Some contributions to supervised classification of hyperspectral data

X. Hadoux

THÈSE

Pour obtenir le grade de Docteur

Délivré par l'**Université de Montpellier II**

Préparée au sein de l'école doctorale **I2S**
Et de l'unité de recherche **ITAP, Irstea**

Spécialité: **Informatique**

Présentée par **Xavier Hadoux**

Some Contributions to Supervised Classification of Hyperspectral Data

Soutenue le 12 Novembre 2014 devant le jury composé de

Pr. Tom Fearn	University College London	Rapporteur
Pr. Grégoire Mercier	Telecom Bretagne	Rapporteur
Pr. Jocelyn Chanussot	INP Grenoble	Examinateur
Dr. Aoife Gowen	University College Dublin	Examinatrice
Dr. Olivier Strauss	Université Montpellier 2	Examinateur
Dr. Jean-Michel Roger	Irstea, UMR ITAP	Examinateur
Dr. Nathalie Gorretta	Irstea, UMR ITAP	Encadrante
Dr. Gilles Rabatel	Irstea, UMR ITAP	Directeur

“Be as you wish to seem.”

Aristote

Acknowledgements

I would first like to thank the UMR ITAP from Irstea and its Director, Tewfik Sari, for providing such an amazing working environment for research. This PhD thesis was directed by Gilles Rabatel, co-directed by Olivier Strauss and supervised on a day-to-day basis by Nathalie Gorretta, I really want to thank all of you to have trusted and supported me in order to achieve this PhD.

I also want to thank the other members of my PhD committee, Jean-Michel Roger, Ryad Bendoula, Bruno Tisseyre and Tewfik Sari, who, thanks to their knowledge, discussion and understanding, helped me in formalising and orientating this thesis work. The working environment would not have been the same without the team spirit that we had in the COMIC team, wonderfully managed by Alexia Gobrecht.

I would like to thank Tom Fearn (University College London) and Grégoire Mercier (Telecom Bretagne) to have reviewed and evaluated this PhD thesis, Aoife Gowen (University College Dublin) and Jean-Michel Roger (Irstea, UMR ITAP) to have examined this thesis and Jocelyn Chanussot to have reviewed the thesis as well as presided the jury.

I particularly want to thank Jean-Michel Roger, who, despite his high level of responsibility, always left his door open and tried to answer all my chemometric questions.

Ryad Bendoula, merci à toi d'avoir essayé de m'expliquer l'optique... il y a encore du boulot! Et évidemment merci aussi pour ces innombrables discussions qui m'ont fait "grandir" comme tu le dis si bien.

Daniel Moura, finalement on aura pris plus de cafés que travailler ensemble, mais cela aurait pu être bien quand même... donc merci à toi aussi!

Sylvain Jay, it was not that bad to have to share the same office... I won't thank you for your Matlab skills because it might be a bit hard, but for the rest, it was a real pleasure to work with you and I hope to keep doing so!

I also want to thank Veronique Bellon-Morel, scientific director at Irstea, for the organisation of the international conference of spectroscopy NIR 2013. Thanks to you, I had the opportunity at a very early stage to have a view from the inside of the scientific world!

I want to thank again Olivier Strauss who offered me the opportunity to teach for three different subjects at a Master level during this PhD. This was a very valuable experience for me, and I am really grateful you chose and trusted me to do this job.

Other than the scientific aspect, which only accounts for a part of the three years of a PhD, the administrative parts are also of major importance (and can often be even more complex...). Without these very friendly, efficient and patient persons, I would certainly not be here today. Of course, I think of Michèle Egea, Élodie Borg, Danielle Hoarau, Isabelle Paysan and Virginie Tuset for Irstea and Nicolas Serrurier for the Doctoral School of Montpellier University.

These three years would certainly not have been as funny or interesting without these colleagues/friends: Bastien Nguyen, Sarah Bellini, Benoit Moulin, Rebecca Kabore, Ana Herrero, Faten Ammari, Constance Demestihas, Léa Causse, Anne Montalban, Marjorie Davin, Priscillia Lamounier, Sylvain Jay, and of course the patatas called Silvia Mas!

I want to thank everyone who helped me in making this document more understandable through their suggestions and comments. In particular, I want to thank my friend Heather Wardlaw who spent her time proofreading this document.

For my research collaboration with the Royal Melbourne Institute of Technology (RMIT), Australia, I would like to thank the Department of International Relations (DRI) of Irstea, that trusted and gave me the chance to realise this international collaboration. I want to thank Dinesh Kant Kumar, head of the Biosignal Laboratory at RMIT to have been so welcoming and interested in my project. Thanks to this collaboration I also had the fortune of meeting Marc Sarossy, Anne Weymouth and Jonathan Crowston. I am looking forward to establishing further research opportunities with them.

I of course want to thank all my family and friends for being so supportive and loving, without them nothing would have been possible in the first place!

I finally want to thank my love Rachael, who supported and helped me everyday from the beginning of this thesis.

Thank you!

Contents

List of Figures

List of Tables

Notations

a, A	Scalar
$\mathbf{a}$	Column vector
$\mathbf{a}^\top$	Row vector
$\mathbf{A}$	Matrix of size $n \times m$
$\mathbf{I}_N$	Identity matrix of $\mathbb{R}^N$
$\mathbf{1}_N$	Column vector of N ones
$\mathbf{E}_k\big(\mathbf{A}\big)$	Matrix $\big(m \times k\big)$ containing the k eigenvectors of $\mathbf{A}$ associated to its k largest eigenvalues
$\mathcal{R}\big(\mathbf{A}\big)$	Range of $\mathbf{A}$: Subspace spanned by the columns of $\mathbf{A}$
$P_{\mathbf{P}}(\mathbf{A})$	Projection of the column vectors of $\mathbf{A}$ onto a subspace spanned by the matrix $\mathbf{P}$. In $\mathbb{R}^n$, $P_{\mathbf{P}}(\mathbf{A}) = \mathbf{P}\big(\mathbf{P}^\top\mathbf{P}\big)^{-1}\mathbf{P}^\top\mathbf{A}$. In $\mathbb{R}^m$, $P_{\mathbf{P}}(\mathbf{A}) = \mathbf{A}\mathbf{P}\big(\mathbf{P}^\top\mathbf{P}\big)^{-1}\mathbf{P}^\top$
$P_{\mathbf{P}}^{\perp}(\mathbf{A})$	Projection of the column vectors of $\mathbf{A}$ orthogonally to a subspace spanned by the matrix $\mathbf{P}$. In $\mathbb{R}^n$, $P_{\mathbf{P}}^{\perp}(\mathbf{A}) = \big(\mathbf{I}_n - \mathbf{P}\big(\mathbf{P}^\top\mathbf{P}\big)^{-1}\mathbf{P}^\top\big)\mathbf{A}$. In $\mathbb{R}^m$, $P_{\mathbf{P}}^{\perp}(\mathbf{X}) = \mathbf{X}\big(\mathbf{I}_m - \mathbf{P}\big(\mathbf{P}^\top\mathbf{P}\big)^{-1}\mathbf{P}^\top\big)$.
$P_{\mathbf{P},k}(\mathbf{A})$	Projection of $\mathbf{A}$ onto the subspace spanned by the k main directions of $\mathcal{R}\big(\mathbf{P}\big)$.
$P_{\mathbf{P},k}^{\perp}(\mathbf{A})$	Projection of $\mathbf{A}$ orthogonal to the subspace spanned by the k main directions of $\mathcal{R}\big(\mathbf{P}\big)$.

N	Number of observations in the training set

N_i	Number of observations in the i^{th} class in the training set
P	Number of variables (wavelengths)
C	Number of classes
$\mathbf{X}$	Matrix of size $\left(N \times P\right)$ containing the training set observations $\mathbf{X}^\intercal = \left[\mathbf{x}_1, \cdots, \mathbf{x}_N\right]$
$\mathbf{Y}$	Matrix of size $\left(N \times C\right)$ coding the class membership of the observations $\mathbf{Y}^\intercal = \left[\mathbf{y}_1, \cdots, \mathbf{y}_N\right]$ for example, $\mathbf{y} = \left[0, 0, 1, 0\right]$ codes for the third class among four
$\mathcal{E} = \mathcal{R}\left(\mathbf{X}\right)$	Individual space ($\mathcal{E} \subseteq \mathbb{R}^N$)
$\mathcal{F} = \mathcal{R}\left(\mathbf{X}^\intercal\right)$	Feature (Variable) space ($\mathcal{F} \subseteq \mathbb{R}^P$)
$\mathbf{D}$	Dimension reduction loading matrix $\left(P \times Q\right)$ with $Q \leq P$
$\mathbf{S}$	Score matrix matrix $\left(N \times Q\right)$ containing the N realizations of $Q \leq P$ variable vectors such that $\mathbf{S} = \mathbf{XD}$

$\mathbf{T}\left(\mathbf{X}\right)$	Total scatter of $\mathbf{X}$; $\mathbf{T}\left(\mathbf{X}\right) = \mathbf{X}^\intercal\mathbf{X}$
$\mathbf{B}\left(\mathbf{X}, \mathbf{Y}\right)$	Between-class scatter $\mathbf{B}\left(\mathbf{X}, \mathbf{Y}\right) = \mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}$
$\mathbf{W}\left(\mathbf{X}, \mathbf{Y}\right)$	Within-class scatter $\mathbf{W}\left(\mathbf{X}, \mathbf{Y}\right) = \mathbf{X}^\intercal\mathbf{X} - \mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}$ When there is no risk of confusion, the symbols $\mathbf{T}$, $\mathbf{B}$ and $\mathbf{W}$ are used without arguments.

