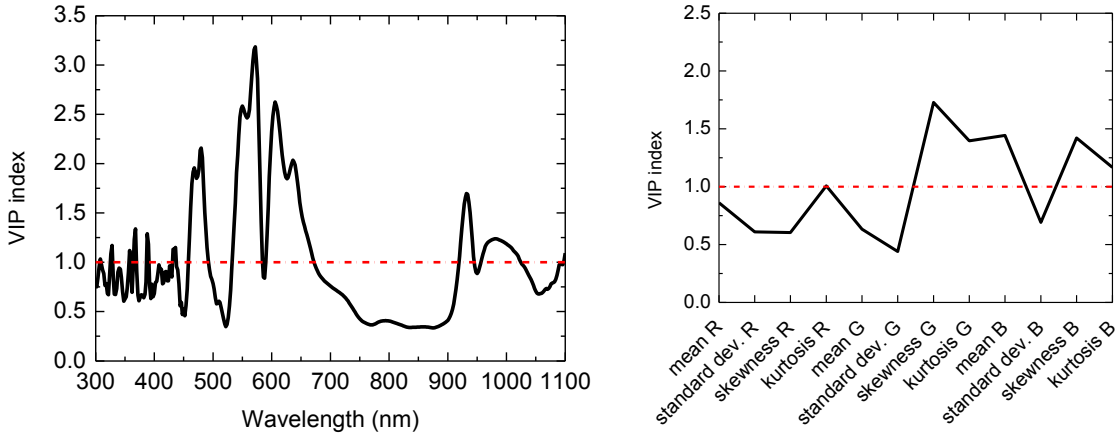
The results of the fresh/frozen-thawed classification obtained from each of the available analytical technologies are given in Table 7.3.

Table 3. *Fresh versus frozen-thawed PLS-DA classification results expressed in terms of sensitivity (Se) and specificity (Sp) towards the fresh class.*

	Calibration		**Validation**		**Validation after 24h**	
	Se	*Sp*	*Se*	*Sp*	*Se*	*Sp*
$\mathbf{X}_{NIR}$	100	100	100	100	60	100
$\mathbf{X}_{color}$	100	98	98	100	80	80
$\mathbf{X}_{texture}$	57	60	44	44	60	60
$\mathbf{X}_{shear}$	95	79	96	31	80	20

As anticipated in the preliminary data analysis, image texture ($\mathbf{X}_{texture}$) is not informative on the status of the fish, returning a poor classification performance. For the results shown in Table 7.3, the percentage of variance explained by the $\mathbf{X}_{texture}$ model in calibration and cross-validation was very poor, i.e. below 1%. Information on shear stress $\mathbf{X}_{shear}$, instead, is more informative, though characterized by low specificity values especially in the validation datasets. The results of Table 7.3 contrast with those reported by Zhu *et al*. (2013), who found image texture to be predictive of the fish status. The discrepancy could be due to the much higher variability considered in the present study (as discussed in the previous Section), the different classifier employed (linear versus non-linear), the type of images and their resolution (RGB versus hyperspectral) and the species considered in the analysis (goatfish versus halibut). Furthermore, the selection of a limited portion of the whole fillet (Figure 7.2) clearly introduces an approximation that can negatively affect the classification results (Sivertsen *et al*., 2011). This approximation, together with the quality of the images, could also explain the discrepancy between the performances of the models centered on $\mathbf{X}_{texture}$ and $\mathbf{X}_{shear}$.

Excellent classification performances were obtained from the VIS/NIR spectra ($\mathbf{X}_{NIR}$) and the color features ($\mathbf{X}_{color}$), though some errors were still observed in the validation set of the 10 samples analyzed 24 hours later (Table 7.1). The VIP index for the two PLS-DA models is shown in Figure 7.7.



(a) *(b)*

Figure 7.7. *VIP index for the PLS-DA classification models obtained from (a) NIR spectra and (b) color features. The dashed red line indicates the threshold value (VIP = 1).*

For the VIS/NIR spectra, the VIP index (Figure 7.7a) indicated the wavelength regions around 470, 560, 610, 930 and 970 nm as the most predictive of the fish status. These results were consistent with those reported from other researchers on different species. The wavelengths around 470 nm could be associated to the metmyoglobin (metMb). Liu and Chen (2001), though referring to chicken samples, suggested that the variation in the absorbance of this region is related to an early enzymatic degradation of metMb during thawing. Additionally, Fasolato *et al.* (2012) related the region between 450 and 520 nm to deoxymyoglobin.

The peak centered around 560 nm (corresponding to the heme-pigments) was associated with various forms of myoglobin or hemoglobin, while the shift observed around 610 nm (reported also by Sivertsen *et al*., 2011, in samples of *Gadus morhua*) was linked to a different amount of the oxy-hemoglobin. The wavelength at 930 nm was related to the third overtone C–H stretch in lipid and protein, as observed in *Psetta maxima* samples (Zhu et al., 2012). The region around 970 nm (the second overtone band of O–H stretch) was highlighted also by Uddin *et al*. (2005) and Zhu *et al*. (2013). Particularly, Uddin *et al*. (2005) reported that the higher absorbance showed by fresh samples could be ascribed to the different scattering effect on the intact tissue of fresh flesh. Hence, the region around 970 nm can be linked to the exudates released, a conclusion supported by the drip loss analysis.

For the color features extracted from the RGB images, the VIP index (Figure 7.7b) indicated the descriptors of the light intensity distribution of the green and blue channels (G and B skewness and kurtosis) as the most informative. It was difficult to relate the outcomes of the VIP index with the appearance of the fillets: looking at the images, it seemed that frozen-thawed fillets appeared to be more opaque than the fresh ones. Generally speaking, several authors reported color changes that fish fillets may undergo

due to several factors such as storage, slaughtering, muscle pH, etc. (Ruff *et al*., 2002; Guillerm-Regost *et al*., 2006; Roth *et al*., 2007; Roth *et al*., 2009), and these changes are typically species-dependent. Considering frozen products, the oxidation of the heme pigments (from myoglobin to met-myoglobin) has been usually linked to changes in the meat appearance (e.g. tuna, swordfish). Addis *et al.* (2012) related recently the effect of the freezing treatment to the red and blue RGB channels on bluefin tuna (Thunnus thynnus). Commonly a yellowish discoloration and an opaque appearance were described in lean fish (Connell, 1995). A change from translucent to opaque was reported also during the frozen storage of trout fillets, and linked to protein denaturation (Ozbay *et al*., 2006).

It should be stressed that both for the spectra and for the images the information related to the redness of the sample (wavelength around 700 nm and features of the red channel R) were not relevant within the classification model, consistently with the results of the seasonality analysis (which was not explicitly taken into account in Addis *et al*., 2012).

The classification results on the analyses carried out with a 24-hour delay were found to be poor for all the analytical technologies considered (at least two misclassifications; see Table 7.3), though the analysis of the model residuals revealed that the samples could be adequately described by the model. The misclassifications may be related to the physicochemical changes occurring in both fresh and frozen-thawed fillets during the early time of storage. Though such changes (such as the heme-pigment oxidations, the change in fillets hardness and appearance due to the release of liquids, etc.) are not detected from the models, they can negatively contribute to their performance.

An effect of the variability within the data was observed within the model structure. Both the PLS-DA models on $\mathbf{X}_{NIR}$ and on $\mathbf{X}_{color}$ were built on a relatively high number of LVs (13 and 5, condensating the 401 variables of $\mathbf{X}_{NIR}$ and the 12 variables of $\mathbf{X}_{color}$, respectively), as a confirmation that the information about the status of the fish fillets was somehow hidden within the data.

As discussed earlier, two major sources of variability were shown to characterize the available data, namely sample seasonality and size. When a fresh/frozen-thawed discrimination model is built, it is very important to explicitly take these factors into account, although this has been seldom (if ever) done so far (Zhu *et al*., 2013). Other factors may in principle affect the fresh/frozen-thawed classification model performance, such as the storage and freezing conditions (both temperature and time) and the product shelf life. However, with respect to the former factors, Fasolato *et al*. (2012) and Zhu *et al*. (2013) have shown that these conditions have limited impact on the possibility of correctly discriminating between fresh and frozen-thawed samples using either VIS/NIR spectra or NIR imaging. With respect to the product shelf life, Sivertsen *et al*. (2013) have drawn similar conclusions using VIS/NIR spectra and imaging: though the wavelengths related to

the sample freshness partially overlapped to those responsible for the fresh/frozen-thawed classification, the authors reported a classification accuracy close to 90%.

7.1.3.3 Data fusion for fresh/frozen-thawed classification

Information from different analytical technologies were fused using a multiblock framework with the aim of improving the classification performance of the second validation dataset (i.e., the 24 h validation data). Image texture data $\mathbf{X}_{texture}$ were excluded because of the limited content of information retained for classification purposes. All possible combinations of the three data matrices were fused to evaluate the classification performance. The classification results are shown in Table 7.4.

Table 7.4. *Multiblock PLS-DA classification results expressed in terms of sensitivity (Se) and specificity (Sp) towards the fresh class for all the fusion combinations explored.*

Fused data combination	Calibration		Validation		Validation after 24h	
	Se	*Sp*	*Se*	*Sp*	*Se*	*Sp*
$\mathbf{X}_{NIR}$, $\mathbf{X}_{color}$	100	100	98	100	100	100
$\mathbf{X}_{NIR}$, $\mathbf{X}_{shear}$	99	100	98	100	80	100
$\mathbf{X}_{color}$, $\mathbf{X}_{shear}$	100	100	98	100	100	100
$\mathbf{X}_{NIR}$, $\mathbf{X}_{color}$, $\mathbf{X}_{shear}$	100	100	98	100	100	100

As can be clearly seen from Table 7.4, the fusion of different information significantly improved the classification results, particularly for the 24 h validation data. Note that $\mathbf{X}_{shear}$, which returned poor specificity values if considered alone (Table 7.3), turned out to be informative when coupled with other data. Among the possible combinations of Table 7.4, the the first one is of utmost importance for two reasons: the camera and the VIS/NIR spectrometer avoid to destroy the sample, and the measurements are very fast. Hence, the fusion of digital camera and VIS/NIR spectrometer would be preferable for potential in-line/at-line inspections. Figure 8 presents the VIP index for the multiblock PLS-DA fusing spectra and color features.

It is shown that the color features have an higher impact within the multiblock classification model than the spectra. As for the color features, the profile of the VIP index resembles that of Figure 7.7b, though the importance of the features of the red channel (in terms of mean and kurtosis) was found to be different. This can be explained considering that the multiblock approach (and, as a consequence, the VIP index) describes also the correlation between the variables of different blocks (e.g., the mean of the red channel and the region of the spectra around 700 nm).

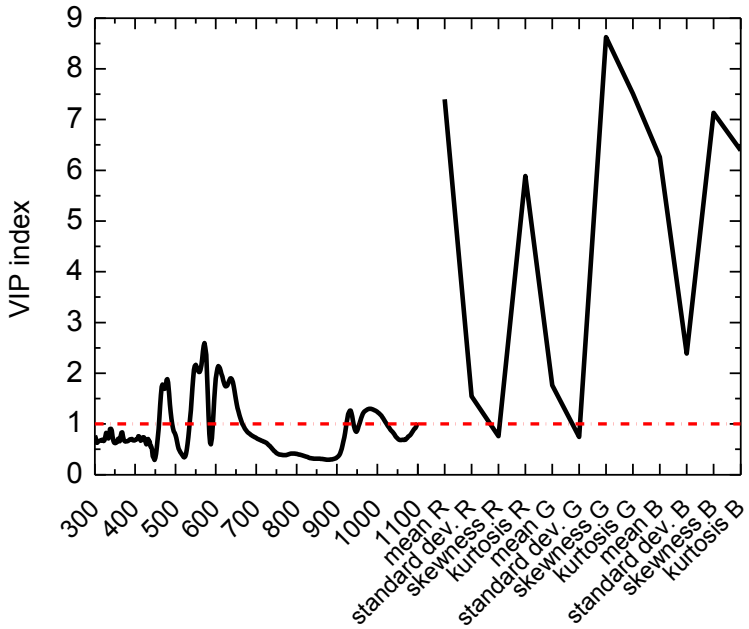


Figure 7.8. *VIP index for the multiblock PLS-DA classification model fusing NIR spectra and color features.*

7.2 Data fusion for the authentication of rainbow trout fillets

7.2.1 Problem statement

Rainbow trout (*Oncorhinchus mykiss*) is a North American salmonid which typically lives in oxygenated and clear waters. Thanks to its resistance to temperatures up to 20 °C, however, it proves capable of adaptation to many areas. The top ten countries in the farming of freshwater trout (*O. mykiss* and *S. trutta* above all) are Turkey, Iran, France, Italy, USA, Denmark, Spain, Germany, Poland, and China. In 2006, these countries produced about 75% of all farmed freshwater trout, for an overall value of about 1.3 billion USD.

The purpose of this study was to evaluate the performance of NIRS as a fast, cost-effective and non-destructive method for the assessment of both raw and cooked rainbow trout (*Oncorhynchus mykiss*) fillet quality. Samples of five different genetic strains from three different rearing farms were considered. Following Gjerde and Martens (1987), who showed that water absorption bands might interfere with important spectral bands of other analytes, fillets were freeze-dried before NIR analysis. Since NIR spectra classification was found to be less accurate in regard to genetic strain, a data fusion approach was adopted to improve the results. Spectral information was fused with mechanical properties and colorimetric data within a multiblock framework resulting in higher classification

accuracy. To the author knowledge, this is the first study that attempts to classify samples by genetic strain using NIR spectra.

7.2.2 Materials and methods

7.2.2.1 Sampling

A total of N =150 farmed rainbow trout (*Oncorhynchus mykiss*) fillet samples was used in this study. Samples of five different genetic strains (indicated as IT1, IT2, IT3, USA and UK, according to origin) and three different rearing farms (in Trentino Alto Adige region in northeast Italy, indicated as farms A, B and C) were considered, for a total of ten samples per farm per genetic strain, i.e. N = 10 (samples) × 3 (farms) × 5 (genetic strains). Farm characteristics were as follows: farm A - indoor rearing tanks supplied with well water at a constant temperature (range: 11-14 °C) throughout the year; farm B - outdoor rearing (temperature range: 9-11 °C); and farm C - outdoor rearing (temperature range: 3-14 °C).

Fish were collected after reaching average weight greater than 600 g (i.e., their commercial size). Twenty-four hours *post mortem,* fish were filleted and fillets were transported in refrigerated condition to the laboratory and immediately processed. Left and right fillets of each specimen were both weighed and analyzed: the former were used to evaluate raw fillet properties; the latter were used to evaluate cooked fillet properties. As regards the latter, prior to physiochemical analyses, each sample was wrapped in aluminum foil and boiled in a steamer for 10 minutes, then cooled at room temperature and weighed after broth removal. Cooking loss was then calculated and expressed as percentage of weight decrease.