Introduction

Hyperspectral imaging devices can record images with a very detailed spectral information for each pixel. The spectral information provided by these sensors been related to the biochemical properties of the measured sample, they have been used extensively for non-destructive measurement in the scientific and industrial fields for the last decades. At Irstea, and especially in the research unit ITAP, this spatialized spectral information allows to increase the characterization possibilities for environment and agrosystems already offered by spectrometers and classical color cameras. Indeed, while the spectral dimension provides a detail source of information regarding crop state (e.g., physiological, pathological), the spatial information helps to retrieve structural information (development stage, presence of weed, ...). The implementation of HS technology and the processing of the obtained data are however complex and thus require adapted procedures.

In the framework of classification, the biochemical differences of spectral pixels can be exploited to create a classification model that can assign each pixel of the HS image to a unique label. For supervised classification, training samples of known labels are required to define the assignment rule.

At the beginning of this thesis, the specific context of weed discrimination within wheat crops was investigated. In particular, a comparison of different spectral pretreatments with respect to classification methods was proposed in the context of in-field proximal detection. However, it appeared that some of the issues encountered could be addressed in a more generic way.

As a result, in the second part of this thesis, some more general issues regarding supervised classification of hyperspectral image were studied and constitute the bulk of the present document. For instance, three main contributions are developed in the following of this thesis. The first one is a new supervised dimension reduction method that can deal with the high dimensionality and collinearity of spectral data. The second one is a spectral-spatial approach that uses a spatial regularization method in combination with a supervised dimension reduction in order to optimize its effect on classification performances. The third and last one is an automatic method that allows radiance hyperspectral images to be classified even with varying lighting conditions and thus avoids the reflectance correction.

Notice that other methods have also been investigated, that we have chosen not to include in the present document. In particular, a collaboration with Pr. Dinesh Kant Kumar and Dr. Marc Sarossy at the Royal Melbourne Institute of Technology of Melbourne, Australia, was accomplished in order to develop a multi-resolution approach for spectral analysis. All these contributions can be found in Annex A.

This PhD manuscript is organized in five chapters. The first chapter introduces the background of supervised classification of hyperspectral images. The most important descriptions and definitions related to hyperspectral images are given. Then, the specific background regarding supervised classification is detailed. At the end, the main issues hyperspectral image classification methods have to face are summarized, i.e., the high dimensionality and collinearity of spectral data, the way of introducing spatial information in the classification process and the main correction stages to obtain a reflectance hyperspectral image. The second chapter gives a state-of-the-art of the main methods that tackle the previously mentioned issues. It finally allows to statuate on the drawbacks in existing approaches. In the third chapter, we propose three original contributions to tackle these issues. The fourth chapter is dedicated to the validation of the proposed approaches by presenting some results with real examples. Finally, the fifth chapter concludes this thesis by synthesizing the most important points of the developed approaches and proposes some perspectives and future research directions to continue this work.

Chapter 1

Introduction to hyperspectral image classification

Contents

This introductory chapter focuses on generic issues associated with classification of objects using hyperspectral imagery. We first give some background on hyperspectral imagery and explain the type of information it can bring in a general context. We then give main vocabulary and definition of classification, and detail some popular classification methods. The main issues associated with classification of hyperspectral images are finally given, i.e., the high-dimensionality and collinearity of spectral data, the use of spatial information in the classification process and the image reflectance calibration.

1.1 Hyperspectral imaging

1.1.1 Light-matter interaction

The interaction between electromagnetic (EM) waves and matter has for a long time been used to retrieve information about objects. Depending on the wavelength, different information can be retrieved. For illustration, Figure 1.1 represents the electromagnetic spectrum classified by range of wavelengths.

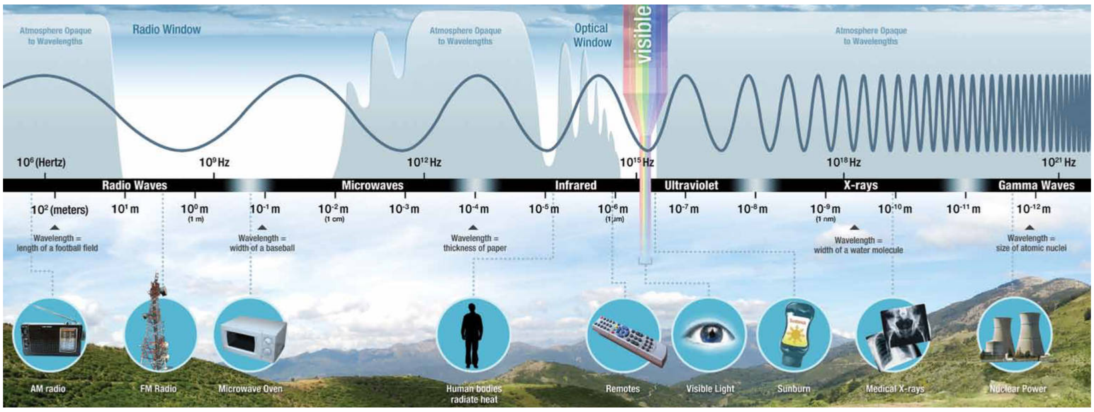


Figure 1.1: Electromagnetic spectrum [NASA, 2010].

Most ranges of the EM spectrum are now widely used in everyday life. For example, X-ray waves due to their high penetration depth in objects are now used daily for medical diagnoses and in airport security to check luggage contents. Ultraviolet (UV) waves are used to reveal fake notes due to fluorescence that re-emits light in the visible domain. Human eyes use the response of objects in the visible domain to perceive different colors. Near-infrared (NIR) waves are used to distinguish between vegetation and soil in agricultural applications [Brown and Noble, 2005], measure fat content in food [Osborne et al., 1984], inspect fruit quality [Nicolaï et al., 2007] and to measure the amount of oxygenated blood in retinal vessels [Schweitzer et al., 1999], etc. Far-infrared or thermal-infrared is used for atmosphere monitoring [Clerbaux et al., 2009] and for target detection in defense applications [Schwartz et al., 1996]. Finally, because the atmosphere is very transparent in this domain, micro and radio waves are used for telecommunication and radar for target detection (Figure 1.1).

For our environmental and agricultural applications, we mostly focus on the visible and near-infrared (VNIR) part of the EM spectra that ranges from 400 nm to 2500 nm. Note that at these wavelengths, the term *light* is usually employed instead of EM wave. The VNIR spectral region is particularly interesting because most organic constituents have specific absorption bands [Williams and Norris, 2001]. For instance, Vigneau [2010] measured the nitrogen content in wheat leaves and Gorretta et al. [2006] detected defaults of wheat kernels using a VNIR hyperspectral (HS) camera. For similar reasons, the VNIR region is also widely used in the food industry [Nicolaï et al., 2007] and medicine [Lu and Fei, 2014].

When light interacts with an object, three different interaction modes are usually described:
Transmission is when light passes through the object. The change in direction that occurs at the interface is governed by Fresnel equations and depends on the optical index of the object's material.
Reflection which can be either specular or diffuse depending on the nature of the interface. *Diffuse reflection* is when the incident light is reflected in all directions when it goes through a media composed of fine particles or when it is reflected on a rough surface. *Specular reflection* corresponds to the reflection of the incident light in a unique direction. The direction is governed by Descartes Law: that is the angle of incidence equals the angle of reflection with respect to the normal of the surface.
Absorption is when the object absorbs a part of the received energy. The absorption corresponds to molecular vibration, rotation, twisting and bending that involves specific energy levels (and thus wavelengths), which are characteristic of object chemical composition.

The proportion of each interaction mode depends on many parameters such as particle sizes, surface state, presence of absorbing compounds in the studied object.

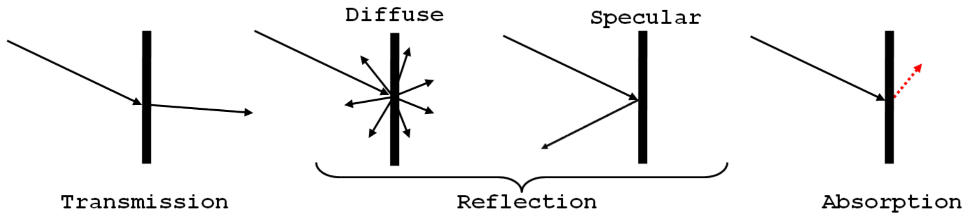


FIGURE 1.2: The three interaction modes between EM wave and matter

The fundamental uses of light-matter interaction in spectroscopy is through absorption. Indeed, from the Beer-Lambert Law, we know that absorption is related to the concentration of absorbing compounds. The Beer-Lambert Law expresses the absorption of a material at the wavelength λ as:

$$A(\lambda) = -\log \frac{I(\lambda)}{I_0(\lambda)} = \epsilon(\lambda).l.C \tag{1.1}$$

where the ratio $I_0(\lambda)$ and $I(\lambda)$ are respectively the incident and transmitted beam intensities as a function of the wavelength λ. The attenuation coefficient $\epsilon(\lambda)$ is an intrinsic property of the object, the path length l is the object's thickness, and C is the chemical concentration of absorbing species. This relation is very useful in analytical chemistry since by only knowing l and $\epsilon(\lambda)$, the concentration can be retrieved by measuring light intensity $I(\lambda)$ that goes through the material.

When there is no possible way of measuring through the object, the surface diffuse reflectance can also be used to retrieve the chemical concentration of absorbing species, but the relation is more complex as described in [Dahm and Dahm, 2001]. The reflectance or reflectivity of the surface is thus characteristic of the observed objects. Therefore, measuring the reflectivity of the object's surface in function of the wavelength defines a reflectance spectrum, which is often referred to as the object's *spectral signature*. For example, in Figure 1.3 are represented the reflectance spectra of four different objects. Note that it is these differences in spectral reflectance that are used to classify objects through spectrometry or hyperspectral imaging.

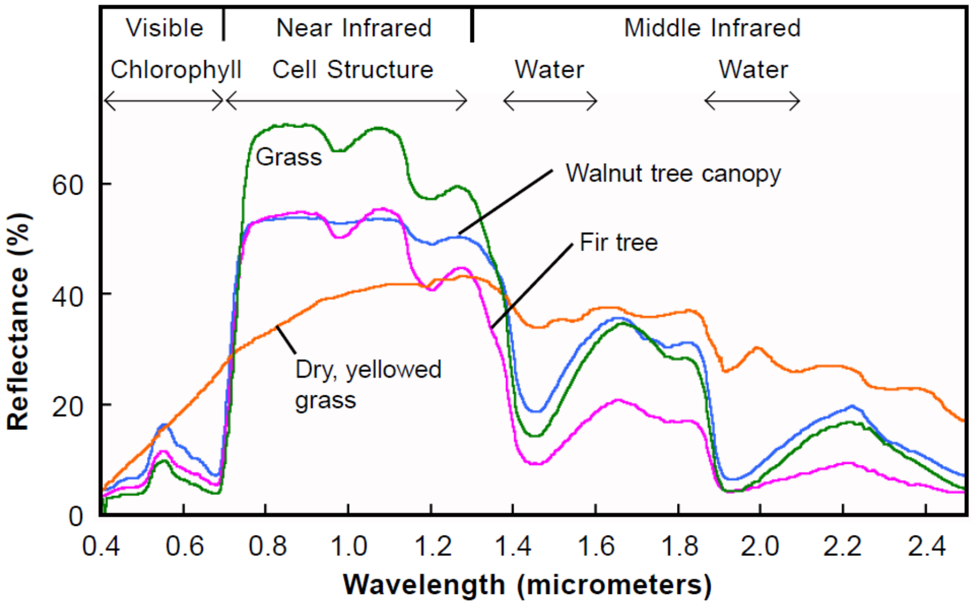


Figure 1.3: Four vegetation reflectance spectra plotted as a function of the wavelength in the VNIR domain.[Smith 2001 Microimage]

1.1.2 Definitions

Hyperspectral imaging, also known as chemical imaging and imaging spectroscopy, is a relatively recent imaging technology that enables the acquisition of both spectral and spatial information of targeted objects. Hyperspectral (HS) images are multivariate images than can be represented as data-cubes with two spatial dimensions (x, y) and one spectral dimension (λ) (Figure 1.4). Each spectral pixel in the resulting image contains a sampled spectral measurement of radiance, which can be interpreted to identify the material presents in the scene.

This representation is usually seen in two equivalent ways:

- From a spectrometric point of view, the HS image content is seen as *spatialized* spectral information: spectrometers are spatially resolved.
- From the image processing point of view, the HS image content is seen as *spectralized* spatial information: image pixels are spectrally resolved.

Either way, each spatial position in the HS image is associated to a spectrum that contains chemical information of the imaged object.

Note that originally HS imaging was developed for large scale remote sensing of environment using satellite as an improvement of multi-spectral imaging [Goetz

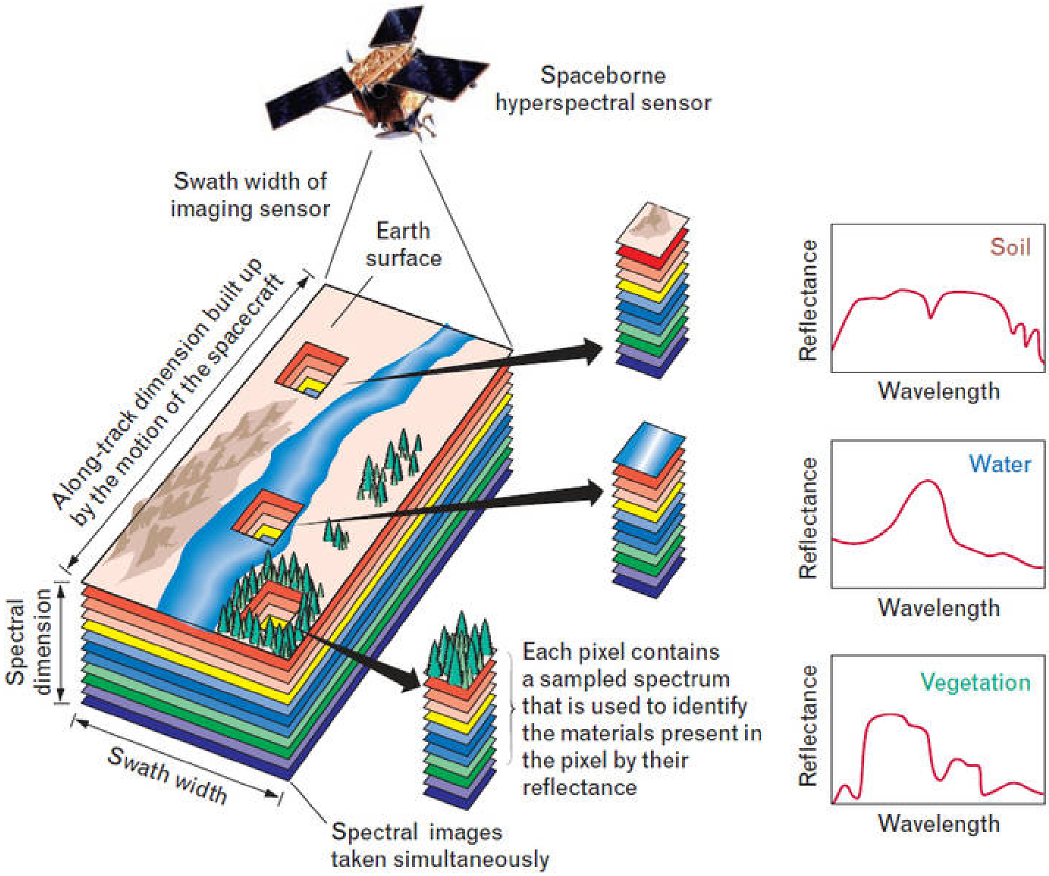

Spaceborne hyperspectral sensor
Swath width of imaging sensor
Earth surface
Along-track dimension built up by the motion of the spacecraft
Spectral dimension
Swath width
Spectral images taken simultaneously
Each pixel contains a sampled spectrum that is used to identify the materials present in the pixel by their reflectance
Reflectance
Soil
Wavelength
Reflectance
Water
Wavelength
Reflectance
Vegetation
Wavelength

Figure 1.4: Hyperspectral image concept. Multi-variate image with simultaneous access to spectral wavebands over a large area in a ground-based scene. The graphs in the figure illustrate the spectral variation in reflectance for soil, water, and vegetation. (from Shaw and Burke [2003])

et al., 1985]. Therefore, some definitions of hyperspectral imaging due to [Kruse, 2000] and [Chang, 2007] generalized this concept of multi-spectral imaging: Multi-spectral devices can record bands of different spectral widths that can be irregularly distributed. HS imaging devices record contiguous, regularly distributed and narrow spectral bands, which leads to an almost continuous spectral measurement for each pixel.