7.2.2.2 NIR spectroscopy data

After the freeze-drying process, fillets were ground twice with a Retsch Grindomix GM 200 (Retsch GmbH, Hann,Germany) at 4000 rpm and then at 8000 rpm per 10s. Two aliquots per sample were placed in a 50 mm diameter ring cup and scanned in reflectance mode at 2 nm intervals from 1100 nm to 2500 nm using a scanning monochromator NIRSystem 5000 (FOSS NIRSystem, Silver Spring, MD, USA). For each sample aliquot, a mean spectrum was obtained by averaging from 32 multiple scans; the spectrum of the sample was then obtained by averaging those of the two aliquots.

7.2.2.3 Chemical analysis

For each sample, the chemical properties analyzed were moisture (method 934.01; AOAC, 2002), protein (method 992.15; AOAC, 1993), total lipid content (method 920.39; AOAC, 2002), and ash (942.05; AOAC, 2002). Fatty acid profiles of freeze-dried samples were

analyzed by gas chromatography (Morrison and Smith, 1964) after Folch extraction (Folch *et al.*, 1957).

7.2.2.4 Physical analysis

Texture and color information were collected using a Zwick-Roell® texture analyzer (Zwick-Roell, Ulm, Germany) and a Spectro-color® meter (Dr. Lange, Düsseldorf, Germany), respectively.

The compression test was repeated three times in three different fillet positions (epaxial, ventral and caudal) using a cylindrical probe, a 200 N load cell, and at 20 mm/min (constant) speed. The shear stress test was carried out in the middle of the fillet using a linear blade, a 200 N load cell, and at 30 mm/min (constant) speed. Data were collected in terms of compression force or shear stress at different percentage of deformation (with respect to the original dimension) and at different absolute deformation (in mm).

CIELAB *L*, *a* and *b* (i.e. the three color indexes obtained from the colorimeter; CIE, 1974) were obtained by averaging from three replicates for each measurement point. Hue angle and chroma values were derived from *a* and *b*.

7.2.2.5 Data analysis

Several chemometric tools were used to analyze the available data: PCA for preliminary data analysis, PLS for estimating chemical properties from NIR spectra, and PLS-DA, LDA, QDA and *k*NN for their classification. Furthermore, in order to improve classification accuracy, data from different instruments (spectra, mechanical properties, color information, etc.) were fused within a multiblock framework as described in Section 7.1.2.5. Please note that while PLS-DA models are calibrated directly on the available data, LDA, QDA and *k*NN models are calibrated on the scores obtained from their PCA decomposition. In the latter case, cross-validation is used to optimize the number of PCA factors (and, for *k*NN, the number *k* of neighbors to consider; Balabin *et al.*, 2010).

In order to validate the proposed models, the data were split into two groups: 120 samples were used in the calibration step, while the remaining 30 (2 samples per farm per genetic strain) were used for model validation. Model parameters were selected in cross-validation (Wold, 1978) of the calibration data using a venetian blind algorithm.

Please note that except for the PCA models used in the preliminary data analysis, all models were built on raw and cooked freeze-dried fillets separately.

7.2.3 Results and discussion

7.2.3.1 Preliminary considerations

The average NIR spectra for both raw and cooked fillets are shown in Figure 7.9.

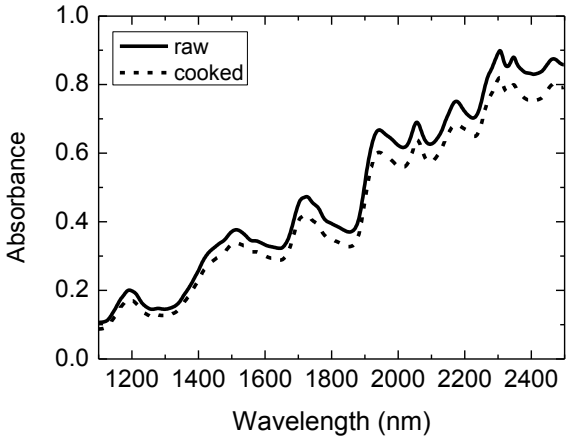


Figure 7.9. *Average raw and cooked freeze-dried rainbow trout samples.*

The main result of the cooking process is a downshift of the spectra. Figure 7.9 reveals the existence of two regions (around 1400 nm and around 1900 nm) in which the difference is minimal, and this is consistent with the freeze-drying treatment of the samples because water absorbance is usually reported for these regions (Murray, 1986).

The score plots of a 3 PCs PCA model calibrated on the [300×700] matrix of the spectra (raw and cooked fillets, with no spectra pretreatments applied) are shown in Figure 7.10.

PC1, explaining 95% of the total variance, mainly accounts for the difference between raw and cooked samples (see Figure 7.10a). The loading values on PC1, in fact, are almost the same for the entire spectral range considered, indicating that the difference between raw and cooked samples can be mainly related to average absorbance clearly observed in Figure 7.9.

In Figure 7.10b-c, the scores of the raw samples are highlighted by their respective farm (Figure 7.10b) and genetic strain (Figure 7.10c). The plots suggest that at least in the PC1-PC2 plane, samples of different farms or genetic strains are not linearly separable, and that non-linear classifiers (such as QDA and *k*NN) might be necessary. Similar behavior was observed also for cooked samples.

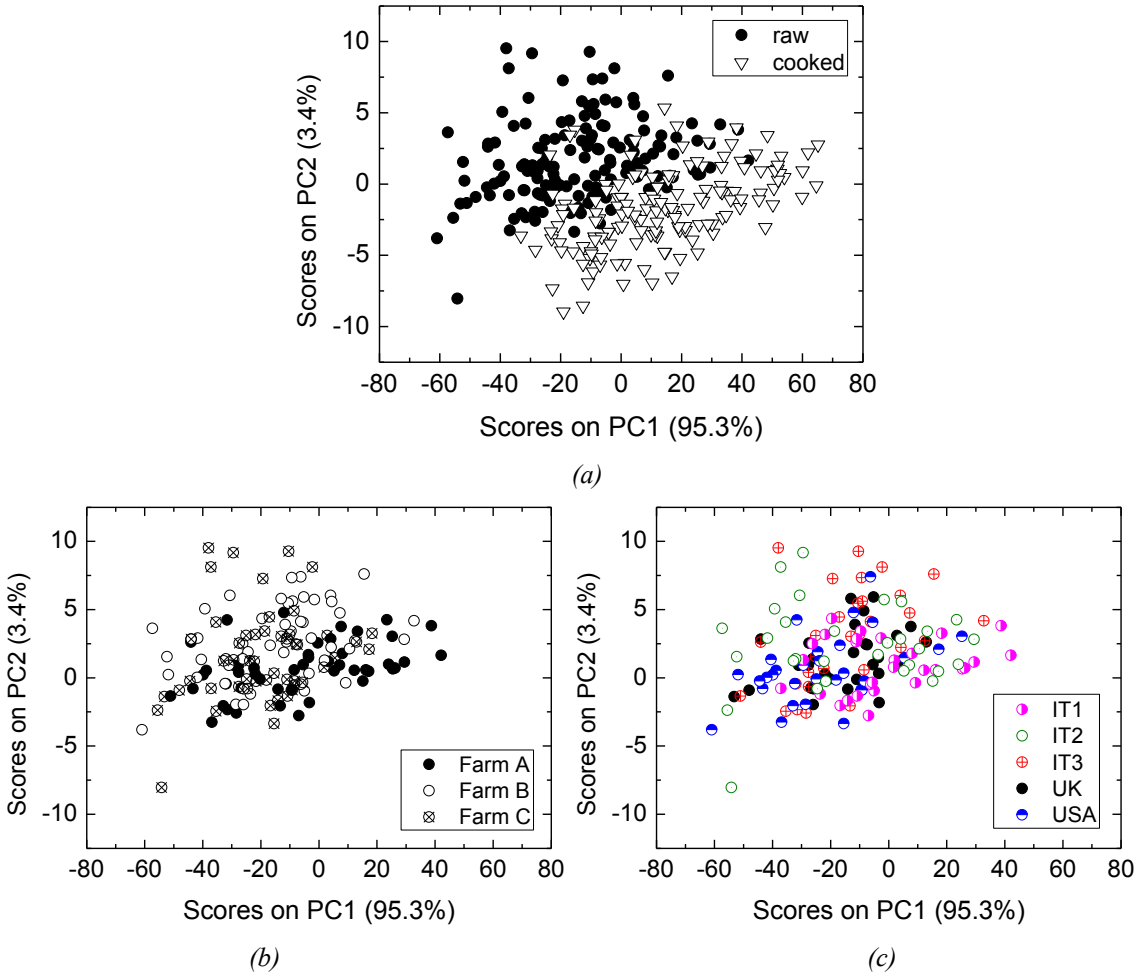


Figure 7.10. *PC1-PC2 score plot of the preliminary PCA model built on the spectra matrix. In (a) raw and cooked samples are highlighted differently. In (b) and (c) raw samples belonging to different farms and genetic strains (respectively) are highlighted differently.*

7.2.3.2 Classification

Results of the four classification strategies are given in the following for the raw freeze-dried fillets. Results for the cooked freeze-dried fillets are not reported here for the sake of conciseness. The accuracy of the models was approximately the same in the two cases, i.e. the cooking process did not alter the discriminating capabilities previously observed.

For each classification strategy, model parameters (LVs of the PLS-DA models, PCs of the PCA models for LDA, QDA and *k*NN, and *k* for *k*NN), spectra pretreatment (no pretreatment, SNV and/or its combinations with derivatives D1 and D2, i.e. the four combinations tested), and calibration, cross-validation and validation accuracies are presented.

Results for classification by farm using NIR spectra are given in Table 7.5. With the exception of the PLS-DA model, classification accuracies were found to be fairly similar among models. It is interesting to note that NIR permitted not only discrimination between

farm A (indoor rearing, hence with more uniform and controlled fish farming conditions) and farms B and C (outdoor rearing), but also between farms B and C, which differed in terms of water temperature, altitude (400 and 700 m a.s.l., respectively), and dissolved oxygen (8.25 and 10.35 ppm, respectively).

Table 7.5. *Classification by farm from NIR spectra of raw samples: results*.

Model	Model parameters	Spectra pretreatment	Calibration accuracy (%)	Cross-validation accuracy (%)	Validation accuracy (%)
PLS-DA	5 LV	SNV & D2	85.8	69.2	86.7
LDA	16 PC	No preprocessing	100	100	100
QDA	12 PC	No preprocessing	100	100	100
*k*NN	*k* = 5, 16 PC	No preprocessing	98.3	98.3	96.7

In confirmation of the conclusions drawn from the preliminary analysis (see Section 7.2.3.1), a high number of PCs were retained for classification (with the difference between the PLS-DA model and the other models consisting mainly of the pretreatment on the spectra).

Results for classification of raw samples by genetic strain are given in Table 7.6 for NIR spectra and in Table 7.7 for proximate composition, color and mechanical properties.

Table 7.6. *Classification by genetic strain from NIR spectra of raw samples: results*.

Model	Model parameters	Spectra pretreatment	Calibration accuracy (%)	Cross-validation accuracy (%)	Validation accuracy (%)
PLS-DA	17 LV	SNV & D2	62.5	44.2	36.7
LDA	21 PC	No preprocessing	92.5	66.7	60.0
QDA	9 PC	No preprocessing	85.8	62.5	53.3
*k*NN	*k* = 5, 14PC	No preprocessing	90.8	90.8	60.0

Table 7.7. *Classification by genetic strain from proximate composition and color and mechanical properties of raw samples: results*.

Data	Model	Model parameters	Calibration accuracy (%)	Cross-validation accuracy (%)	Validation accuracy (%)
Proximate composition	PLS-DA	4 LV	40.8	33.3	26.7
	LDA	4 PC	38.3	33.3	23.3
	QDA	3 PC	35.0	30.0	30.0
	*k*NN	*k* = 3, 1 PC	89.2	89.2	26.7
Color	PLS-DA	15 LV	65.0	53.3	60.0
	LDA	14 PC	70.0	54.2	53.3
	QDA	10 PC	76.7	49.2	56.7
	*k*NN	*k* = 1, 12 PC	81.7	81.7	63.3
Mechanical properties	PLS-DA	20 LV	64.2	41.7	43.3
	LDA	11 PC	55.8	45.8	53.3
	QDA	11 PC	82.5	47.5	46.7
	*k*NN	*k* = 3, 4 PC	82.5	82.5	43.3

Cross-validation and validation accuracies (which more closely resemble the practical use of the models on unknown samples) were found to be unsatisfactory, with values generally below 60% (with few exceptions). Please note that as suggested by PCA analysis, non-linear classifiers (*k*NN in particular) provided better performance.

In order to improve the results, the available information (proximate composition, NIR spectra, color and mechanical information) were fused together. Since *k*NN returned the highest classification accuracy (see Tables 7.6 and 7.7), it was used also to classify the combined information. Results are given in Table 7.8 for three different data combinations. It should be noted that higher classification accuracies were obtained when fusing the available information (compared to accuracies obtained using each piece of information separately).

Table 7.8. *Multi-block kNN classification by genetic strain from all available information: results for raw samples*.

Fused data	Model parameters	Calibration accuracy (%)	Cross-validation accuracy (%)	Validation accuracy (%)
Color, NIR	k = 3, 8 PC	84.2	84.2	66.7
Color, NIR, Mechanical properties	k = 3, 7 PC	91.7	91.7	83.3
Color, NIR, Mechanical properties, Proximate composition	k = 1, 13 PC	91.7	91.7	73.3

The combination of NIR spectra, color and mechanical information represented the best choice, given that proximate composition analyses (whose addition to the fused information did not improve the cross-validation accuracy on which the selection of the best model was based) are much more time consuming. The fact that the addition of the proximate composition did not improve classification accuracy was somewhat expected, as the information on the genetic strain carried by the compositional data was found to be very poor.

The confusion matrix for the validation data for the best model of Table 7.8 is given in Table 7.9.

Table 7.9. *Confusion matrix for the best multi-block kNN classifier (raw samples).*

	IT1	IT2	IT3	UK	USA
IT1	5	0	0	1	0
IT2	0	3	1	1	1
IT3	0	0	5	0	1
UK	0	0	0	6	0
USA	0	0	0	0	6

Five out of thirty samples were misclassified, and the majority of errors involved samples of the genetic strain IT2. Please also note that the use of the *k*NN classifier limits the interpretability of the results obtained, given that no statistics, such as the VIP index are available.

7.3 Conclusions

The problem of authenticating fresh West African goatfish (*Pseudupeneus prayensis*) fillets and raw and cooked freeze-dried rainbow trout (*Oncorhynchus mykiss*) fillets was addressed in this Chapter. Fusing the information derived from different analytical instruments (a portable VIS/NIR spectrometer, a digital camera, a texture analyzer and a colorimeter) through a multiblock approach was shown to return higher classification accuracies with respect to those obtained considering each analytical instrument separately. In the first case study, the combination of the spectral data and the color features extracted from the images allowed to obtain a 100% accuracy on both the calibration and validation datasets. Robustness of the classification models to the multiple sources of variability within the data (due to the seasonality effect, size) was ensured thanks to a D-optimal selection of the training samples. In the second case study, the combination of spectral, colorimetric and mechanical data increased the classification accuracies from 90.8 and 60%, respectively in calibration and validation, to 91.7 and 73.3%.

Conclusions and future perspectives

The possibility of relying on fast, cost-effective, non-invasive, reproducible and multivariate tools for the assessment of product quality is crucial in our global economical system, where the capability of delivering high quality products and protecting them from fraud, adulteration and counterfeiting is a key competitive factor. However, the industrial practice is still far from such ideal measurement systems. For instance, in the food and pharmaceutical industries the assessment of product quality still relies on the judgment of a panel of trained experts.

Machine vision systems and absorption spectroscopy have been demonstrated to encompass the majority of the abovementioned requirements of an ideal product quality assessment system. In order to exploit the information embedded in digital images and spectra, however, appropriate statistical tools are needed. Latent variable models (LVMs) are specifically intended to analyze these types of data, which are characterized by strong correlations (between wavelengths in a spectrum and between pixels in a digital image). The use of such modeling tools has been continuously growing in the last decades, but in most published contributions *tailored* solutions are provided. Under this perspective, this Dissertation has proposed some techniques to overcome some of the limitations that currently hinder the diffusion of these tools in the common industrial practice. Table 1 summarizes the main achievements, with the indication of the application considered, origin of the data and reference to related papers that have been published or are in press.

With respect to the use of **computer vision** systems, Chapter 3 demonstrated through an industrial case study the effectiveness of image analysis not only for the assessment of product quality, but also for process understanding purposes. Multivariate image (MIA) and wavelet texture analysis (MWTA) were used to quantify the elegance of film-coated tablets in terms of color uniformity and surface erosion. The metrics developed were then regressed against the coating operating conditions. Given the PLS regression model, the score plot was shown to be useful for monitoring the quality of the end product, while the analysis of the loadings allowed one to evaluate the effect of different physical phenomena on surface erosion.

In Chapter 4, a strategy to ensure the reproducibility of the results derived from image analysis applications was presented. Computer vision systems are usually deemed to be quick, accurate, objective and able to return reproducible results.

Table 1. *Summary of the main achievements of this Dissertation, with the indication of the considered application, of the data origin and of the relevant references.*

Chapter	Main achievement	Application	Data origin	Reference
Chapter 3	Use of image analysis for process understanding and troubleshooting	Tablet film-coating process	Industrial	Ottavian, M., M. Barolo and S. García-Muñoz. Multivariate Image and Texture Analysis to Investigate the Erosion Mechanism of Film-Coated Tablets. *J. Pharm. Innov.*, in press.
Chapter 4	Standardization of machine vision systems	Tablet film-coating process	Laboratory	Ottavian, M., M. Barolo and S. García-Muñoz (2013). Maintenance of Machine Vision Systems for Product Quality Assessment. Part I: Addressing Changes in Lighting Conditions. *Ind. Eng. Chem. Res.*, **52**, 12309-12318.
				Ottavian, M., M. Barolo and S. García-Muñoz. Maintenance of Machine Vision Systems for Product Quality Assessment. Part II: Addressing Camera Replacement. *Ind. Eng. Chem. Res.*, in press. DOI: 10.1021/ie402910z.
Chapter 5	Multispectral data classification strategy that avoids trial-and-error data preprocessing	Fraud detection	Laboratory	Ottavian, M., P. Facco, L. Fasolato and M. Barolo (2012). Multispectral Data Classification Using Similarity Factors. *Chemom. Intell. Lab. Syst.*, **118**, 13-23.
Chapter 6	Foodstuff authentication from NIR spectral data	Fraud detection	Laboratory	Ottavian, M., L. Fasolato, P. Facco and M. Barolo (2013). Foodstuff Authentication From Spectral Data: Toward a Species-Independent Discrimination Between Fresh and Frozen-Thawed Fish Samples. *J. Food Eng.*, **119**, 765-775.
	Multispecies approach for fraud detection			Ottavian, M., P. Facco, M. Barolo, P. Berzaghi, S. Segato, E. Novelli and S. Balzan (2012). Near-Infrared Spectroscopy to Assist Authentication and Labeling of Asiago d'Allevo Cheese. *J. Food Eng.*, **113**, 289-298.
				Ottavian, M., P. Facco, L. Fasolato, E. Novelli, M. Mirisola, M. Perini and M. Barolo (2012). Use of Near-Infrared Spectroscopy for Fast Fraud Detection in Seafood: Application to the Authentication of Wild European Sea Bass (*Dicentrarchus labrax*). *J. Agric. Food Chem.*, **60**, 639-648.
Chapter 7	Multi-block approach for low-level data fusion	Fraud detection	Laboratory	Ottavian, M., L. Fasolato, L. Serva, P. Facco and M. Barolo. Data Fusion for Food Authentication: Fresh/Frozen-Thawed Discrimination in West African Goatfish (*Pseudupeneus prayensis*) Fillets. *Food Bio. Tech.*, in press. DOI: 10.1007/s11947-013-1157-x.
				Dalle Zotte, A., M. Ottavian, A. Concollato, L. Serva, R. Martelli and G. Parisi. Authentication of Raw and Cooked Freeze-Dried Rainbow Trout (*Onchorynchus mykiss*) by Means of Near-Infrared Spectroscopy and Data Fusion. *Food Res. Int.*, in press. DOI: 10.1016/jfoodres.2013.10.033.

This latter advantage, however, is true only as long as the conditions, under which the quality assessment model has been calibrated, remain unchanged. These conditions include both the lighting system and the camera itself. A strategy to monitor the status of the computer vision system was proposed. The strategy is based on a MIA model built on images of four color standards. Every time one needs to operate the computer vision system to interrogate an existing quality assessment model, an image of the color standards needs to be collected, and projected onto the MIA model. If it conforms to it, the status of the system is deemed to be unchanged (within a given statistical confidence). If it does not conform to it, the use of the dynamic time warping (DTW) was proposed to align the images collected under the new conditions. The effectiveness of the DTW correction was demonstrated both in case of changes in lighting conditions and in the camera.

With respect to the use of **absorption spectroscopy**, in Chapter 5 a novel strategy for multispectral data classification was proposed. The strategy, which is based on similarity factors defined from PCA models built on the spectra, works through an assigned sequence of pretreatments of the spectral data involving the use of the discrete wavelet transform. The main advantage of the proposed strategy, which has been tested on three food engineering case studies, is the possibility of avoiding the time consuming trial-and-error procedure usually employed for the selection of the best preprocessing strategy.
The main contribution presented in Chapter 6 is a multi-species (and, possibly, species-independent) fraud detection approach. Considering the specific problem of detecting the fresh/frozen-thawed substitution fraud in sea food, published approaches typically tailor the classification model on the specific species being analyzed. The use of the orthogonal PLS decomposition was shown in Chapter 6 to be effective in calibrating a multi-species classification model, returning a very good classification accuracy also when applied to samples of species not used in the calibration step.

Eventually, with respect to the use of **data fusion** for product quality characterization, in Chapter 7 two food engineering applications were presented. In both cases, the use of a multiblock approach for coupling the available information (spectral data, digital images, mechanical properties, etc.) was shown to return better classification accuracies than those obtained by using each piece of information separately.

In summary, it has been shown how LVMs can be employed for the analysis of the signals derived from different analytical instruments for product quality characterization purposes. They demonstrated to be able to efficiently extract the information embedded in the signals, compress it and facilitate its interpretation. In addition, they have shown to be

versatile, with a broad range of applicability, i.e. from simple predictive tools to tools to support process understanding and troubleshooting.

The studies carried out in this Dissertation have opened further perspectives and issues, that could be considered in future research. One of the most interesting **areas open to further investigation** is the standardization of computer vision systems. The strategy presented in the Dissertation to transfer a quality assessment model to new illuminating conditions or to new cameras can be applied when the quality features of interest are related to product color. The possibility of transferring the quality assessment model when looking at surface texture (roughness, presence of defects, etc.) remains an open issue. Though MIA-based texture analysis techniques have been developed (Bharati and MacGregor, 2004), the state of the art is represented by wavelet texture analysis, which operates through the extraction of features from the images. When applying wavelet texture analysis, since the MIA decomposition is avoided, the use of the dynamic time warping correction proposed in Chapter 4 of the Dissertation is limited.

Further research is needed in the promising field of data fusion. Combining the available information in all possible ways to find (by trial-and-error) the one returning the best quality characterization is the standard approach usually adopted. A systematic way to select *i*) how and *ii*) which information should be combined would be highly welcomed because, as shown also in Chapter 7, the data fusion approach can enhance the characterization of the quality of the product being analyzed.

Appendix A

Algorithmic notes

In this Appendix some notes are provided on the main algorithms implemented to estimate the parameters of the latent variable models used in this Dissertation and described in Chapter 2.

A.1 Principal component analysis (PCA)

As shown in Chapter 2 (Section 2.1.1), given a dataset $\mathbf{X}$ [$I \times N$] of I samples and N variables, the parameters of a PCA model can be found through the eigenvector decomposition of matrix $\mathbf{X}^T\mathbf{X}$. In this Dissertation, this has mostly been done using singular value decomposition (SVD; Meyer, 2000) or the nonlinear iterative partial least squares algorithm (NIPALS; Wold, 1966).
The first method requires the estimation of covariance (or correlation) matrix and then would compute all the PCs of the system (as many as the variables in $\mathbf{X}$, i.e. N) at once:

$$\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad . \tag{A.1}$$

In (A.1), $\mathbf{V} = \mathbf{U}$ and they include the eigenvectors of $\mathbf{X}^T\mathbf{X}$, namely the PCA loading matrix $\mathbf{P}$. $\mathbf{S}$ is the [$N \times N$] diagonal matrix of the singular values, which coincides with the eigenvalues of $\mathbf{X}^T\mathbf{X}$. The calculation of the $\mathbf{X}^T\mathbf{X}$ matrix requires however that there are no missing data in the $\mathbf{X}$ dataset. Given that real datasets are usually characterized by the presence of missing data, the NIPALS algorithm is usually preferred.
The algorithm computes the scores and loadings of each PC in an iterative way, starting from PC1 and extracting each PC one at a time. As with SVD, PCs are found and ordered according to the amount of variance of the original dataset they capture. Starting from PC1, for each PC the algorithm calculates the scores and loadings vectors $\mathbf{t}$ and $\mathbf{p}$ from the $\mathbf{X}$ matrix. The outer product $\mathbf{tp}$ is then subtracted from $\mathbf{X}$ to give the residual matrix $\mathbf{E}$:

$$\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad . \tag{A.2}$$

$\mathbf{E}$ is then used at the next iteration to extract the scores and loadings vectors for PC2. The algorithm can be summarized in the following steps (Geladi and Kowalski, 1986):

1. Consider a row vector $\mathbf{x}_i$ from $\mathbf{X}$ and set $\mathbf{t} = \mathbf{x}_i$;
2. calculate $\mathbf{p}^T$:

$$\mathbf{p}^T = \frac{\mathbf{t}^T\mathbf{X}}{\mathbf{t}^T\mathbf{t}} \quad ; \tag{A.3}$$

3. normalize $\mathbf{p}^T$ to unit length;
4. calculate $\mathbf{t}$:

$$\mathbf{t} = \frac{\mathbf{X}\mathbf{p}}{\mathbf{p}^T\mathbf{p}} \quad ; \tag{A.4}$$

5. compare $\mathbf{t}$ used in step 2. with $\mathbf{t}$ calculated in step 4. If they differ less than an assigned tolerance, then stop (the method has converged), otherwise restart from step 2. with the last calculated $\mathbf{t}$;
6. if converged, calculate $\mathbf{E}$ according to (A.2), and go back to step 1, by setting $\mathbf{X} = \mathbf{E}$ to calculate the next PC.

The algorithm iterates until the A PCs selected to build the PCA model have been determined. It is demonstrated that the parameters provided by the NIPALS algorithm are the same as the eigenvector solution problem of (2.4) (Chapter 2, Section 2.1.1) (Geladi and Kowalski, 1986). Furthermore, it can feasibly handle datasets with missing data.

Other approaches have been used to calculate the PCA loadings and scores based on optimization frameworks for parameter estimation. In these cases, the PCA loadings are found in order to minimize the sum of squared errors between the data matrix $\mathbf{X}$ (López-Negrete de la Fuente *et al.*, 2010).

A.2 Projection to latent structures (PLS)

Projection to latent structures (PLS) includes a class of algorithms that attempts to summarize the variation in a regressor matrix $\mathbf{X}$ that is in some way predictive of a corresponding matrix $\mathbf{Y}$ of response variables (Chapter 2, Section 2.1.2) (MacGregor *et al.*, 1994). One of the most common algorithms to estimate the PLS model parameters is NIPALS (Wold, 1966; Wold *et al.*, 1983), whose steps are summarized below and illustrated in Figure A.1 (MacGregor *et al.*, 1994).

1. Set $\mathbf{u}$ equal to any column of $\mathbf{Y}$;
2. regress columns of $\mathbf{X}$ on $\mathbf{u}$ to get weights $\mathbf{w}$:

$$\mathbf{w}^T = \frac{\mathbf{u}^T\mathbf{X}}{\mathbf{u}^T\mathbf{u}} \quad ; \tag{A.5}$$

3. normalize **w** to unit length;
4. calculate the scores **t**:

$$\mathbf{t} = \frac{\mathbf{Xw}}{\mathbf{w}^T\mathbf{w}} \quad ; \tag{A.6}$$

5. regress the columns of **Y** on **t**:

$$\mathbf{q}^T = \frac{\mathbf{t}^T\mathbf{Y}}{\mathbf{t}^T\mathbf{t}} \quad ; \tag{A.7}$$

6. calculate the new score vector for **Y**:

$$\mathbf{u} = \frac{\mathbf{Yq}}{\mathbf{q}^T\mathbf{q}} \quad ; \tag{A.8}$$

7. check convergence of **u**: if yes go to step 8; if no go to step 2.;
8. calculate **X** matrix loadings, by regressing columns of **X** on **t**:

$$\mathbf{p}^T = \frac{\mathbf{t}^T\mathbf{X}}{\mathbf{t}^T\mathbf{t}} \quad ; \tag{A.9}$$

9. calculate residual matrices **E** and **F**:

$$\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad , \tag{A.10}$$

$$\mathbf{F} = \mathbf{Y} - \mathbf{t}\mathbf{q}^T \quad ; \tag{A.11}$$

10. to calculate the next set of latent vectors, restart from step 1, replacing **X** and **Y** by **E** and **F**, respectively.

(A.5) and (A.7) allow each dataset latent space to get information about the other one. Various interpretation of the PLS algorithm and its properties are discussed by Höskuldsson (1988).