Grahn and Geladi [2007] similarly defined a HS image as a type of multivariate image that has these two properties:

- many wavelengths or other variable bands, often more than 100;
- the possibility to express a pixel as a spectrum with spectral interpretation, spectral transformation, spectral data analysis, etc

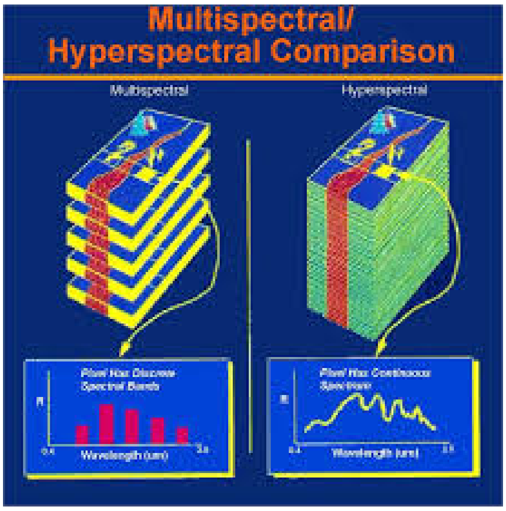


Figure 1.5: From multi-spectral images to hyperspectral images. (from http://rst.gsfc.nasa.gov/)

Gowen et al. [2007] summarized the possibilities offered by HS images compared with classical vision techniques and spectroscopy measures. This summary is presented in Table 1.1.

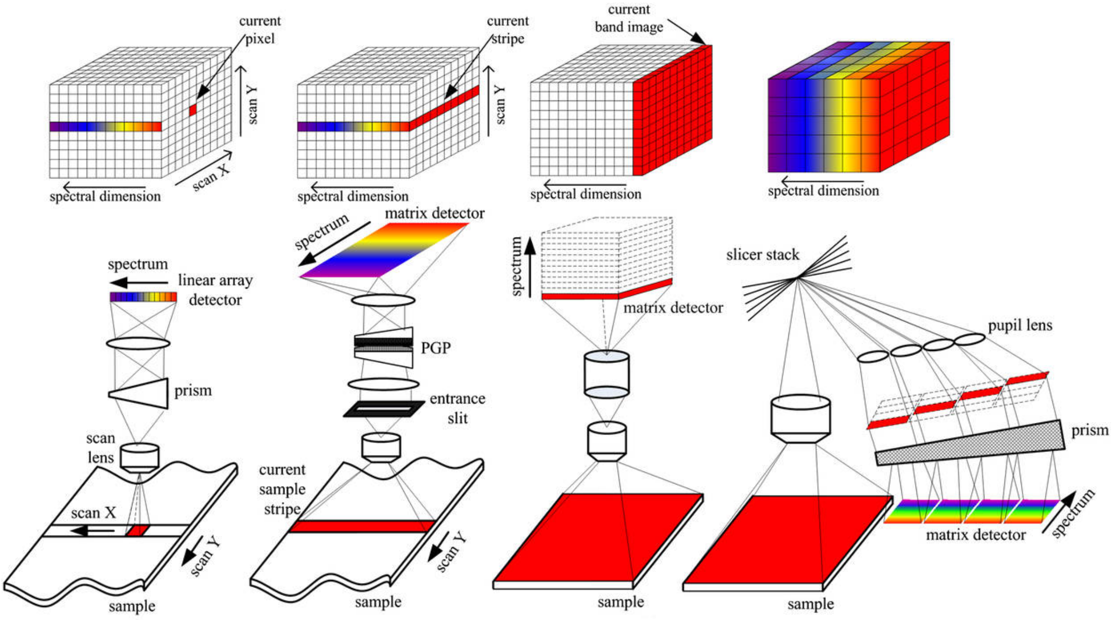
Table 1.1: Comparison of RGB imaging, NIR spectroscopy (NIRS), multi-spectral imaging (MSI) and hyperspectral imaging (HSI). Yes (Y), Limited (L) and No (N). (Modified from [Gowen et al., 2007])

Feature	RGB	NIRS	MSI	HSI
Spatial information	Y	N	Y	Y
Spectral information	N	Y	L	Y
Multi-constituent information	L	Y	L	Y
Sensitivity to minor components	N	N	L	Y

1.1.3 Acquisition

There are four main HS image acquisition techniques that differ on the way to fill the data-cube. These four techniques, i.e., point-scanning, line-scanning, spectral-scanning and non-scanning are schematically represented in Figure 1.6. Each has advantages and drawbacks that are context and application dependent.

Point-scanning or whisk-broom is a 2-dimensional spatial scanning technique that uses a spectrophotometer. With this technique, all wavelengths are acquired simultaneously but for only one pixel at a time. This measure is usually accurate in terms of spectral resolution but often less precise and slower in the spatial directions because of moving parts involved in the scanning.



(A) Point-scanning (B) Line-scanning (C) Spectral-scanning (D) Non-scanning

FIGURE 1.6: Illustration of the four main technologies for hyperspectral image acquisition.(from [Li et al., 2013b])

Line-scanning or push-broom corresponds to a 1-dimensional spatial scanning technique that uses a 2-dimensional sensor. This sensor records one spatial and one spectral dimension at a time. The acquisition of other lines is performed either by moving the sensor over the objects (plane, drone or satellite in remote sensing) or by moving the objects (conveyor belt, translation stage in laboratory or industries).

Spectral-scanning, also called staring or area imaging corresponds to the acquisition of several 2-dimensional images at different wavelengths. Mostly used in laboratories, the stationary object is spectrally scanned by exchanging one filter after another. Wavelength scanning can be made by using either a Fabry–Pérot interferometer or a tunable filter (Acousto Optical Tunable Filter or Liquid Crystal Tunable Filter) [Gat, 2000]. The main advantages of this technique are to be able to choose only the spectral bands of interest and to have a potentially high image resolution. However, if the object is not perfectly still during the acquisition, image channel misregistration creates spectral smearing.

Non-scanning, snapshot or one-shot imaging technique records all dimensions of the HS data cube simultaneously [Hagen et al., 2012, Hagen and Kudenov, 2013]. The data cube is acquired using the perspective projection of the data cube and reconstructed without any moving part involved in the process. The acquisition time is thus largely reduced, but the image resolution is usually worse than with the other approaches.

1.2 Supervised classification

In this section, we present the general context of classification, with its most important definitions and hypotheses. Some of the most popular classification approaches are also given. This thesis focuses only on supervised classification.

1.2.1 Definitions and hypotheses

The objective of classification is to identify the nature of objects in terms of classes based on some characteristics or features. In supervised classification, all classes are assumed to be known and mutually exclusive. Some observations for each class are also supposed to be available to train a model. These observations that form the so-called *training samples* are manually attributed, which necessitate the prior establishment of a ground truth (GT).

With a HS image, features can potentially take multiple forms, e.g., raw spectra, reduced spectral variables, object shapes, textures, or some combinations of these. Let us define a feature space $\mathcal{X} \in \mathbb{R}^P$ and a finite set of all possible classes $\mathcal{Y} = \{\mathcal{Y}_1, \cdots, \mathcal{Y}_C\}$, where $\mathcal{Y}_c$ denotes one of the C classes. The N observations of the training set are gathered in a feature matrix $\mathbf{X} = \{\mathbf{x}_i \in \mathcal{X}\}$ and its associated class or label matrix $\mathbf{Y} = \{\mathbf{y}_i \in \mathcal{Y}\}$, where $i = 1, \cdots, N$. With this notation, $\mathbf{x}_i$ corresponds to the i^{th} feature vector and $\mathbf{y}_i$ to its associated class vector. The class vector is conveniently coded in 'dummy' or disjunctive fashion, e.g. , $\mathbf{y} = [0\ 0\ 1\ 0\ 0]^\intercal$ codes class 3 among 5.

Classification consists in assigning each feature vector to one of the C classes of interest using a function $g : \mathcal{X} \mapsto \mathcal{Y}$. The assignment is hoped to be made as accurately as possible using only the available data in the training set. The objective of classification is to generalize well so that any unseen feature vector $\mathbf{x}$ is also well classified. This corresponds to maximizing the posterior probability which is the probability to obtain a class given a feature vector $\Pr(\mathcal{Y}_c \mid \mathbf{x})$. Using the Bayes Theorem, the posterior class probability can be computed using the class conditional density $Pr(\mathbf{x} \mid \mathcal{Y}_c)$, the prior class probability $Pr(\mathcal{Y}_c)$ and $Pr(\mathbf{x})$ as:

$$Pr(\mathcal{Y}_c \mid \mathbf{x}) = \frac{Pr(\mathbf{x} \mid \mathcal{Y}_c)Pr(\mathcal{Y}_c)}{Pr(\mathbf{x})} \tag{1.2}$$

Note that the class conditional density taken as a function of $\mathcal{Y}_c$ is also called likelihood function and is often noted $f_c(\mathbf{x})$. Similarly, the prior probability of being in class c is also often noted π_c. The denominator $Pr(\mathbf{x}) = \sum_c Pr(\mathbf{x} \mid \mathcal{Y}_c)Pr(\mathcal{Y}_c)$ is the same for every class and is thus usually left apart for further computations. We thus have the classical relation:

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior} \tag{1.3}$$

The function g^* that maximizes the posterior class probability:

$$g^*(\mathbf{x}) = \arg \max_c Pr(\mathcal{Y}_c \mid \mathbf{x}) \tag{1.4}$$

is called *Bayes classifier* and is the best classifier over all measurable functions g for the *zero-one loss function* [Fan et al., 2011]. This loss function given by

$$L\big(\mathbf{y}, g(\mathbf{x})\big) = \begin{cases} 0, & g(\mathbf{x}) = \mathbf{y} \\ 1, & g(\mathbf{x}) \neq \mathbf{y} \end{cases} \tag{1.5}$$

is commonly used in classification since it corresponds to the computation of the misclassification rate.