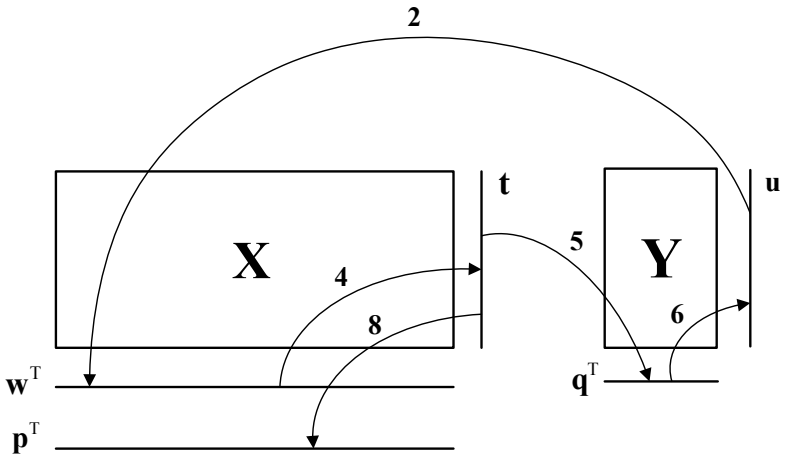


Figure A.1. *Schematic of a PLS algorithm iteration (adapted from MacGregor et al., 1994).*

A.3 Orthogonal PLS (OPLS)

The OPLS algorithm for a single column **Y** response matrix (i.e. $I = 1$) and a predictor matrix **X** (each of them properly scaled) is a modification of the NIPALS algorithm described in section A.2. Once the loading vector **p** has been calculated (step 8., (A.9)), the orthogonal components $\mathbf{w}_\perp$, $\mathbf{t}_\perp$ and $\mathbf{p}_\perp$ are evaluated as follows (Trygg and Wold, 2002).

9. Calculate the orthogonal weigths $\mathbf{w}_\perp$ from **p**:

$$\mathbf{w}_\perp = \mathbf{p} - \left[\frac{\mathbf{w}^T\mathbf{p}}{\mathbf{w}^T\mathbf{w}}\right]\mathbf{w} \quad ; \tag{A.12}$$

10. normalize $\mathbf{w}_\perp$ to unit length;
11. calculate the scores $\mathbf{t}_\perp$:

$$\mathbf{t}_\perp = \mathbf{X}\mathbf{w}_\perp \quad ; \tag{A.13}$$

12. calculate the orthogonal **X** loadings $\mathbf{p}_\perp$:

$$\mathbf{p}_\perp^T = \frac{\mathbf{t}_\perp^T\mathbf{X}}{\mathbf{t}_\perp^T\mathbf{t}_\perp} \quad ; \tag{A.14}$$

13. calculate the residual matrix $\mathbf{E}_{OPLS}$:

$$\mathbf{E}_{OPLS} = \mathbf{X} - \mathbf{t}_\perp\mathbf{p}_\perp^T \quad . \tag{A.15}$$

$\mathbf{E}_{OPLS}$ represents the residual after a portion of orthogonal variability has been removed. For additional orthogonal components $\mathbf{w}_{\perp}$, $\mathbf{p}_{\perp}$, $\mathbf{t}_{\perp}$, the algorithm needs to be repeated by setting $\mathbf{X} = \mathbf{E}_{OPLS}$.

Appendix B

Interpretation of LVM models parameters

This Appendix reports some details on the interpretation of the parameters of a latent variable model (LVM). In particular, some indications are provided on how to interpret the loading and score diagrams in order to get information from the data.

B.1 Interpretation of the score and loading plots

PCA and PLS models (Chapter 2) are usually built on multivariate datasets to gain understanding on the system the data have been generated from. This can be achieved by analyzing the correlation between variables and the similarities between samples. The advantage in using LVMs to this purpose is due to the fact that the model structure is transparent and allows to interpret the correlation structure in a straightforward way. From the analysis of the model parameters, an interpretation on the mechanisms acting on the system can be drawn.

For the purpose of a practical application of PCA and PLS, the analysis of the scores and of the loadings of the model is crucial. In general, this is done by considering plots of the scores and of the loadings, which can be reported in several ways. According to common practice, the scores are reported as scatter plots, in which the scores on a PC (or on a LV, indifferently) are reported versus the scores on another PC. This is usually done for the scores on the first LVs found by the model, because they explain most part of the variability in the data. Bi-dimensional plots are usually used as they are easier to visualize than three-dimensional ones. Figure B.1b reports an example of a score plot.

Loadings are usually reported as bar plots or as scatter plots. In the first case a bar plot of the loadings of the original variables on each PC is reported, as in Figure B.1a. In the second case, as in score plots, the loadings of the variables on a PC are plotted versus the loadings of the same variables on a different PC. This is a useful way to summarize in a single plot more exhaustive information on variable correlation. In general, loading plots are useful for two important reasons: *i*) understanding which are the variables related to the data variability and which are not; *ii*) understanding if there are correlations among the variables. Recalling the meaning of loadings in PCA and weights in PLS (Chapter 2,

Section 2.1.1 and Section 2.1.2), a measured variable which shows a high loading or weight has a significant importance on the related PC/LV, thus being responsible of a significant part of the variability in the data. Therefore, loadings in PCA and weights in PLS help in identifying the "most important" variables for the system, and to rank them by importance order. If this information is combined with physical knowledge on the system, one can obtain additional physical insights on the system under investigation, by understanding which are the driving forces linked to the physical and/or chemical phenomena that drive the system. When two variables have similar loadings on a PC, they are said to be correlated. If the loading absolute values are similar but the values are opposite, they are said to be inversely related (or anti-correlated). This means that it is expected that, considering the data used to build the model, an increase in one variable results in a decrease of the other variable. Figure B.1 gives an example of this occurrence. For example, in the top bar plot it can be clearly seen that variable x_1 and variable x_3 are the most significant variables on this direction, and they are inversely related as their loadings are opposite. Differently, on the second bar plot (which refers to PC2), x_3 has the highest loading and looks inversely related to x_1 and x_2, which have a lower importance. Note that the PCA loadings and the PLS weights on each PC/LV are independent. Therefore, the information obtained from the analysis of one latent component is not contrasting with the other ones, but it simply provides a different type of information (namely, it identifies a different driving force for the process).

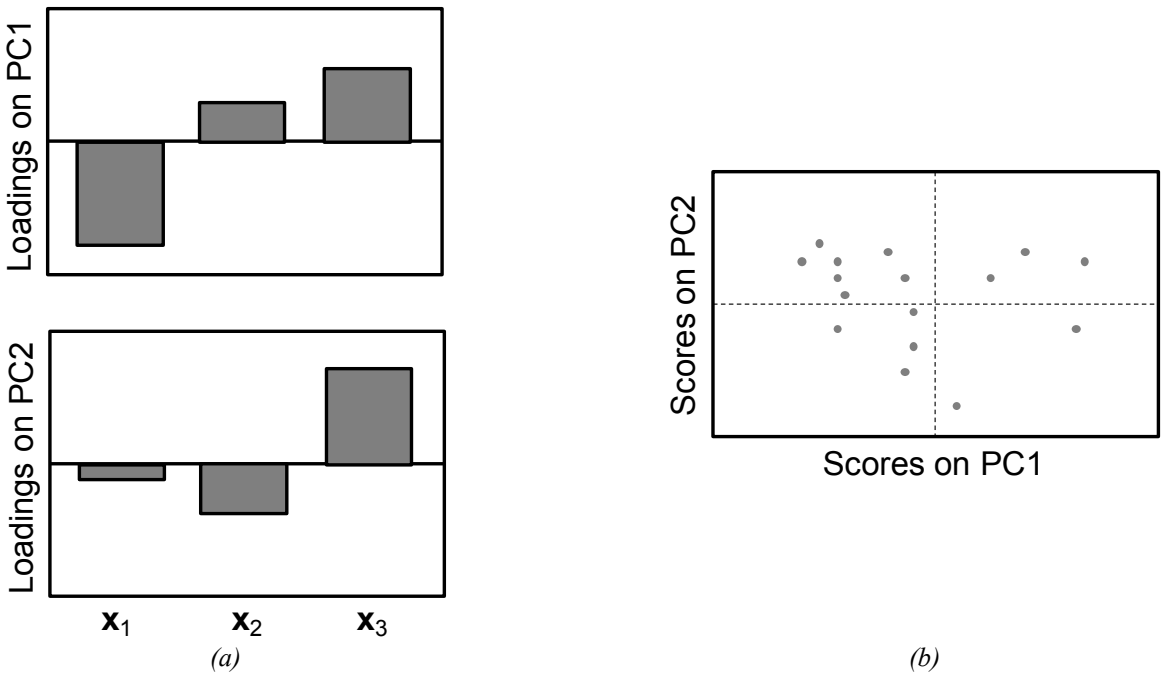


Figure A.1. *Example of (a) loading bar plots and (b) score plot for a model with 2 PCs.*

Score plots as the one reported in Figure B.1b are useful to identify similarities between samples. This means that samples with similar characteristics fall in the same region of the score plot. Moreover, the pattern observed in a score plot reflects the correlation structure identified by the variable loadings. For example, in Figure B.1b three main clusters can be observed along PC1. Samples are therefore grouped according to their similarities or differences in the variables that have the highest loading on PC1. By analyzing the loading plot, one can identify which these variables are (x_1 and x_3 in this case). Therefore, samples having a high positive score on PC1 will have higher x_3 values and lower x_1 values on average, because x_3 has a positive loading on PC1 whereas x_1 has a negative one. The situation is opposite in the case of samples with negative PC1 scores. A similar analysis can be done also for the other PCs.

Finally, note that in the PLS case it is more useful to analyze jointly the model weights with the loadings of **Y** (**Q** loadings). This analysis allows to identify cross-correlations among variables (i.e., how the regressors are related with the responses), which is of particular interest considering that PLS is a regression model built to predict the responses from the inputs. The following scaling is usually applied prior to build the scatter plot:

$$\mathbf{w}_a^* = \frac{\mathbf{w}_a^*}{\sqrt{\frac{1}{N}}\left\|\mathbf{w}_a^*\right\|} \quad , \tag{B.1}$$

$$\mathbf{q}_a = \frac{\mathbf{q}_a}{\sqrt{\frac{1}{M}}\left\|\mathbf{q}_a\right\|} \quad . \tag{B.2}$$

References

Adamopoulos, K.G., A.M. Goula and H.J. Petropakis (2001). Quality Control During Processing of Feta Cheese: NIR Application. *J. Food Compos. Anal*., **14**, 431-440.

Addis, P., M. Secci, I. Locci and A. Cau (2012). Harvesting, Handling Practices and Processing of Bluefin Tuna Captured in the Trap Fishery: Possible Effects on the Flesh Quality. Col. *Vol. Sci. Pap. ICCAT*, **67**, 390-398.

Addison, P.S. (2002). *The Illustrated Wavelet Transform Handbook*, Taylor & Francis, New York (U.S.A).

Afseth, N. K., V. H. Segtnan and J. P. Wold (2006). Raman Spectra of Biological Samples: a Study of Preprocessing Methods. *Appl. Spectrosc*., **60**, 1358-1367.

Alasalvar, C., K. D. A. Taylor, E. Zubcov, F. Shahidi and M. Alexis (2002). Differentiation of Cultured and Wild Sea Bass (*Dicentrarchus labrax*): Total Lipid Content, Fatty Acid and Trace Mineral Composition. *Food Chem*., **79**, 145-150.

Alcala, M., M. Blanco, M. Bautista and J.M. Gonzales (2010). On-Line Monitoring of a Granulation Process by NIR Spectroscopy. *J. Pharm. Sci*., **99**, 336-345.

Amigo, J.M., J.Cruz, M. Bautista, S. Maspoch, J. Coello and M. Blanco (2008). Study of Pharmaceutical Samples by NIR Chemical Image and Multivariate Analysis. *TRAC – Trend Anal. Chem.*, **27**, 696-713.

Andersen, C. M. and R. Bro (2010). Variable Selection in Regression – a Tutorial. *J. Chemom.*., **24**, 728-737.

AOAC (2002). Official Methods of Analysis of the Association of Official Analytical Chemists. Association of Official Analytical Chemists, Arlington (U.S.A.).

Arandjelović, O. and R. Cipolla (2009). A Pose-Wise Linear Illumination Manifold Model for Face Recognition Using Video. *Comput. Vis. Image Und*., **113**, 113-125.

Arvanitoyannis, I. S., E. V. Tsitsika and P. Panagiotaki (2005). Implementation of Quality Control Methods (Physicochemical, Microbiological and Sensory) in Conjunction with Multivariate Analysis Towards Fish Authenticity. *Int. J. Food Sci. Technol*., **40**, 237–263.

Aujol, J.-F., G. Aubert and L. Blanc-Feraud (2003). Wavelet-Based Level Set Evolution for Classification of Textured Images. *IEEE Trans. Image Process*., **12**, 1634-1641.

Aursand, M., L. Jørgensen and H. Grasdalen (1995). Positional Distribution of ω3 Fatty Acids in Marine Lipid Triacylglycerols by High-resolution ^{13}C Nuclear Magnetic Resonance Spectroscopy. *JAOCS*., **72**, 293-297.

Azzouz, .K, Y. Diatta, S. Mansour, M. Boumaïza, M. M. Ben Amor and C. Capapé (2011). First Record of the West African Goatfish Pseudupeneus prayensis (*Osteichthyes:*

Mullidae) Off the Tunisian Coast (Central Mediterranean). *Acta Ichthyol. Piscat.*, **41**, 133–136.

Bachelor, B.G. (1985). *Lighting and Viewing in Automated Visual Inspection*. IFS Publication Ltd., Bedford (U.K.).

Balabin, R. M., R. Z. Safieva and E. I. Lomakina (2010). Gasoline Classification Using Near Infrared (NIR) Spectroscopy Data: Comparison of Multivariate Techniques. *Anal. Chim. Acta*, **671**, 27-35.

Baldwin, E.A., J. Bai, A. Plotto and S. Dea (2011). Electronic Noses and Tongues: Applications for the Food and Pharmaceutical Industries. *Sensors*, **11**, 4744-4766.

Ballard, D.A. and C.M. Brown (1982). *Computer Vision*. Prentice-Hall, Englewood Cliffs, NJ (U.S.A.).

Barbin, D. F., D.-W. Sun and C. Su (2013). NIR Hyperspectral Imaging as a Non-Destructive Evaluation Tool for the Recognition of Fresh and Frozen-Thawed Porcine Longissimus Dorsi Muscles. *Innov. Food Sci. Emerg.*, **18**, 226-236..

Barbin, D. F., G. Elmasry, D.-W. Sun and P. Allen (2012). Near-Infrared Hyperspectral Imaging for Grading and Classification of Pork. *Meat Sci*., **90**, 259-268.

Barker, M. and W. Rayens (2003). Partial Least-Squares for Discrimination. *J. Chemom*., **17**, 166-173.

Barnes, R. J., M. S. Dhanoa and S. J. Lister (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.*, **43**, 727–777.

Begley, T.H., E. Lanza, K.H. Norris and W.R. Hruschka (1984). Determination of Sodium Chloride in Meat by Near-Infrared Diffuse Reflectance Spectroscopy. *J. Agric. Food Chem*., **32**, 984-987.

Bell, J. G., T. Preston, R. J., Henderson, F. Strachan, J. E. Bron, K. Cooper and D. J. Morrison (2007). Discrimination of Wild and Cultured European Sea Bass (*Dicentrarchus labrax*) Using Chemical and Isotopic Analyses. *J. Agric. Food Chem*., **55**, 5934-5941.

Bell, J. G., T. Preston, R. J., Henderson, F. Strachan, J. E. Bron, K. Cooper and D. J. Morrison (2007). Discrimination of Wild and Cultured European Sea Bass (*Dicentrarchus labrax*) Using Chemical and Isotopic Analyses. *J. Agric. Food Chem*., **55**, 5934-5941.

Berardo, N., V. Pisacane, P. Battilani, A. Scandolara, A. Pietri and A. Marocco (2005). Rapid Detection of Kernel Rots and Mycotoxins in Maize by Near-Infrared Reflectance Spectroscopy. *J. Agric. Food Chem*., **53**, 8128-8134.

Bharati, M. and J. F. MacGregor (2003). Softwood Lumber Grading Through On-line Multivariate Image Analysis Techniques. *Ind. Eng. Chem. Res*., **42**, 5345-5353.

Bharati, M. H., J. J. Liu and J. F. MacGregor (2004). Image Texture Analysis: Methods and Comparisons. *Chemom. Intell. Lab. Syst*., **72**, 57-71.

Bozzetta, E., M. Pezzolato, E. Cencetti, K. Varello, F. Abramo, F. Mutinelli, F. Ingravalle and E. Teneggi (2012). Histology as a Valid and Reliable Tool to Differenziate Fresh from Frozen-Thawed Fish. *J. Food Protect.*, **75**, 1536-1541.

Braspenning, P. J. (1995). *Artificial neural networks: an introduction to ANN theory and practice. Lecture notes in computer science*, Springer, Berlin.

Broad, N., P. Graham, P. Hailey, A. Hardy, S. Holland, S. Hughes, D. Lee, K. Prebble, N. Salton and P. Warren (2006). Guidelines for the Development and Validation of Near-infrared Spectroscopic Methods in the Pharmaceutical Industry in *Handbook of Vibrational Spectroscopy*, John Wiley & Sons, New York (U.S.A.).

Brosnan, T. and D.-W. Sun (2004). Improving Quality Inspection of Food Products by Computer Vision-a Review. *J. Food Eng*., **61**, 3-16.

Bruwer, M.-J., J. F. MacGregor and W. M. Bourg Jr. (2007) Fusion of Sensory and Mechanical Testing Data to Define Measures of Snack Food Texture. *Food Qual. Prefer*., **18**, 890-900.

Büning-Pfaue, H. (2003). Analysis of Water in Food by Near Infrared Spectroscopy. *Food Chem.*, **82**, 107–115.

Burger, J. and P. Geladi (2006). Hyperspectral NIR Image Regression Part II: Dataset Preprocessing Diagnostic. *J. Chemom.*, **20**, 106-119.

Burnham, A.J., R. Viveros and J.F. MacGregor (1996). Frameworks for Latent Variable Multivariate Regression. *J. Chemom.*, **10**, 31-45.

Burns, D.A. and E.W. Ciurczak (2008). *Handbook of Near-Infrared Analysis*. Taylor & Francis, New York (U.S.A.).

Bylesjö, M., M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes and J. Trygg (2006). OPLS Discriminant Analysis: Combining the Strengths of PLD-DA and SIMCA Classification. *J. Chemom.*, **20**, 341-351.

Candolfi, A., R. De Maesschalck, D. Jouan-Rimbaud, P. A. Hailey and D. L. Massart (1999). The Influence of Data Pre-processing in the Pattern Recognition of Excipients Near-Infrared Spectra. *J. Pharmaceut. Biomed*., **21**, 115-132.

Casale, M., C. Casolino, P. Oliveri and M. Forina (2010) The Potential of Coupling Information Using Three Analytical Technologies for Identifying the Geographical Origin of Ligurian Extra Virgin Oil. *Food Chem.*, **118**, 163-170.

Castiñeira, D., B. C. Rawlings and T. F. Edgar (2012). Multivariate Image Analysis (MIA) for Industrial Flare Combustion Control. *Ind. Eng. Chem. Res*., **51**, 12642-12652.

Chalmers, J.M. and P.R. Griffiths (2002). *Handbook of Vibrational Spectroscopy*. John Wiley & Sons Inc., New York (U.S.A.).

Chen, Q., J. Zhao and H. Lin (2009). Study on Discrimination of Roast Green Tea (*Camellia Sinensis L.*) According to Geographical Origin by FT-NIR Spectroscopy and Supervised Pattern Recognition. *Spectrochim. Acta A*, **72**, 845-850.

Chen, Y.R., W.R. Hruschka and H. Early (2003). Online Inspection of Poultry Carcasses Using a Visible/Near-Infrared Spectrophotometer. *P. SPIE*, **3544**, 146-155.

Chong, I.-G. and C. H. Jun (2005). Performance of Some Variable Selection Methods When Multicollinearity is Present. *Chemom. Intell. Lab. Syst.*, **78**, 103-112.

Christie, W. (1982). *Lipid analysis. Isolation, Separation, Identification and Structural Analysis of Lipids*. Pergamon Press, Oxford (U.K.).

CIE Colorimetry Committee (1974). Working Program on Color Differences. *J. Opt. Soc. Am.*, **64**, 896–899.

Cocchi, M., R. Seeber and A. Ulrici (2001). WPTER: Wavelet Packet Transform for Efficient Pattern Recognition of Signals. *Chemom. Intell. Lab. Syst.*, **57**, 97-119.

Cole, G., J.E. Hogan and M.E. Aulton (1995). *Pharmaceutical Coating Technology*. London: Taylor & Francis.

Collomb, M., R. Sieber, U. Bütikofer, W. Stoll, H. Sollberger, D. Wechsler and P. Gallmann (2003). Impact of Oilseeds and Grass-feeding at Different Altitudes on Fatty Acid Composition of Milk Lowlands, Mountains and Highlands, and Typical Winter. Presented at *5th International Meeting on Mountain Cheeses*, Arèches (France) 11-12 September.

Commision Regulation (EC) No 1107/96 of 12 June 1996 on the Registration of Geographical Indications and Designations of Origin Under the Procedure Laid Down in Article 17 of Council Regulation (EEC) No 2081/92. L148 of the 21/06/1996 (pp. 1-10).

Commission of the European Communities (2008). Green Paper on Agricultural Product Quality: Product Standards, Farming Requirements and Quality Schemes.

Commission Regulation (EC) No. 2065/ 2001 of October 2001 laying down detailed rules for the application of Council Regulation (EC) No. 104/2000 as regards informing consumers about fishery and aquaculture products.

Conlin, A. K., E. B. Martin and A. J. Morris (2000) Confidence Limits for Contribution Plots. *J. Chemom*., **14**, 725-736.

Connell, J. J. (1995). Quality Deterioration and Effects in Products. In: *Control of Fish Quality* (4th ed). Wiley-Blackwell.

Costa, C., S. D'Andrea, R. Russo, F. Antonucci, F. Pallottino and P. Menesatti (2010). Application of Non-invasive Techniques to Differentiate Sea Bass (*Dicentrarchus labrax*, L. 1758) Quality Cultured Under Different Conditions. *Aquacult. Int*., **19**, 765-778.

Cozzi,, G., J. C. Ferlito, G. Pasini, B. Contiero and F. Gottardo (2009). Application of Near-Infrared Spectroscopy as an Alternative to Chemical and Color Analysis to Discriminate the Production Chains of Asiago d'Allevo Cheese. *J. Agric. Food Chem*., **57**, 11449-11454.

Cozzolino, D. and I. Murray. (2012). A Review on the Application of Infrared Technologies to Determine and Monitor Composition and Other Quality Characteristics in Raw Fish, Fish Products and Seafood. *Appl. Spectrosc. Rev*., **47**, 207-218.

Cozzolino, D., E. Restaino, A. La Manna, E. Fernandez and A. Fassio (2009). Usefulness of Near Infrared Reflectance (NIR) Spectroscopy and Chemometrics to Discriminate Between Fishmeal, Meat Meal and Soya Meal Samples. *Cienc. Investig. Agrar.*, **36**, 209-214.

Cozzolino, D., I. Murray and J. R. Scaife (2002). Near Infrared Reflectance Spectroscopy in the Prediction of Chemical Characteristics of Minced Raw Fish. *Aquacult. Nutr*., **8**, 1-6.

Cozzolino, D., M. Parker, R.G. Dambergs, M. Herderich and M. Gishen (2006). Chemometrics and Visible-Near Infrared Spectroscopic Monitoring of Red Wine Fermentation in a Pilot Scale. *Biotechnol. Bioeng*., **95**, 1101-1107.

Cross, G. and A. Jain (1983). Markov Random Field Texture Models. *IEEE T. Pattern Anal.*, **5**, 25-39.

de Aguiar, P. F., B. Bourguignon, M. S. Khots, D. L. Massart, R. Phan-Than-Luu (1995). D-Optimal Designs. *Chemom. Intell. Lab. Syst*., **30**, 199-210.

de Beer, T.R.M., C. Bodson, B. Dejaegher, B. Walczak, P. Vercruysse, A. Burggraeve, A. Lemos, L. Delattre, Y.V. Heyden, J.P. Remon, C. Vervaete and W.R.G. Baeyens (2008). Raman Spectroscopy as a Process Analytical Technology (PAT) Tool for the In-line Monitoring and Understanding of a Powder Blending Process. *J. Pharm. Biomed. Anal*., **48**, 772-779.

de la Roza-Delgado, B., A. Soldado, A. Martinez-Fernández, F. Vicente, A. Garrido-Varo, D. Pérez-Marin, M. J. de la Haba and J. E. Guerrero-Ginel (2007). Application of Near-Infrared Microscopy (NIRM) for the Detection of Meat and Bone Meals in Animal Feeds: a Tool for Food and Feed Safety. *Food Chem*., **105**, 1164-1170.

Deisingh, A.K. (2005). Pharmaceutical Counterfeiting. *Analyst*, **130**, 271-279.

Duchesne, C., J. J. Liu and J. F. MacGregor (2012). Multivariate Image Analysis in the Process Industries: a Review. *Chemom. Intell. Lab. Syst*., **117**, 116-128.

Duflos, G., B. Le Fur, V. Mulak, P. Becel and P. Malle (2002). Comparison of Methods of Differentiating Between Fresh and Frozen-Thawed Fish or Fillets. *J. Sci. Food Agr.*, **82**, 1341-1345.

El-Hagrasy, A.S., F. D'Amico and J.K. Drennen (2006). A Process Analytical Technology Approach to Near-Infrared Process Control of Pharmaceutical Powder Blending. Part I:

D-Optimal Design for Characterization of Powder Mixing and Preliminary Spectra Data Evaluation. *J. Pharm. Sci*., **95**, 392-406.

El-Hagrasy, A.S., H.R. Morris, F. D'Amico, R.A. Lodder and J.K. Drennen (2001). Near-Infrared Spectroscopy and Imaging for the Monitoring of Powder Blend Homogeneity. *J. Pharm. Sci*., **90**, 1298-1307.

Elmasry, G., M. Kamruzzaman, D.-W. Sun and P. Allen (2012). Principles and Applications of Hyperspectral Imaging in Quality Evaluation of Agro-food Products: A Review. *Cr. Rev. Food Sci*., **52**, 999-1023.

Eriksson, L., E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström and S. Wold (2006). *Multi- and megavariate data analysis. Part I. Basic principles and applications*. Umetrics AB, Umeå (Sweden).

Esbensen, K. and P. Geladi (1989). Strategy of Multivariate Image Analysis (MIA). *Chemom. Intell. Lab. Syst*., **7**, 67-86.

European Parliament Legislative Resolution of July, 6, 2011 on the Provision of Food Information to Consumers, amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 and repealing Directives 87/250/EEC, 90/496/EEC, 1999/10/EC, 2000/13/EC, 2002/67/EC, 2008/5/EC and Regulation (EC) No 608/2004 (17602/1/2010 – C7-0060/2011 – 2008/0028(COD).

Facco, P., E. Tomba, M. Roso, M. Modesti, F. Bezzo and M. Barolo (2010). Automatic Characterization of Nanofiber Assemblies by Image Texture Analysis. *Chemom. Intell. Lab. Syst.*, **103**, 66-75.

Fagan, C.C., C. P. O'Donnell, D. J. O'callaghan, G. Downey, E. M. Sheehan, C. M. Delahunty, C. Everard, T. P. Guinee and V. Howard (2007). Application of Mid-Infrared Spectroscopy to the Prediction of Maturity and Sensory Texture Attributes of Cheddar Cheese. *Food Bio. Tech.*, **72**, 130-137.

Fasolato, L., A. Manfrin, C. Corrain, A. Perezzani, G. Arcangeli, M. Rosteghin, E. Novelli, L. M. Lopparelli, S. Balzan, M. Mirisola, L. Serva, S. Segato and E. Bianchi (2008). Assessment of Quality Parameters and Authentication in Sole (*Solea vulgaris*) by NIRS (Near Infrared Reflectance Spectroscopy). *Ind. Aliment.-Italy*, **47**, 355-361.

Fasolato, L., E. Cencetti, R. Riovanto, M. Mirisola, E. Novelli, S. Balzan, L. Serva, J. C. Ferlito, F. Benozzo, E. M. Teneggi and P. Berzaghi (2010b). Validation of Near Infrared Spectroscopy Analysis in Authentication of Fresh and Frozen-Thawed Fish Products. Edited preprints of *14th International Conference on NIR Spectroscopy* (S. Saranwong, S. Kasemsumran, W. Thanapase and P.Williams, Eds.), Bangkok (Thailand), 7-16 November.

Fasolato, L., E. Novelli, L. Salmaso, L. Corain, F. Camin, M. Perini, P. Antonetti, and S. Balzan (2010). Application of Nonparametric Multivariate Analysis to the Authentication of Wild and Farmed European Sea Bass (*Dicentrarchus labrax*). Results

of a Survey on Fish Sampled in the Retail Trade. *J. Agric. Food Chem*., **58**, 10979-10988.

Fasolato, L., S. Balzan, K. Valentini, J. C. Ferlito, R. Riovanto, M. Mirisola, E. Cencetti, L. Serva, F. Benozzo, E. M. Teneggi, P. Berzaghi and E. Novelli (2010a). Non-Destructive Non Touch Visible-NIR Transmittance Spectroscopy for Identification of Fresh and Frozen-Thawed Fish. Presented at *4th Conference NIR On The Go*, Legnaro (PD, Italy), May 27-28.

Fasolato, L., S. Balzan, R. Riovanto, P. Berzaghi, M. Mirisola, J. C. Ferlito, L. Serva, F. Benozzo, R. Passera, V. Tepedino and E. Novelli (2012). Comparison of Visible and Near-Infrared Reflectance Spectroscopy to Authenticate Fresh and Frozen-Thawed Swordfish (*Xiphias gladius L*). *J. Aquat. Food Prod. T*., **21**, 493-507.

FDA (2004a). Good Manufacturing Practices (GMPs) for the 21st Century – Food Processing.

FDA (2004b). Guidance for Industry. PAT – A Framework for Innovative Pharmaceutical Development, Manufacturing and Quality Assurance. Center for Drug Evaluation and Research, U.S. Food and Drug Administration.