The risk of misclassification is defined as:

$$\textit{risk}(g) = \mathbb{E}\Big[L\big(\mathbf{y}, g(\mathbf{x})\big)\Big] \tag{1.6}$$

where $\mathbb{E}$ is the mathematical expectation for every $\mathbf{x}$. Because classes are assumed to be mutually exclusive, this risk is decomposed as:

$$\textit{risk}(g) = \mathbb{E}\left[\sum_{c=1}^{C} L\big(\mathbf{y}, g(\mathbf{x})\big) Pr(\mathcal{Y}_c \mid \mathbf{x})\right] \tag{1.7}$$

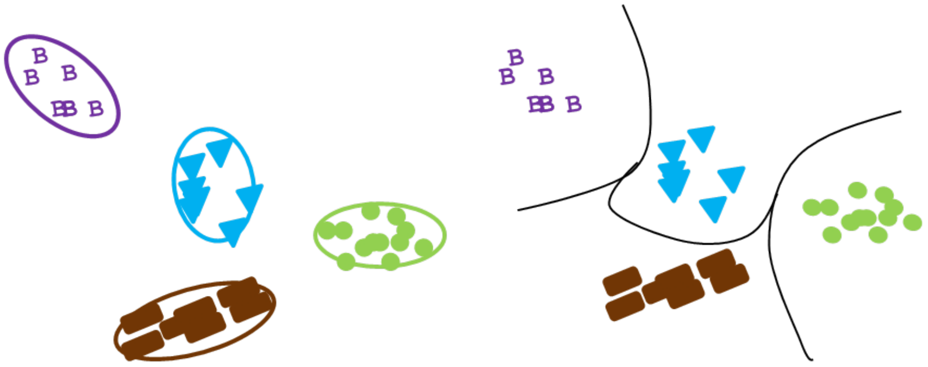
Using the zero-one loss function, the expected risk becomes:

$$\textit{risk}(g) = 1 - \mathbb{E}\big(Pr(\mathcal{Y}_c \mid \mathbf{x})\big) \tag{1.8}$$

which means that the minimum risk is obtained for the Bayes classifier and is thus called *Bayes risk*. When g^* can be computed, $risk(g^*)$ is used as a benchmark for other classifiers .

The classification procedure is usually done in two successive stages [Bishop, 2007]. (1) *The inference stage* consists in learning the posterior class probability using the available training samples. It can be done in two different ways which are used to categorize classification methods [Bishop et al., 2007]. *Generative classifiers* learn both the class likelihood and the class prior probability and use Bayes Theorem to retrieve the posterior probability whereas *discriminative classifiers* directly learn the class posterior probability.
(2) *The decision stage* uses this posterior class probability to make the optimal class assignment.

Classification methods differ on the choice for the inference model: linear *vs.* non-linear, parametric *vs.* non-parametric and on the decision rule.



(A) Generative classifiers model each class probability density function

(B) Discriminative classifiers directly model the posterior class probability.

FIGURE 1.7: Conceptual visualisation of the two main classification approaches.

1.2.2 Generative classifiers

Generative classifiers model the posterior class probability using Bayes rule; that is by modeling the likelihood function by making the assumption of the distribution for each class and estimating the class prior probability. The prior is usually estimated by either the empirical proportion $\widehat{\pi}_c = N_c/N$, where N_c is the number of training sample of class c, or, if no assumption on the proportion is preferred, given by $\widehat{\pi}_c = 1/C$.

Depending on the classifier, different assumptions are made to estimate the likelihood function.

The Naive Bayes Classifier makes a strong conditional assumption of independence of $\mathbf{x}$ variables for the estimation of the likelihood function.

$$Pr(\mathbf{x} \mid \mathcal{Y}_c) = Pr(x_1, \cdots, x_P \mid \mathcal{Y}_c) = \prod_{i=1}^{P} Pr(x_i \mid \mathcal{Y}_c) \tag{1.9}$$

This is a very strong assumption, especially with spectral data that are smooth functions of the wavelength for which the variables are clearly not independent of one another. It however allows a drastic simplification of the complexity and can potentially be used with transformed variables. Using the independance assumption, the classification function is simply given by:

$$g(\mathbf{x}) = \arg\max_c \left(\pi_c \prod_{i=1}^{P} Pr(x_i \mid \mathcal{Y}_c) \right) \tag{1.10}$$

Linear Discriminant Analysis (LDA) makes the assumption that class densities are multivariate Gaussian with class mean vectors $\boldsymbol{\mu}_c = \mathbb{E}\big[\mathbf{x} \mid \mathcal{Y}_c\big] = \Big(\mathbb{E}\big[\mathbf{x}_1 \mid \mathcal{Y}_c\big], \cdots, \mathbb{E}\big[\mathbf{x}_P \mid \mathcal{Y}_c\big]\Big)^\intercal$ and equal covariance matrices $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_c = \mathbb{E}\big[\mathbf{x} - \boldsymbol{\mu}_c \mid \mathcal{Y}_c\big]\mathbb{E}\big[\mathbf{x} - \boldsymbol{\mu}_c \mid \mathcal{Y}_c\big]^\intercal$ for every class $c \in \mathcal{Y}$. For an observation $\mathbf{x}$ that belongs to the class c, its likelihood function is given by:

$$f_c(\mathbf{x}) = (2\pi)^{-P/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\big) \tag{1.11}$$

Because it is usually numerically more stable and does not change the outcome, the negative of the log-likelihood:

$$\ell_c(\mathbf{x}) = -2\log\big(f_c(\mathbf{x})\big) = P \cdot \log(2\pi) + \log\big(|\boldsymbol{\Sigma}|\big) + (\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) \tag{1.12}$$

is usually minimized. The two first terms on the right hand side are constant with respect to the class c. Therefore, maximizing the likelihood corresponds to minimizing the Mahalanobis distance [De Maesschalck et al., 2000], which is given by:

$$d_M(\mathbf{x}, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)} \tag{1.13}$$

Note that when the covariance matrix is the identity matrix $\boldsymbol{\Sigma} = \mathbf{I}$, the Mahalanobis distance is equivalent to the Euclidean distance between $\mathbf{x}$ and $\boldsymbol{\mu}_c$:

$$d_E(\mathbf{x}, \boldsymbol{\mu}_c) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_c)^\intercal (\mathbf{x} - \boldsymbol{\mu}_c)} \tag{1.14}$$

It can be shown by expanding the term $(\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)$ that LDA defines linear class boundaries in the feature space [Fukunaga, 1990].

To get the posterior probability, both the prior and the Gaussian parameters for the likelihood function have to be estimated from the training samples.

Class mean vectors and the covariance matrix are estimated from the training samples as:

$$\widehat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{j=1, \mathbf{Y}_j \in \mathcal{Y}_c}^{N_c} \mathbf{x}_j^\intercal \tag{1.15}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{c=1}^{C} \sum_{j=1, \mathbf{Y}_j \in \mathcal{Y}_c}^{N_c} (\mathbf{x}_j^\intercal - \widehat{\boldsymbol{\mu}}_c)(\mathbf{x}_j^\intercal - \widehat{\boldsymbol{\mu}}_c)^\intercal \tag{1.16}$$

Minimizing the negative log-posterior probability gives the LDA classification function for any input vector $\mathbf{x}$:

$$g_{LDA}(\mathbf{x}) = \arg\min_c \Big((\mathbf{x} - \widehat{\boldsymbol{\mu}_c})^\intercal \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}_c}) - 2\log(\widehat{\pi}_c) \Big) \tag{1.17}$$

Note that g_{LDA} is the best classifier for the zero-one loss function under the assumption that $Pr(\mathbf{x} \mid \mathcal{Y}_c)$ follows a Normal distribution $\mathcal{N}(\widehat{\boldsymbol{\mu}_c}, \widehat{\boldsymbol{\Sigma}})$.