FDA (2004c). Pharmaceutical CGMPs for the 21st Century – A Risk Based Approach. Final Report. U.S. Department of Health and Human Services. U.S. Food and Drug Administration.

FDA (2005). Food CGMP Modernization – A Focus on Food Safety. Center for Food Safety and Nutrition. U.S. Food and Drug Administration.

FDA (2010). Guidance for Industry. Standards for Securing the Drug Supply Chain – Standardized Numerical Identification for Prescription Drug Packages. Center for Drug Evaluation and Research, U.S. Food and Drug Administration.

FDA (2011a). Guidance for Industry. Incorporation of Physical-Chemical Identifiers into Solid Oral Dosage Form Drug Products for Anticounterfeiting. Center for Drug Evaluation and Research, U.S. Food and Drug Administration.

FDA (2011b). Pathway to Global Product Safety and Quality. U.S. Department of Health and Human Services. U.S. Food and Drug Administration.

Fernández-Segovia, I., A. Fuentes, M. Aliño, R. Masot, M. Alcañiz and J. M. Barat (2012). Detection of Frozen-Thawed Salmon (*Salmo salar*) by a Rapid Low-Cost Method. *J. Food Eng*., **113**, 210–216.

Fissore, D., R. Pisano and A.A. Barresi (2013). Applying Quality-by-Design to Develop a Coffee Freeze-Drying Process. *J. Food Eng*., **123**, 179-187.

Flores-Cerillo, J. and J.F. MacGregor (2003). Within-Batch and Batch-to-Batch Inferential-Adaptive Control of Semibatch Reactors: a Partial Least Squares Approach. *Ind. Eng. Chem. Res*., **42**, 3334-3345.

Flores-Cerillo, J. and J.F. MacGregor (2004). Control of Batch Product Quality by Trajectory Manipulation Using LV Models. *J. Process Contr.*, **14**, 539-553.

Flores-Cerillo, J. and J.F. MacGregor (2005). Latent Variable MPC for Trajectory Tracking in Batch Processes. *J. Process Contr.*, **15**, 651-663.

Folch, J., M. Less and G.H.S. Stanley (1957). A Simple Method for the Isolation and Purification of Total Lipids from Animal Tissues. *J. Biol. Chem*., **226**, 497-509.

Forshed, J., I. Schuppe-Koistinen and S. P. Jacobsson (2003). Peak Alignment of NMR Signals by Means of a Genetic Algorithm. *Anal. Chim. Acta*, **487**, 189-199.

Freireich, B., W.R. Ketterhagen and C. Wassgren (2011). Intra-Tablet Coating Variability for Several Pharmaceutical Tablet Shapes. *Chem Eng Sci*., **66**, 2535-2544.

García-Muñoz, S. and A. Carmody (2010). Multivariate Wavelet Texture Analysis for Pharmaceutical Solid Product Characterization. *Int. J. Pharm*., **398**, 97-106.

García-Muñoz, S. and C.A. Oksanen (2010). Process Modeling and Control in Drug Development and Manufacturing. *Comput. Chem. Eng.*, **34**, 1007-1008.

García-Muñoz, S. and D. Gierer (2010). Coating Uniformity Assessment for Colored Immediate Release Tablets Using Multivariate Image Analysis. *Int. J. Pharm*., **395**, 104-113.

Garvin, D. (1984). What Does "Product Quality" Really Mean? *Sloan Manage. Rev*., 25-43.

Gazzetta Ufficiale (2006). Disciplinare di Produzione DOP "Asiago". Gazzetta Ufficiale della Repubblica Italiana N. 190 del 17/08/2006 (in Italian).

Geladi, P. (1992). Some Special Topics in Multivariate Image Analysis. *Chemom. Intell. Lab. Syst.*, **45**, 375-390.

Geladi, P. and B. R. Kowalski (1986). Partial Least-squares Regression: a Tutorial. *Anal. Chim. Acta*, **185**, 1-17.

Geladi, P. and K. Esbensen (1989). Can Image Analysis Provide Information Useful in Chemistry? *J. Chemom*., **3**, 419-429.

Geladi, P., H. Isaksson, L. Lindqvist, S. Wold and K. Esbensen (1989). Principal Component Analysis of Multivariate Images. *Chemom. Intell. Lab. Syst*., **5**, 209-220.

Gjerde, B., and H. Martens (1987). Predicting Carcass Composition of Rainbow Trout by Near-Infrared Spectroscopy. *J. Anim. Breed. Genet*., **104**, 121–136.

Golic, M. and K.B. Walsh (2006). Robustness of Calibration Models Based on Near Infrared Spectroscopy for the In-Line Grading of Stonefruit for Total Soluble Content. *Anal. Chim. Acta*, **555**, 286-291.

Gonzales, R.C., R.E. Woods and S.L. Eddins (2009). *Digital Image Processing Using MATLAB*. Gatesmark Publishing, USA.

Gonzáles-Martín, M. I., P. Severiano Pérez, I. Revilla, A. M. Vivar-Quintana, J. M. Hernández- Hierro, C. Gonzáles- Pérez and I. A. Lobos-Ortega (2011). Prediction of

Sensory Attributes of Cheese by Near-infrared Spectroscopy. *Food Chem*., **127**, 256-263.

Gordon, K.C. and C.M. McGoverin (2011). Raman Mapping of Pharmaceuticals. *Int. J. Pharm.*, **417**, 151-162.

Gray, V., G. Kelly, M. Xia, C. Butler, S. Thomas and S. Mayock (2009). The Science of USP 1 and 2 Dissolution: Present Challenges and Future Relevance. *Pharm. Res*., **26**, 1289-1302.

Guillerm-Regost, C., T. Haugen, R. Nortvedt, M. Carlehöug, B. T. Lunestad, A. Kiessling and A. M. Rørár (2006). Quality Characterization of Farmed Atlantic Halibut During Ice Storage. *J. Food Sci*., **71**, 83-90.

Gümüs, B. and M. O. Balaban (2012). Prediction of the Weight of Aquacultured Rainbow Trout (*Oncorhynchus mykiss*) by Image Analysis. *J. Aquat. Food Prod. T.*, **19**, 227-237.

Gunther, J. C., J. Baclaski, D. E. Seborg and J. S. Conner (2009). Pattern Matching in Batch Bioprocesses – Comparisons Across Multiple Products and Operating Conditions. *Computers Chem. Eng*., **33**, 88-96.

Halliday, D., R. Resnick and J. Walker (2012). *Fundamentals of physics*. John Wiley & Sons, New York (U.S.A.).

Haralick, R., K. Shanmugan and I. Dinstein (1973). Textural Features for Image Classification. *IEEE T. Syst. Man Cyb.*, **3**, 610-621.

He, D.J., T. Maekawa and H. Morishima (2001). Detecting Device for On-Line Detection of Internal Quality of Fruits Using Near-Infrared Spectroscopy and Related Experiments. *Transactions of the Chinese Society of Agricultural Engineering*, **17**, 146-148.

Herland, H., M. Esaiassen, M. Cooper and R. L. Olsen (2010). Quality of Farmed Cod: Effects of Season and Storage. *Aquac. Res*., **41**, 1203–1210.

Hernández, J.A., B. Heyd and G. Trystam (2008). On-line Assessment of Brightness and Surface Kinetics During Coffee Roasting. *J. Food Eng*., **87**, 314-322.

Hildrum, K.I., B.N. Nilsen, F. Westad and N.M. Wahlgren (2004). In-Line Analysis of Ground Beef Using a Diode Array Near Infrared Instrument on a Conveyor Belt. *J. Near Infrared Spec*., **12**, 367-376.

Ho, L., R. Mueller, M. Roemer, K.C. Gordon, J. Heinämäki, P. Kleinebudde, M. Pepper, T. Rades, Y.C. Shen, C.J. Strachan, P.F. Taday and J.A. Zeitler (2007). Analysis of Sustained-Release Tablet Film Coats Using Terahertz Pulsed Imaging. *J Control Release*., **119**, 253-261.

Höskuldsson, A. (1988). PLS Regression Methods. *J. Chemom.*, **2**, 211-228.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psycol.*, **24**, 417-441.

Hu, W., M. Hu, X. Zhou, T. Tan, J. Lou and S. Maybank (2006). Principal Axis-Based Correspondence Between Multiple Cameras for People Tracking. *IEEE T. Pattern Anal*., **28**, 663-671.

Huang, H., H. Yu, H. Xu and Y. Ying (2008). Near Infrared Spectroscopy for On/In-Line Monitoring of Quality in Foods and Beverages: a Review. *J. Food. Eng*., **87**, 303-313.

ICH (2006). Quality Risk Management Q9.

Ilie, A. and G. Welch (2005). Ensuring Color Consistency Across Multiple Cameras. Edited preprints of *10th IEEE International Conference on Computer Vision*, Beijing (China), 15-21 October.

International Standards Organization (2005). ISO 9000. Quality Management Systems – Fundamentals and Vocabulary.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. John Wiley & Sons Inc., New York (U.S.A.).

Jacquet, J.L. and D. Pauly (2008). Trade Secrets: Renaming and Mislabeling of Seafood. *Mar. Policy*, **32**, 309-318.

Jaeckle, C.M. and J.F. MacGregor (1998). Product Design Through Multivariate Statistical Analysis of Process Data. *AIChE J*., **44**, 1105-1118.

Jaeckle, C.M. and J.F. MacGregor (2000a). Industrial Application of Product Design Through the Inversion of Latent Variable Models. *Chemom. Intell. Lab. Syst*., **50**, 199-210.

Jaeckle, C.M. and J.F. MacGregor (2000b). Product Transfer Between Plants Using Historical Process Data. *AIChE J*., **46**, 1989-1997.

Jahr, I. (2007). Lighting in Machine Vision. In: *Handbook of Machine Vision* (Hornberg, A. Ed.), Wiley-Vch, Weinheim (Germany).

Javed, O., Z. Rasheed, K. Shafique and M. Shah (2003). Tracking Across Multiple Cameras with Disjoint Views. Edited preprints of *9th IEEE Conference on Computer Vision*, Nice (France), 13-16 October.

Jennerich, R. J. (1977). Stepwise Discriminant Analysis In Statistical Methods for Digital Computers, John Wiley & Sons, New York (U.S.A.).

Jiang, S. T. and T. C. Lee (2006). Processing Frozen Seafoods. In: *Handbook of Food Products Manufacturing* (Y. H. Hui Ed.), John Wiley & Sons, Hoboken (NJ, U.S.A.).

Johnson, R.A. and D. W. Wichern (2007). *Applied multivariate statistical analysis (6th ed.)*. Pearson Education, Inc., Upper Saddle River, NJ (U.S.A.).

Johnston, I. A. (2001) Implications of Muscle Growth Patterns for the Colour and Texture of Fish Flesh. In: *Farmed Fish Quality* (S. C. Kestin and P. D. Warris Eds.), Fishing New Books, London, (U.K.).

Karoui, R. and J. De Baerdemaeker (2007). A Review of the Analytical Methods Coupled with Chemometric Tools for the Determination of the Quality and Identity of Dairy Products. *Food Chem*., **102**, 621-640.

Karoui, R., B. Lefur, C. Grondin, E. Thomas, C. Demeulemester, J. De Baerdemaeker and A.-S. Guillard (2007). Mid-Infrared Spectroscopy as a New Tool for the Evaluation of Fish Freshness. *Int. J. Food Sci. Tech*., **42**, 57–64.

Karoui, R., E. Thomas and E. Dufour (2006). Utilisation of a Rapid Technique Based on Front-Face Fluorescence Spectroscopy for Differentiating Between Fresh and Frozen–Thawed Fish Fillets. *Food Res. Int*., **39**, 349–355.

Karoui, R., L. Pillonel, E. Schaller, J. O. Bosset and J. De Baerdemaeker (2006). Prediction of Sensory Attributes of European Emmental Cheese Using Near-Infrared Spectroscopy: A Feasibility Study. *Food Chem.*, **101**, 1121–1129.

Kassidas, A., J. F. MacGregor and P. A. Taylor (1998). Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J*., **44**, 864-875.

Kawamura, S., M. Natsuga, K. Takekura and K. Itoh (2003a). Development of an Automatic Rice-Quality Inspection System. *Comput. Electron. Agr*., **40**, 115-126.

Kawamura, S., M. Tsukahara, M. Natsuga and K. Itoh (2003b). On-Line Near Infrared Spectroscopic Sensing Technique for Assessing Milk Quality During Milking. *T. ASAE*, paper 033026.

Kennedy, J.F., C.A. White and A.J. Browne (1985). Application of Infrared Reflectance Spectroscopy to the Analysis of Milk and Dairy Products. *Food Chem*., **16**, 115-131.

Ketterhagen, W.R. (2011). Modeling the Motion and Orientation of Various Pharmaceutical Tablet Shapes in a Film Coating Pan Using DEM. *Int J Pharm*., **409**, 137-49.

Khan, S. and M. Shah (2003). Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View. *IEEE T. Pattern Anal*., **25**, 1355-1360.

Kim, M., H. Chung, Y. Woo and M.S. Kempes (2007). A new non-invasive, quantitative Raman technique for the determination of an active ingredient in pharmaceuical liquids by direct measurement through a plastic bottle. *Anal. Chim. Acta*, **587**, 200-207.

Kimiya, T., A. H. Sivertsen and K. Heia (2013). VIS/NIR Spectroscopy for Non-Destructive Freshness Assessment of Atlantic Salmon (*Salmo salar L*) Fillets. *J. Food Eng*., 116, 758-764.

Krzanowski, W. J. (1979). Between-Groups Comparison of Principal Components. *J. Am. Stat. Assoc.*, **74**, 703-707.

Laintinen, N., O. Antikainen, J. Rantanen and J. Yliruusi (2004). New Perspectives for Visual Characterization of Pharmaceutical Solids. *J. Pharm. Sci*., **93**, 165-176.

Lakshminarayanan, S., H. Fuji, B. Grosman, E. Dassau, and D.R. Lewin (2000). New Product Design via Analysis of Historical Databases. *Comput. Chem. Eng*., **24**, 671-676.

Lawless, H.T. and H. Heymann (2010). *Sensory Evaluation of Food. Principle and Practice*. Springer, New York (U.S.A.).

Leduc, F., F. Krzewinski, B. Le Fur, A. N'Guessan, P. Malle, O. Kol and G. Duflos (2012). Differentiation of Fresh and Frozen/Thawed Fish, European Sea Bass (*Dicentrarchus labrax*), Gilthead Seabream (*Sparus aurata*), Cod (*Gadus morhua*) and Salmon (*Salmo salar*), Using Volatile Compounds by SPME/GC/MS. *J. Sci. Food Agr.*, **92**, 2560-2568.

Lewis, J.P. (1995). Fast Template Matching. Edited preprints of *Vision Interface*, Quebec City (Canada), May 15-19.

Libermen, H. and L. Lachman (1981). *Pharmaceutical Dosage Forms:Tablets*. New York: Marcel Dekker Inc.

Lignitto, L., V. Cavatorta, S. Balzan, G. Gabai, G. Galaverna, E. Novelli, S. Sforza and S. Segato (2010). Angiotensin-Converting Enzyme Inhibitory Activity of Water-Soluble Extracts of Asiago d'Allevo Cheese. *Int. Dairy J.*, **20**, 11-17.

Liu, D., X.-A. Zeng and D.-W. Sun (2013). NIR Spectroscopy and Imaging Techniques for Evaluation of Fish Quality – A Review. *Appl. Spectrosc. Rev.*, **48**, 609-628.

Liu, J. J. (2004). Machine Vision for Process Industries: Monitoring, Control, and Optimization of Visual Quality of Processes and Products. *Ph.D. Thesis*, McMaster University, Canada.

Liu, J.J. and C. Han (2011). Wavelet Texture Analysis in Process Industries. *Korean J. Chem. Eng.*, **28**, 1814-1823.

Liu, J.J. and J.F. MacGregor (2007). On the Extraction of Spatial and Spectral Information from Images. *Chemom. Intell. Lab. Syst.*, **85**, 119-130.

Liu, Y. and Y.-R. Chen (2001). Two-Dimensional Visible/Near-Infrared Correlation Spectroscopy Study of Thawing Behavior of Frozen Chicken Meats Without Exposure to Air. *Meat Sci.*, **57**, 299-310.

López-Negrete de la Fuente, R., S. García-Muñoz and L.T. Biegler (2010). An Efficient Nonlinear Programming Strategy for PCA Models with Incomplete Data Sets. *J. Chemom.*, **24**, 301-311.

Lucas, A., D. Andeuza, E. Rock, and B. Martin (2008). Prediction of Dry Matter, Fat, pH, Vitamins, Minerals, Carotenoids, Total Antioxidant Capacity, and Color in Fresh and Freeze-Dried Cheeses by Visible-Near-Infrared Reflectance Spectroscopy. *J. Agric. Food Chem.*, **56**, 6801-6808.

Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S. and J. A. K. Suykens (2010). A Tutorial on Support Vector Machine-Based Methods for Classification Problems in Chemometrics. *Anal. Chim. Acta*, **665**, 129-145.

Ma, H. and C.A. Anderson (2008). Characterization of Pharmaceutical Powder Blends by NIR Chemical Imaging. *J. Pharm. Sci.*, **97**, 3305-3320.

Majolini, D., A. Trocino, G. Xiccato and A. Santulli (2009). Near Infrared Reflectance Spectroscopy (NIRS) Characterization of European Sea Bass (*Dicentrarchus labrax*) from Different Rearing Systems. *Ital. J. Anim. Sci*., **8**, 860-862.

Malacarne, M., P. Formaggioni, P. Franceschi, A. Summer and P. Mariani (2006). Proteolysis and Lipolysis of Parmigiano-Reggiano Cheese at Different Ripening Periods: 12, 24, 55 and 96 months. *Ann. Fac. Medic. Vet. di Parma*, XXVI, 145-164.

Mallat, S.G. (1989). A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE T. Pattern Anal.*, **11**, 674-693.

Mamani-Linares, L.W., C. Gallo and D. Alomar (2012). Identification of Cattle, Llama and Horse Meat by Near Infrared Reflectance or Transflectance Spectroscopy. *Meat Sci*., **90**, 378–385.

Mantanus, J., E. Ziémons, P. Lebrun, E. Rozet, R. Klinkenberg, B. Streel, B. Evrard and P. Hubert (2010). Active content determination of non-coated pharmaceutical pellets by near-infrared spectroscopy: method development, validation and reliability evaluation. *Talanta*, **80**, 1750-1757.

Marchesini, G., S. Balzan, S. Segato, E. Novelli and I. Andrighetto (2009). Colour Traits in the Evaluation of the Ripening Period of Asiago Cheese. *Ital. J. Anim. Sci*., **8**, 411-413.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate analysis*. Academic Press Limited, London (U.K.).

Martens, H. and T. Naes (1989). *Multivariate calibration*, John Wiley & Sons Ltd (U.S.A.).

Martinez, I. (2006). Revision of Analytical Methodologies to Verify the Production Method of Fish. In: *Seafood from Fish to Dish, Quality, Safety and Processing of Wild and Farmed Fish* (J. B. Luten, C. Jacobsen, K. Bekaert, A. Sæbo, J. Oehlenschlager, Eds.), Wageningen Academic Publishers, Wageningen, (The Netherlands), pp. 541-550.

Martinez, I., I. Stendhal, M. Aursand, Y. Yamashita and M. Yamashita (2009). Chapter 14: Analytical Methods to Differentiate Farmed from Wild Seafood in *Handbook of Seafood and Seafood Products Analysis* (L. M. L. Nollet, F. Toldrá. Eds) CRC Press Boca Raton, Florida, (U. S. A.), pp. 215 -232.

Martinez, I., M. Aursand, U. Erikson, T.E. Singstad, E. Veliyulin and C. van der Zwaag (2003). Destructive and Non-Destructive Analytical Techniques for Authentication and Composition Analyses of Foodstuffs. *Trends Food Sci. Tech*., **14**, 489-498.

Masoum, S., C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi and D. N. Rutledge (2007). Application of Support Vector Machines to ^{1}H NMR Data of Fish Oils: Methodology for the Confirmation of Wild and Farmed Salmon and Their Origins. *Anal. Bioanal. Chem*., **387**, 1499-1510.

Matero, S., F. van den Berg, S. Poutiainen, J. Rantanen and J. Pajander (2013). Towards Better Process Understanding: Chemometrics and Multivariate Measurements in Manufacturing of Solid Oral Dosage Forms. *J. Pharm. Sci*., **102**, 1385-1403.

McClure, W. F. and D. L. Stanfield (2002). Near-infrared Spectroscopy of Biomaterials. In: *Handbook of vibrational spectroscopy*. John Wiley & Sons Ltd (U.S.A.).

McGlone, V.A., P.J. Martinses, C.J. Clark and R.B. Jordan (2005). On-Line Detection of Brownheart in Braeburn Apples Using Near Infrared Transmission Measurements. *Postharvest Bio. Tech*., **37**, 142-151.

McSweeney, P. L. H. (2004). Biochemistry of Cheese Ripening. *Int. J. Dairy Technol.*, **57**, 127-144.

Mendoza, F. and J.M. Aguilera (2004). Application of Image Analysis for Classification of Ripening Bananas. *J. Food Sci*., **69**, 471-477.

Meyer, C.D. (2000). *Matrix analysis and applied linear algebra*. SIAM, Philadelphia, PA (U.S.A.).

Miller, W.M. and M. Zude (2002). NIR-based Sensing Coupled with Physical/Color Features to Identify Brix Level of Florida Citrus. *T. ASAE*, paper 026037.

Misimi, E., U. Erikson and A. Skavhaug (2008). Quality Grading of Atlantic Salmon (Salmo salar) by Computer Vision. *J. Food Sci*., **73**, 211-217.

Mitra, S. and T. Acharya (2007). Gesture Recognition: a Survey. *IEEE T. Syst. Man CY. C.*, **37**, 311-324.

Moffat, A.C., S. Assi and R.A. Watt (2010). Identifying Counterfeit Medicines Using Near Infrared Spectroscopy. *J. Near Infrared Spec*., **18**, 1-15.

Montero, D., L. Robaina, M. J. Caballero, R. Ginés and M. S. Izquierdo (2005). Growth, Feed Utilization and Flesh Quality of European Sea Bass (*Dicentrarchus labrax* L.) Fed Diets Containing Vegetable Oils: a Timecourse study on the Effect of a Re-feeding Period with a 100% Fish Oil Diet. *Aquaculture*, **24**, 121-134.

Montgomery, D. C. and G. C. Runger (2010). *Applied statistics and probability for engineers*. John Wiley & Sons, New York (U.S.A.).

Moore, J.C., J. Spink and M. Lipp (2012). Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010. *J. Food Sci*., **77**, 118-126.

Morrison, J., T. Preston, J. E. Bron, R. J. Henderson, K. Cooper, F. Strachan and J. G. Bell (2007). Authenticating Production Origin of Gilthead Sea Bream (*Sparus aurata*) by Chemical and Isotopic Fingerprinting. *Lipids*, **42**, 537-545.

Morrison, W.R. and L.M. Smith (1964). Preparation of Fatty Acid Methyl Esters and Dimethylacetals from Lipids with Boron Fluoride-Methanol. *J. Lipid. Res*., **3**, 600-608.

Mukhopadhyay, R. (2007). The Hunt for Counterfeit Medicine. *Anal. Chem*., **79**, 2622-2627.

Navrátil, M., C. Cimander and C.F. Mandenius (2004). On-Line Multisensor Monitoring of Yogurt and Filmjölk Fermentations on Production Scale. *J. Agric. Food Chem*., **52**, 415-420.

Nomikos, P. and J.F. MacGregor (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, **37**, 41-59.

Nott, K. P., S. D. Evans and L. D. Hall (1999). Quantitative Magnetic Resonance Imaging of Fresh and Frozen-Thawed Trout. *Magn. Reson. Imaging*, **17**, 445-455.

Nummiaro, K., E. Koller-Meier, T. Svoboda, D. Roth and L. Van Gool (2003). Color-Based Object Tracking in Multi-Camera Environments. Edited preprints of *25th DAGM Symposium*, Magdeburg (Germany), 10-12 September.

Ozbay, G., K. Spencer and T. A. Gill (2006). Investigation of Protein Denaturation and Pigment Fading in Farmed Steelhead (*Onchorhynchus mykiss*) Fillets During Frozen Storage. *J. Food Process. Pres*., **30**, 208–230.

Pallottino, F., P. Menesatti, C. Costa, G. Paglia, F.R. De Salvador and D. Lolletti (2010). Image Analysis Techniques for Automated Hazelnut Peeling Determination. *Food Bio. Tech*., **3**, 155-159.

Park, B., Y.-R. Chen, W. R. Hruschka, S. D. Shackelford and M. Koohmaraie (2001). Principal Component Regression of Near-Infrared Reflectance Spectra for Beef Tenderness Prediction. *T. ASAE*, **44**, 609-615.

Pathare, P.B., U.L. Opara and F.A.-J. Al-Said (2013). Color Measurement and Analysis in Fresh and Processed Foods: a Review. *Food Bio. Tech*., **6**, 36-60.

Pavlov, A. (2007). Changes in the Meat from Aquaculture Species During Storage at Low Temperature and Attempts for Differentiation Between Thawed-Frozen and Fresh Chilled Meat. A Review. *Bulg. J. Vet. Med*., **10**, 67-75.

Pereira, A. C., M. S. Reis and P. Saraiva (2009). Quality Control of Food Products Using Image Analysis and Multivariate Statistical Tool. *Ind. Eng. Chem. Res*., **48**, 988-998.

Pérez-Ramoz, J.D., W.P. Findlay, G. Peck and K.R. Morris (2005). Quantitative Analysis of Film Coating in a Pan Coater Based on In-Line Sensor Measurements. *AAPS Pharm Sci Tech.*, **6**, 127-136.

Poli, B.M., G. Parisi, G. Zampacavallo, M. Mecatti, P. Lupi, M. Gualtieri and O. Franci (2001). Quality Outline of European Sea Bass (Dicentrarchus labrax) Reared in Italy: Shelf Life, Edible Yield, Nutritional and Dietetic Traits. *Aquaculture*, **202**, 303-315.

Pomerantsev, A.L. and O.Y. Rodionova (2012). Process Analytical Technology: a Critical View of the Chemometricians. *J. Chemom*., **26**, 299-310.

Porikli, F. (2003). Inter-Camera Color Calibration by Correlation Model Function. Edited preprints of *ICIP - International Conference on Image Processing*, Barcelona (Catalonia, Spain), 14-17 September.

Prats-Montalbán, J.M. and A. Ferrer (2007). Integration of Colour and Textural Information in Multivariate Image Analysis: Defect Detection and Classification Issues. *J. Chemom*., **21**, 10-23.

Prats-Montalbán, J.M., A. de Juan and A. Ferrer (2011). Multivariate Image Analysis: a Review with Applications. *Chemom. Intell. Lab. Syst*., **107**, 1-23.

Pratt, W. K. (1991). *Digital Image Processing*. John Wiley & Sons, New York, (U.S.A.).

Rajalahti, T. and O.M. Kvalheim (2011). Multivariate Data Analysis in Pharmaceutics: a Tutorial Review. *Int. J. Pharm*., **417**, 280-290.

Rantanen, J., E. Rasanen, O. Antikainen, J.P. Mannermaa and J. Yliruusi (2001). In-line Moisture Measurement During Granulation with a Four-Wavelength Near Infrared Sensor: an Evaluation of Process-Related Variables and Development of a Non-Linear Calibration Model. *Chemom. Intell. Lab. Syst*., **56**, 51-58.

Rantanen, J., H. Wikstrom, R. Turner and L.S. Taylor (2005). Use of In-line Near-Infrared Spectroscopy in Combination with Chemometrics for Improved Understanding of Pharmaceutical Processes. *Anal. Chem*., **77**, 556-563.

Ravn, C., E. Skibsted and R. Bro (2008). Near-Infrared Chemical Imaging (NIR-CI) on Pharmaceutical Solid Dosage Forms – Comparing Common Calibration Approaches. *J. Pharm. Biomed. Anal.*, **48**, 554-561.

Reeves, V. and D. Bednar (1994). Defining Quality: Alternatives and Implications. *Acad. Manage. Rev.*, **19**, 419-445.

Reis, M. and A. Bauer (2009). Wavelet Texture Analysis of On-line Acquired Images for Paper Formation Assessment and Monitoring. *Chemom. Intell. Lab. Syst*., **95**, 129-137.

Rezzi, S., I. Giani, K. Héberger, D. K. Axelson, V. M. Moretti, F. Reniero and C. Guillou (2007). Classification of Gilthead Sea Bream (*Sparus aurata*) From ^{1}H NMR Lipid Profiling Combined with Principal Component and Linear Discriminant Analysis. *J. Agric. Food Chem*., **55**, 9963-9968.

Rodriguez-Saona, L.E. and M.E. Allendorf (2011). Use of FTIR for Rapid Authentication and Detection of Adulteration of Food. *Annu. Rev. Food Sci. Tech*., **2**, 467-483.

Roggo, Y, N. Jent, A. Edmond, P. Chalus and M. Ulmshneider. Characterizing Process Effects on Pharmaceutical Solid Forms Using Near-Infrared Spectroscopy and Infrared Imaging. *Eur J Pharm Biopharm*., **61**, 100-110.

Romero-Torres, S., J.D. Pérez-Ramoz, K.R. Morris and E.R. Grant (2006). Raman Spectroscopy for Tablet Coating Thickness Quantification and Coating Characterization in the Presence of Strong Fluorescent Interference. *J Pharmaceut Biomed*., **41**, 811-819.

Roth, B., A. Foss and A. K. Imsland (2009). Relationship Between Muscle pH and Flesh Color of Atlantic Halibut. *J. Food Sci*., **74**, 123-125.

Roth, B., M.D. Jenssen, T.M. Jonassen, A. Foss and A.K. Imsland (2007). Change in Flesh Quality Associated with Early Maturation of Atlantic Halibut (*Hippoglossus hippoglossus*). *Aquac. Res*., **38**, 757–763.