Quadratic Discriminant Analysis is similar to LDA but the constraint on equal class covariance matrix is relaxed. The likelihood and negative log-likelihood are thus written as:

$$f_c(\mathbf{x}) = (2\pi)^{-P/2} |\boldsymbol{\Sigma}_c|^{-1/2} \exp\big(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) \big) \tag{1.18}$$

and,

$$\ell_c(\mathbf{x}) = -2\log\big(f_c(\mathbf{x})\big) = P \cdot \log(2\pi) + \log\big(|\boldsymbol{\Sigma}_c|\big) + (\mathbf{x} - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) \tag{1.19}$$

for which it can similarly be shown that the boundaries are quadratic functions in the feature space [Fan et al., 2011]. The C class covariance matrices are estimated from the training samples as:

$$\widehat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c} \sum_{j=1, \mathbf{Y}_j \in \mathcal{Y}_c}^{N_c} (\mathbf{x}_j^{\top} - \widehat{\boldsymbol{\mu}}_c)(\mathbf{x}_j^{\top} - \widehat{\boldsymbol{\mu}}_c)^{\top} \tag{1.20}$$

Minimizing the negative log-posterior probability gives the QDA classification function for any input vector $\mathbf{x}$:

$$g_{QDA}(\mathbf{x}) = \arg\min_c \left((\mathbf{x} - \widehat{\boldsymbol{\mu}_c})^{\top} \widehat{\boldsymbol{\Sigma}_c}^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}_c}) + \log\left(|\widehat{\boldsymbol{\Sigma}_c}|\right) - 2\log(\widehat{\pi}_c) \right) \tag{1.21}$$

which is the best classifier for the zero-one loss function under the assumption that $Pr(\mathbf{x} \mid \mathcal{Y}_c)$ follows a Normal distribution $\mathcal{N}(\widehat{\boldsymbol{\mu}_c}, \widehat{\boldsymbol{\Sigma}}_c)$.

Note on Fisher's Linear Discriminant (FDA)

This discrimination approach developed by Fisher [1936] and extended by Rao [1948] does not require Normally distributed classes nor equal class covariances as in LDA. It was indeed developed as a dimension reduction method that finds the linear subspace that maximally separates the class centroids while minimizing the class spread. In this subspace, whose maximum dimension is given by $\min(C-1, P)$, classes are thus better separated than in the original space, which in turn leads to higher classification performances [Hastie and Tibshirani, 1996]. In the feature space, the directions $\mathbf{D}$ that best discriminate classes, maximize the variance ratio of between- to within-group scatter, i.e.,

$$\frac{\mathbf{D}^{\top}\mathbf{B}\mathbf{D}}{\mathbf{D}^{\top}\mathbf{W}\mathbf{D}}, \tag{1.22}$$

with $\mathbf{W} = N\widehat{\boldsymbol{\Sigma}}$, $\mathbf{B} = \sum_{c=1}^{C} N_c(\widehat{\boldsymbol{\mu}_c} - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{\mu}_c} - \widehat{\boldsymbol{\mu}})^{\top}$ and where $\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{\top}$ is the mean vector over all samples.

These directions are given by the $\min(C-1, P)$ leading eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$. It can however be shown [Hastie and Tibshirani, 1996] that maximizing the likelihood of LDA with some rank constraints on the mean vector matrix is equivalent to this Fisher's Linear Discriminant.

1.2.3 Discriminative model

Discriminative models learn the boundaries between classes without estimating class likelihood as illustrated in Figure 1.7b.

K-Nearest Neighbors (KNN) [Cover and Hart, 1967] is one of the simplest non-parametric discriminative classifier. KNN tends to construct the posterior class probability $Pr(\mathcal{Y}_c \mid \mathbf{x})$ without making any statistical assumption on class distributions. KNN finds the K-closest neighbors of a given vector $\mathbf{x}$ and uses a majority of voting to assign the class label. K is set as a positive integer that is usually small, e.g., between 1 to 7 are typical values. Cross-validation procedure can help to chose the optimal value that depends on the required complexity of the classification frontier. Note that the term 'closest' depends on the chosen distance, which is usually either the Euclidean distance, or, more frequently with spectral data, the Spectral Angle [Yuhas et al., 1992].

KNN naturally manages non-convex and non linearly separable classes but is however relatively slow and requires to store the whole training samples for classification.

The Support Vector Machine (SVM) is a linear binary classifier that aims at finding the furthest separating hyperplan to the closest point in both classes directly in the feature space [Vapnik, 1998]. In SVM, class labels are noted $Y_i = \pm 1$. The separating hyperplan $H_P \in \mathbb{R}^P$ is defined by its normal vector $\mathbf{w} \in \mathbb{R}^P$ and its bias $b \in \mathbb{R}$:

$$\mathbf{w}^\intercal \mathbf{x} + b = 0, \ \forall \mathbf{x} \in H_P \tag{1.23}$$

The distance from $\mathbf{x}$ to H_P is given by:

$$f(\mathbf{x}) = \frac{\mid \mathbf{w}^\intercal \mathbf{x} + b \mid}{\parallel \mathbf{w} \parallel} \tag{1.24}$$

In the linearly separable case, the optimal hyperplan parameters are given by:

$$\arg\max_{\mathbf{w},b} \left(\frac{1}{\parallel \mathbf{w} \parallel} \min_i \left(Y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \right) \right) \tag{1.25}$$

This complex optimization problem can be broken into a simpler quadratic optimization $\arg\min_{\mathbf{w},b} \left(\frac{1}{2} \parallel \mathbf{w} \parallel^2 \right)$ under the inequality constraints $Y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geqslant 1, \ i = 1, \cdots, N$. The optimal solution is then computed using Lagrange multipliers [Schölkopf and Smola, 2002].

Note that in some linearly non-separable cases, an additional term $\xi_i,\ i = 1, \cdots, N$ is added to the optimization problem to allow some of the training vectors to lie in the 'wrong' side of the separating hyperplan [Vapnik, 1998].

However, the key for the success of SVM is that in case of non-linearly separable classes, the use of the Kernel Trick (see Section 1.2.5) is directly applicable.

1.2.4 Multi-class

Binary classifiers such as SVM do not naturally enable multi-class classification problems to be solved. Two strategies are commonly considered to solve this issue:

One-versus-all strategy creates C binary classifiers $g_c,\ c = 1, \cdots, C$ during the training phase which distinguish each class from all the other ones. A new observation $\mathbf{x}$ is then classified with the label of the class of highest score $y = \arg\max_c \big(g_c(\mathbf{x})\big)$.

One-versus-one strategy creates $C(C-1)/2$ binary classifiers to distinguish each pair of class i and j. The classification is given by $y = \arg\max_i \big(\sum_{j \neq i} g_{i,j}(\mathbf{x})\big)$.

One-versus-one strategy requires more classifiers to be trained but is usually preferred as C is rarely large enough to be computationally too demanding. One-versus-one strategy also enables less complex class separators to be found.

1.2.5 The Kernel Trick

Linear classifiers cannot, by definition, properly classify non-linearly separable classes directly in the feature space. In such cases, if the classifier only depends on dot products, it can benefit from the so-called *Kernel Trick* [Vapnik, 1998]. It consists in mapping the vector from the feature space to a higher dimensional space in which classes become linearly separable. The mapping is performed by a kernel function that has to respect *Mercer conditions* [Vapnik, 1998]. The two most used Kernels are:

- Polynomial: $\phi(\mathbf{x}, \mathbf{x}') = \big(\mathbf{x} \cdot \mathbf{x}' + c\big)^a$
- Gaussian: $\phi(\mathbf{x}, \mathbf{x}') = \exp\big(- \parallel \mathbf{x} - \mathbf{x}' \parallel^2 / (2\sigma^2)\big)$

where c, a, and σ correspond to the kernel parameters that have to be tuned. The great idea behind this Kernel Trick is that computations do not have to be made explicitly in the high dimensional space [Schölkopf and Smola, 2002].

1.2.6 Training and assessing a classifier performance

When setting a classification model the question of complexity is a topic of major interest that has been discussed thoroughly in [Esbensen and Geladi, 2010]. The core idea is that increasing the complexity of a model by only observing the error made with the training data is prone to overfitting and has to be avoided. A good model should be complex enough to fit well the training set as well as be generic enough to classify accurately also an independent test set that would be acquired in the same experimental conditions. This is illustrated in Figure 1.8. As the model complexity increases better performances are obtained in both the training and the test set until a certain trade-off complexity region is reached. In this region, a minimum error is obtained for the test set while the training set error keeps decreasing. This particular region corresponds to the best trade-off between fitting to the training set and generalizability to unseen data. A more complex model loses in generalizability because the training set 'noise' is learned instead of real discriminatory features. In this figure, examples of classification boundaries are also given for a low, optimal and too complex models.

As it is often difficult to get independent test samples, internal cross-validation (CV) is often used to assess the model complexity using only the training samples. The simplest CV is to keep a part of the training set apart and use it as a validation set. Another technique, called leave-p-out is a CV technique for which p observations are left out to test the model that is build on the $N - p$ remaining training observations. This process is repeated with p other observations until all observations have been left out only once.

The CV procedure is useful to find the optimal model complexity and to tune its parameters. However, in order to provide an estimation of the predictive ability of the trained classifier on future samples acquired in the same experimental conditions, it is highly recommended to use an independent test set that has not been used yet.