Roth, B., S.J.S. Johansen, J. Suontama, A. Kiessling, O. Leknes, B. Guldberg and S. Handeland (2005). Seasonal Variation in Flesh Quality, Comparison Between Large and Small Atlantic Salmon (*Salmo salar*) Transferred into Seawater as 0+ or 1+ Smolts. *Aquaculture*, **250**, 830-840.

Ruff, N., R.D. Fitzgerald, T.F. Cross and J.P. Kerry (2002). Comparative Composition and Shelf-Life of Fillets of Wild and Cultured Turbot (*Scophthalmus maximus*) and Atlantic Halibut (*Hippoglossus hippoglossus*). *Aquacult. Int*., **10**, 241-256.

Ruotsalainen, M., J. Heinämäki, H. Guo, N. Laitinen and J. Yliruusi (2003). A Novel Technique for Imaging Film Coating Defects in the Film-Core Interface and Surface of Coated Tablets. *Eur J Pharm Biopharm*., **56**, 381-388.

Russ, J.C. (1999). *The Image Processing Handbook*. Florida, CRC Press.

Sasic, S. (2007). An In-Depth Analysis of Raman and Near-Infrared Chemical Images of Common Pharmaceutical Tablets. *Appl Spectrosc*., **61**, 239-250.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem*., **36**, 1627-1639.

Schievano, E., G. Pasini, G. Cozzi and S. Mammi (2008). Identification of the Production Chain of Asiago d'Allevo Cheese by Nuclear Magnetic Resonance Spectroscopy and Principal Component Analysis. *J. Agric. Food Chem*., **56**, 7208-7214.

Schubring, R. (2010). Quality Assessment of Fish and Fishery Products by Color Measurement. In: *Sensory Analysis of Foods of Animal Origin* (L. M. L. Nollet and F. Toldrá Eds), CRC Press, Boca Raton, Florida, (U.S.A.).

Seber, G. A. F. (1984). *Multivariate Observations*, John Wiley & Sons, Hoboken, NJ (U.S.A.).

Seitavuopio, P., J. Heinämäki, J. Rantanen and J. Yliruusi (2006). Monitoring Tablet Surface Roughness During the Film Coating Process. *AAPS Pharm Sci Tech*., **7**, 31.

Sekulic, S.S., J. Wakeman, P. Doherty and P.A. Hailey (1998). Automated System for the On-line Monitoring of Powder Blending Processes Using Near-Infrared Spectroscopy. Part II. Qualitative Approaches to Blend Evaluation. *J. Pharm. Biomed. Anal*., **17**, 1285-1309.

Sharaf, M. A., D. L. Illman and B. R. Kowalski (1986). *Chemometrics*, John Wiley & Sons, New York (U.S.A.).

Shi, Z., R.P. Cogdill, S.M. Short and C.A. Anderson (2008). Process Charactertization of Powder Blending by Near-Infrared Spectroscopy: Blend End-point and Beyond. *J. Pharm. Biomed. Anal*., **47**, 738-745.

Sigal, L., S. Sclaroff, S. and V. Athitsos (2004). Skin Color-Based Video Segmentation Under Time-Varying Illumination. *IEEE T. Pattern Anal*., **26**, 862-877.

Simon, L. L., K. A. Oucherif, Z. K. Nagy and K. Hungerbuhler (2010). Bulk Video Imaging Based Multivariate Image Analysis, Process Control Chart and Ccoustic Signal Assisted Nucleation Detection. *Chem. Eng. Sci*., **65**, 4893-4995.

Simon, L. L., T. Merz, S. Dubuis, A. Lieb and K. Hungerbuhler (2012). In Situ Monitoring of Pharmaceutical and Specialty Chemicals Crystallization Processes Using Endoscopy-stroboscopy and Multivariate Image Analysis. *Chem. Eng. Res. Des*., **90**, 1847-1855.

Sivertsen, A. H., T. Kimiya and K. Heia (2011). Automatic Freshness Assessment of Cod (*Gadus morhua*) Fillets by VIS/NIR Spectroscopy. *J. Food Eng*., **103**, 317–323.

Skibsted, E.T.S., J.A. Westerhuis, A.K. Smilde and D.T. Witte (2007). Examples of NIR Based Real Time Release in Tablet Manufacturing. *J. Pharm. Biomed*., **43**, 1297-1305.

Sonka, M., V. Hlavac and R. Boyle (1999). *Image Processing, Analysis and Machine Vision*. PWS Publishing, CA (U.S.A.).

Svensson, O., T. Kourti and J. F. MacGregor (2002). An Investigation of Orthogonal Signal Correction Algorithms and Their Characteristics. *J. Chemom*., **16**, 176-188.

Tessier, J., C. Duchesne and G. Bartolacci (2007). A Machine Vision Approach to On-Line Estimation of Run-of-Mine Ore Composition on Conveyor Belts. *Miner. Eng*., **20**, 1129-1144.

Tessier, J., C. Duchesne, C. Gauthier and G. Dufour (2008). Estimation of Alumina Content of Anode Cover Material Using Multivariate Image Analysis Techniques. *Chem. Eng. Sci.*, **63**, 1370-1380.

Tomba, E. (2013). Latent Variable Modeling Approaches to Assist the Implementation of Quality-by-Design Paradigms in Pharmaceutical Development and Manufacturing. *PhD Thesis*, University of Padova (Italy).

Tomba, E., M. Barolo and S. García-Muñoz (2012). General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Ind. Eng. Chem. Res*., **51**, 12886-12900.

Tomba, E., P. Facco, F. Bezzo and M. Barolo (2013). Latent Variable Modeling to Assist the Implementation of Quality-by-Design Paradigms in Pharmaceutical Development and Manufacturing: a Review. *Int. J. Pharm*., **457**, 283-297.

Tomba, E., P. Facco, M. Roso, M. Modesti, F. Bezzo and M. Barolo (2010). Artificial Vision System for the Automatic Measurement of Interfiber Pore Characteristics and Fiber Diameter Distribution in Nanofiber Assemblies. *Ind. Eng. Chem. Res*., **49**, 2957-2968.

Trygg, J. and S. Wold (2002). Orthogonal Projections to Latent Structures (O-PLS). *J. Chemom*., **16**, 119-128.

Tu, K.-Y (2009). Analysis of Camera's Images Influenced by Varying Light Illumination for Design of Color Segmentation. *J. Inf. Sci. Eng*., **25**, 1885-1899.

Uddin, M. (2010). Differentiation of Fresh and Frozen-Thawed Fish. In: *Handbook of Seafood and Seafood Product Analysis* (M. Leo, L. M. L. Nollet, F. Toldrá, Eds.), CRC Press, Boca Raton (U.S.A.).

Uddin, M., E. Okazaki, S. Torza, Y. Yumiko, M. Tanaka and Y. Fukuda (2005). Non Distructive Visible/NIR Spectroscopy for Differentiation of Fresh and Frozen-Thawed Fish. *J. Food Sci*., **70**, 506-510.

Ulbricht, T. L. V. and D. A. T. Southgate (1991). Coronary Heart Disease: Seven Dietary Factors. *The Lancet*, **338**, 985-992.

Valle, S., W. Li and S.J. Qin (1999). Selection of the Number of Principal Components: the Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Ind. Eng. Chem. Res.*, **38**, 4389-4401.

Valous, N. A., F. Mendoza, D.-W. Sun and P. Allen (2009). Colour Calibration of a Laboratory Computer Vision System for Quality Evaluation of Pre-Sliced Hams. *Meat Sci*., **81**, 132-141.

van den Berg, F., C.B. Lyndgaard, K.M. Sørensen and S.B. Engelsen (2013). Process Analytical Technology in the Food Industry. *Trends Food Sci. Tech*., **31**, 27-35.

Van Mechelen, I. and A.K. Smilde (2010). A Generic Link-Mode Decomposition Model for Data Fusion. *Chemom. Intell. Lab. Syst*., **104**, 83-94.

Vidaček, S., H. Medića, K. Botka-Petrakb, J. Nežakc and T. Petrak (2008). Bioelectrical Impedance Analysis of Frozen Sea Bass (*Dicentrarchus labrax*). *J. Food Eng*., **88**, 263–271.

Vidal, A., P. Talens, J.P. Prats-Montalbán, S. Cubero, F. Albert and J. Blasco (2013). In-Line Estimation of the Standard Colour Index of Citrus Fruits Using a Computer Vision System Developed for a Mobile Platform. *Food Bio. Tech*., **6**, 3412-3419.

Vornheder, P.F. and W.J. Brabbs (1970). Moisture Determination by Near-Infrared Spectrometry. *Anal. Chem*., **42**, 1454-1456.

Wargo, D.J. and J.K. Drennen (1996). Near-Infrared Spectroscopic Characterization of Pharmaceutical Powder Blends. *J. Pharm. Biomed. Anal*., **14**, 1415-1423.

Waskewitz, P. (2007). Machine Vision in Manufacturing. In: *Handbook of Machine Vision* (Hornberg, A. Ed.), Wiley-Vch, Weinheim (Germany).

Waters, M.E. (1982). Chemical Composition and Frozen Storage Stability of Spot, *Leiostomus xanthurus*. *Mar. Fish. Rev*., **44**, 14–22.

Westerhuis, J.A., T. Kourti and J.F. MacGregor (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *J. Chemom*., **12**, 301-321.

Wilson, K.E. and E. Crossman (1997). The Influence of Tablet Shape and Pan Speed on Intra-Tablet Film Coating Uniformity. *Drug Dev. Ind. Pharm*., **23**, 1239-1243.

Wise, B.M. and N.B. Gallagher (1996). The Process Chemometrics Approach to Process Monitoring and Fault Detection. *J. Process Control*, **6**, 329-348.

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. In *Multivariate analysis*, Academic Press Limited, New York (U.S.A.).

Wold, S. (1978). Cross-Validatory Estimation of Number of Components in Factor and Principal Components Models. *Technometrics*, **20**, 397-405.

Wold, S. and M. Sjöström (1977). *SIMCA: a Method for Analyzing Chemical Data in Terms of Similarity and Analogy* In *Chemometrics: Theory and Applications*, ACS Symposium Series.

Wold, S., H. Martens and H. Wold (1983). The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. *Lecture Notes in Math.*, **973**, 286-293.

Xiccato, G., A. Trocino, F. Tulli and E. Tibaldi (2004). Prediction of Chemical Composition and Origin Identification of European Sea Bass (*Dicentrarchus labrax*) by Near Infrared Reflectance Spectroscopy (NIRS). *Food Chem.*, **86**, 275-281.

Yacoub, .F, J. Lautens, L. Lucisano and W. Banh (2011). Application of Quality by Design Principles to Legacy Drug Product. *J Pharm. Innov.*, **6**, 61-68.

Yacoub, F. and J.F. MacGregor (2004). Robust Processes Through Latent Variable Modeling and Optimization. *AIChE J.*, **57**, 1278-1287.

Yadav, B.K. and V.K. Jindal (2001). Monitoring Milling Quality of Rice by Image Analysis. *Comput. Electron. Agr.*, **33**, 19-33.

Yasuma, F., T. Mitsunaga, D. Iso and S.K. Nayar (2008). Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum. Technical Report, Department of Computer Science, Columbia University, CUCS-061-08 (http://www.cs.columbia.edu/CAVE/databases/multispectral/ last access: 07/27/2011).

Yu, H. and J. F. MacGregor (2003). Multivariate Image Analysis and Regression for Prediction of Coating Content and Distribution in the Production of Snack Foods. *Chemom. Intell. Lab. Syst.*, **67**, 125-144.

Yu, H., J. F. MacGregor, G. Haarsma and W. Bourg (2003). Digital Imaging for Online Monitoring and Control of Industrial Snack Food Processes. *Ind. Eng. Chem. Res.*, **42**, 3036-3044.

Yu, H. and J. Flores-Cerrillo (2013). Latent Variable Model Predictive Control for Trajectory Tracking in Batch Processes: Internal Model Control Interpretation and Design Methodology. *Ind. Eng. Chem. Res.*, **52**, 12437-12450.

Yu, H., and J. F. MacGregor (2004). Monitoring Flames in an Industrial Boiler Using Multivariate Image Analysis. *AIChE J.*, **50**, 1474-1483.

Zeaiter, M., J.M. Roger and V. Bellon-Maurel (2006). Dynamic Orthogonal Projection. A New Method to Maintain the On-Line Robustness of Multivariate Calibrations. Application to NIR-Based Monitoring of Wine Fermentation. *Chemom. Intell. Lab. Syst.*, **80**, 227-235.

Zeitler, J.A., P.F. Taday, D.A. Newnham, M. Pepper and T. Rades (2007). Terahertz Pulsed Spectroscopy and Imaging in the Pharmaceutical Setting – a Review. *J Pharm Pharmacol*., **59**, 209-233.

Zeitler, J.A., Y.C. Shen, C. Baker, P.F. Taday, M. Pepper and T. Rades (2006). Analysis of Coating Structures and Interfaces in Solid Oral Dosage Forms by Three Dimensional Terahertz Pulsed Imaging. *J Pharm Sci*., **96**, 330-340.

Zhang, L., M. Henson and S. Sekulic (2005). Multivariate Data Analysis for Raman Imaging of a Model Pharmaceutical Tablet. *Anal Chim Acta*., **545**, 262-278.

Zhang, X. and Y. Gao (2009). Face Recognition Across Pose: a Review. *Pattern Recogn*., **42**, 2876-2896.

Zhu, F., D. Zhang, Y. He, F. Liu and D.-W. Sun (2013). Application of Visible and Near Infrared Hyperspectral Imaging to Differentiate Between Fresh and Frozen–Thawed Fish Fillets. *Food Bio. Tech*., **6**, 2931-2937.

Ziémons, E., J. Mantanus, P. Lebrun, E. Rozet, B. Evrard and P. Hubert (2010). Acetaminophen determination in low-dose pharmaceutical syrup by NIR spectroscopy. *J. Pharm. Biomed. Anal*., **53**, 510-516.

Web sites

`www.asiagocheese.it` (last access 12/19/2011, in Italian)

`www.ifpma.org/global-health/counterfeits.html` (last access 01/02/2014)

Acknowledgements

There are many people I would like to thank for their direct or indirect contributions to this work and, more generally, to these three years.

Thanks to my supervisor, Prof. Massimiliano Barolo, and to Dr. Pierantonio Facco for their guidance, patience, and for all the helpful discussions that enrich this Dissertation.
Thanks to Dr. Salvador García-Muñoz for his support during my visit at Pfizer and for his priceless contribution to my professional growth.
Thanks to Dr. Luca Fasolato for introducing me to the food technology world and for the interesting collaborations.

Grazie a tutti gli amici del CAPE-Lab per le immancabili pause caffè, e grazie a chi del CAPE-Lab non fa parte per tutti gli scherzi, i kebab e, in generale, per tutte le risate che ci siamo fatti insieme.

Grazie a mamma e papà per il supporto incondizionato che hanno sempre mostrato verso le mie scelte.

Infine, grazie a Elena. Grazie per aver cambiato così tanto la mia vita in così poco tempo.