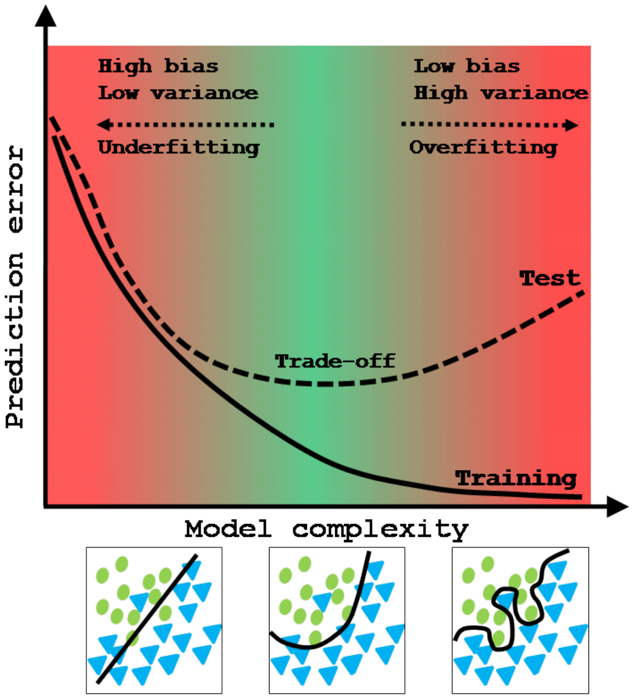


Figure 1.8: Illustration of the overfitting issue in classification.

1.3 Classification issues with HS data

The type and amount of information provided by HS data have to be considered when setting up a classification. For classification purposes, although the large number of spectral bands provided by the HS camera also means potentially more useful discriminatory information, there are some issues with high dimensional spaces. The spatial information also has to be considered for optimal results. The acquired image has to be corrected from different nuisance effects before getting a reflectance image that is related only to the object chemistry.

1.3.1 Problems with the spectral dimension

There are several problems related to the use of spectral data for classification purposes, which are due to the fact that we try to model a low-dimensional 'structure' embedded in a high-dimensional space using only few observations. In practical applications, it is usually impossible to use generative classifiers because of the difficulty associated to the statistical estimation of $Pr(\mathbf{x} \mid \mathcal{Y}_c)$ as the dimension of $\mathbf{x}$ increases. Discriminative classifiers, although generally directly applicable to high dimensions, are affected as well by the high dimensionality in terms of robustness because the space emptiness makes the class boundaries difficult to learn.

The high dimensionality of spectral data is subjected to the so-called *curse of dimensionality*, first named by Bellman and Kalaba [1965] to emphasize their dynamic search strategies for the estimation of multivariate functions. Bellman stated that, as the number of dimensions (P) increases, the number of evaluations needed to estimate a function on a regular grid was correspondingly increasing to the power 2^P. An illustration from Bishop's book on pattern recognition [Bishop, 2007] illustrates this phenomenon on one to three dimensions (Figure 1.9).

A trivial example, in which $\mathbf{x}$ is a Boolean vector of dimension 30, requires the estimation of more than 3 billion parameters [Tom M., 2005]. Typical HS classification problems involve vectors of dimensions of more than a hundred. The estimation thus requires an amount of observations that is unmanageable for any possible application.

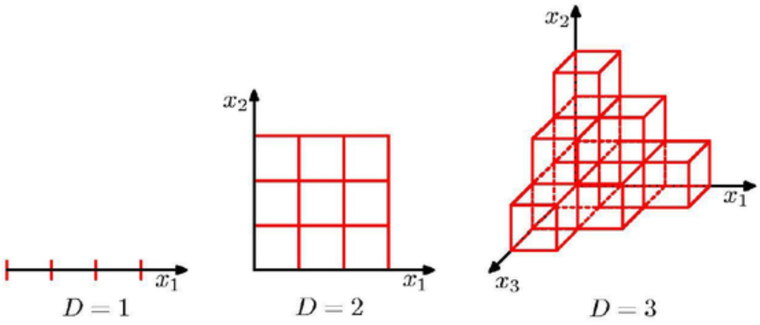


FIGURE 1.9: Illustration of the Curse of dimensionality, showing how the number of regions of a regular grid grows exponentially with the dimensionality D of the space. For clarity, only a subset of the cubical regions are shown for D =3. (from [Bishop, 2007]).

Reference papers that detail the high-dimension issues are: [Donoho, 2000] from a general pattern recognition point of view; [Jimenez et al., 1998] from a HS and multi-spectral point of view; [Tormod and Bjorn-Helge, 2001] from a chemometrics point of view. In the following, we briefly state and illustrate the main problems:

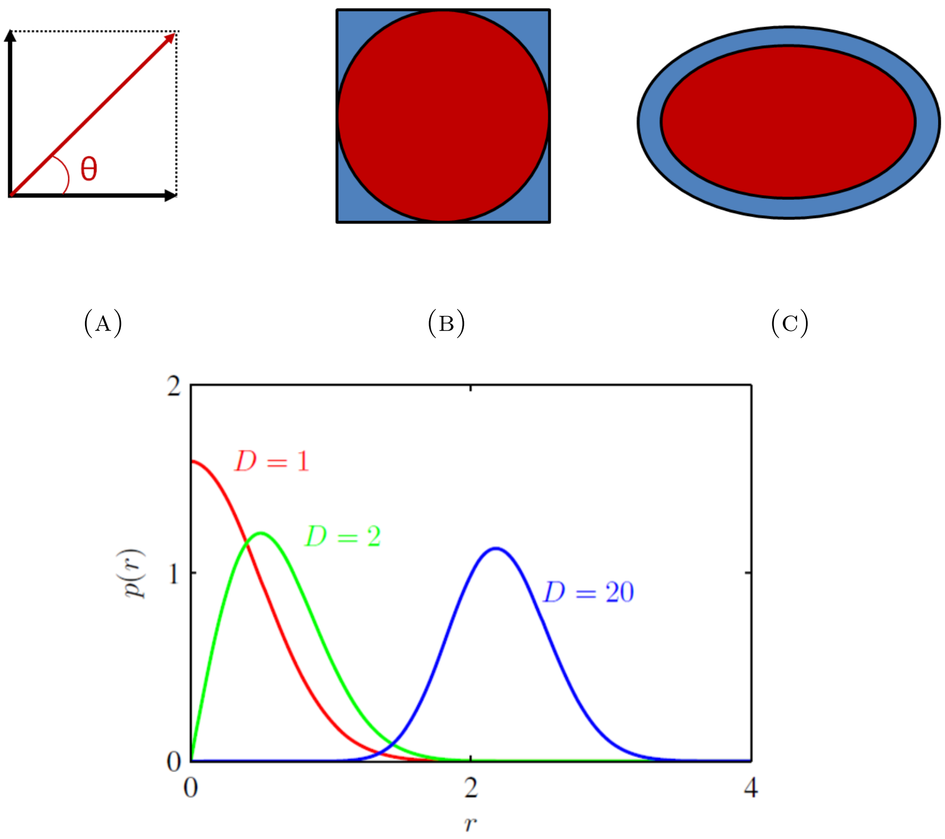
Geometry in high dimensional spaces cannot directly be translated from the usual 3D space to higher dimensions [Kendall, 1961]. Therefore, our intuition is often not right, which does not help in building new methods. A classical example is that the diagonals in high dimensional spaces tend to be orthogonal to the Euclidean coordinate axis as the space dimensionality increases (Figure 1.10a).

The cosine of this angle, given by $cos(\theta_P) = \pm(P)^{-1/2}$, approaches zero as P increases. Thus, projecting some spectra orthogonally to the diagonal vector, which is done when averaging a spectrum [Boulet and Roger, 2012], projects them close to the zero coordinate [Jimenez et al., 1998], losing localization information in the original space.

Another geometrical phenomenon that happens in high-dimensional spaces is called the *concentration of the measure*, which states that high dimensional regions are mostly empty because data tend to concentrate in a thin layer at the boundary of the regions. For instance, it was demonstrated that as the dimensionality P of the space increases:
1) the volume of the hypercube concentrates in its corners (Figure 1.10b).
2) the volume of a hyperellipsoid concentrates in its outside shell (Figure 1.10c).
3) the Normally distributed data tend to concentrates in the tail of the distribution, thus losing its bell shape (Figure 1.10d).
Each observation neighborhood in the feature space is thus likely to be empty. Hence, statistical density estimations have to be made using large bandwidths therefore losing fine spectral details. However, because high dimensional spaces are mostly empty, a lower dimensional structure containing the information is likely to exist.

Statistical estimations require an increasing number of training samples as the dimensionality increases. Hughes [1968] proved that with a limited number of available samples, the accuracy of statistical estimations started decreasing past some dimensions (see Figure 1.11). For a parametric classifier, the required number of training samples was estimated to be linearly related to the dimensionality and to the square for a quadratic classifier [Fukunaga and Hayes, 1989]. Similarly, for non-parametric classifiers, in order to get accurate estimations of multivariate densities, the required number of training samples is exponentially related to the dimensionality of the space. Because of the complexity involved in obtaining large ground truth information, it is often not possible to meet any of these criteria.

Collinearity among variables is a well known problem with spectral data. This problem, related to matrix conditioning, is due to the very high intercorrelation of the measured spectral variables. Therefore, even if the number of training samples was much larger than the number of variables, because of spectral 'smoothness', the actual dimension Q that captures all the spectral variability is smaller than P. The spectral matrix $\mathbf{X}$ is thus rank deficient leading to numerical stability



(A) (B) (C)

(D)

FIGURE 1.10: Geometrical problems in high dimensional spaces. As the space dimensionality increases: (A) the diagonal becomes orthogonal to the basis vectors (θ increases), (B) the volume of the hypercube is increasingly concentrated in its corners (the blue to red volume ratio tends toward zero), (C) the hyperellipsoid volume is concentrated in its outside shell (the blue to red volume ratio tends toward zero) and (D) (from [Bishop, 2007]) the probability density of a D-dimensional Gaussian distribution as a function of the radius r (distance from the mean).

problems for computations. The most problematic case is when the inverse of the covariance matrix $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ has to be computed. Small eigenvalues that only correspond to measurement noise have a large effect on the inversion leading to instability. Figure 1.12a represents the covariance matrix of a typical HS image as an image of $P \times P$ pixels for which low to high values are coded from blue to red. Vectors in this covariance matrix are clearly collinear, which is proven by the corresponding eigenvalues plot of Figure 1.12b.

For practical applications this instability directly affects:

- Classical least squares methods: $(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}$

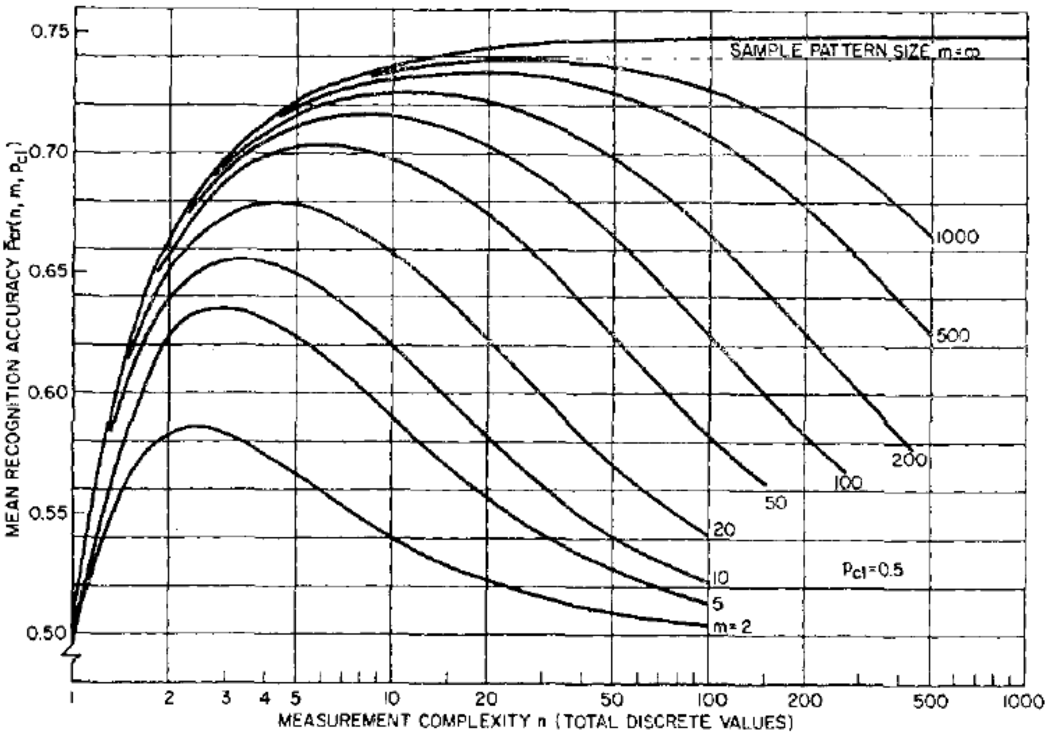
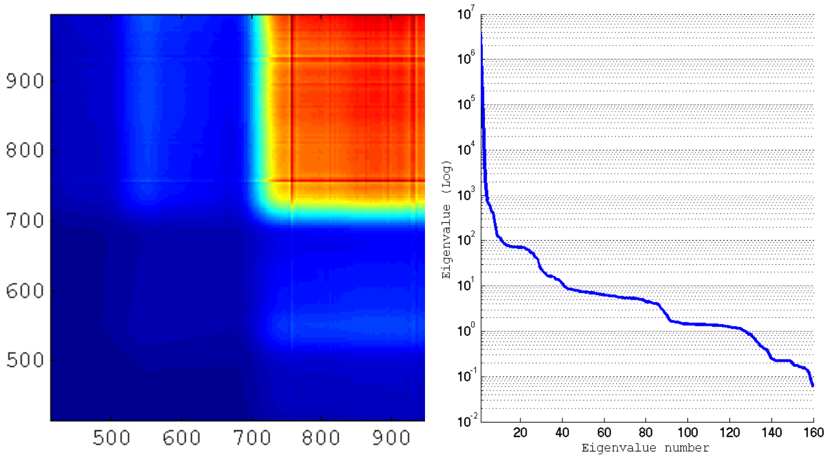


FIGURE 1.11: Accuracy of statistical estimation as the dimensionality of the space increases for various training set sizes. (from [Hughes, 1968])

- Covariance matrix inversion for LDA: $\widehat{\mathbf{\Sigma}}^{-1}$
- Covariance matrices inversion for QDA: $\widehat{\mathbf{\Sigma}_c}^{-1}$, $c = 1, \cdots, C$
- Fisher's Linear Discriminant axes computation: $\mathbf{W}^{-1}\mathbf{B}$

Normality after projection: it was demonstrated that as the dimensionality increases, the linear projection of any data set in a lower dimensional space is likely to be Normally distributed [Diaconis and Freedman, 1984, Hall and Li, 1993]. This fact is highly useful in practice since it does not requires infinitely large space to observe this phenomenon. Jimenez and Landgrebe [1996] confirmed that a uniformly distributed data in high dimension is Normally distributed after projection in low dimension. The implication for classification is that multi-modal classes can behave like a mono-modal class after projection. Figure 1.13 from [Jimenez and Landgrebe, 1996] show the effect of the projection of mono modal uniformly distributed data and bi-modal Normally distributed data in a lower dimensional space.



(A)

(B)

Figure 1.12: Illustration of high collinarity in the covariance matrix of spectral data on a data set containing three classes: wheat, weed and soil.

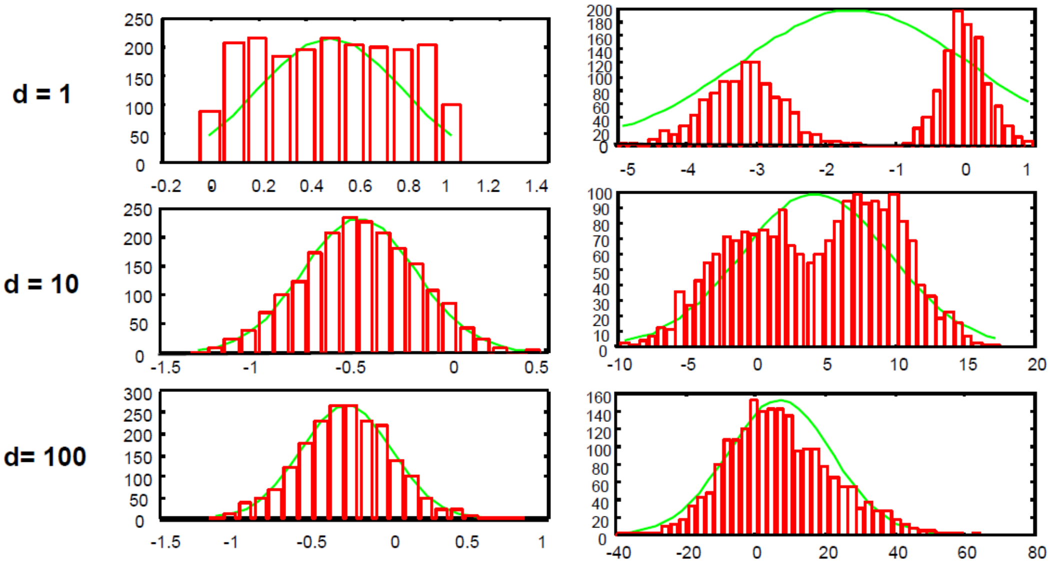


Figure 1.13: Normality after projection (from a space of dimension d) illustrated with generated data [Jimenez and Landgrebe, 1996]. (Left) one class with uniform distribution, (Right) two classes with Normal distribution.

Conclusion

In this section we have described problems related to high dimensionality of spectral data and noticed that thanks to space emptiness and variables collinearity, the data could be reduced with interesting properties. In particular, after dimension reduction, the Normal hypothesis required for QDA can be meet thus approaching the optimal Bayes classifier. Based on these considerations, some of the most important dimension reduction methods are described in the next chapter.

1.3.2 Using spatial information

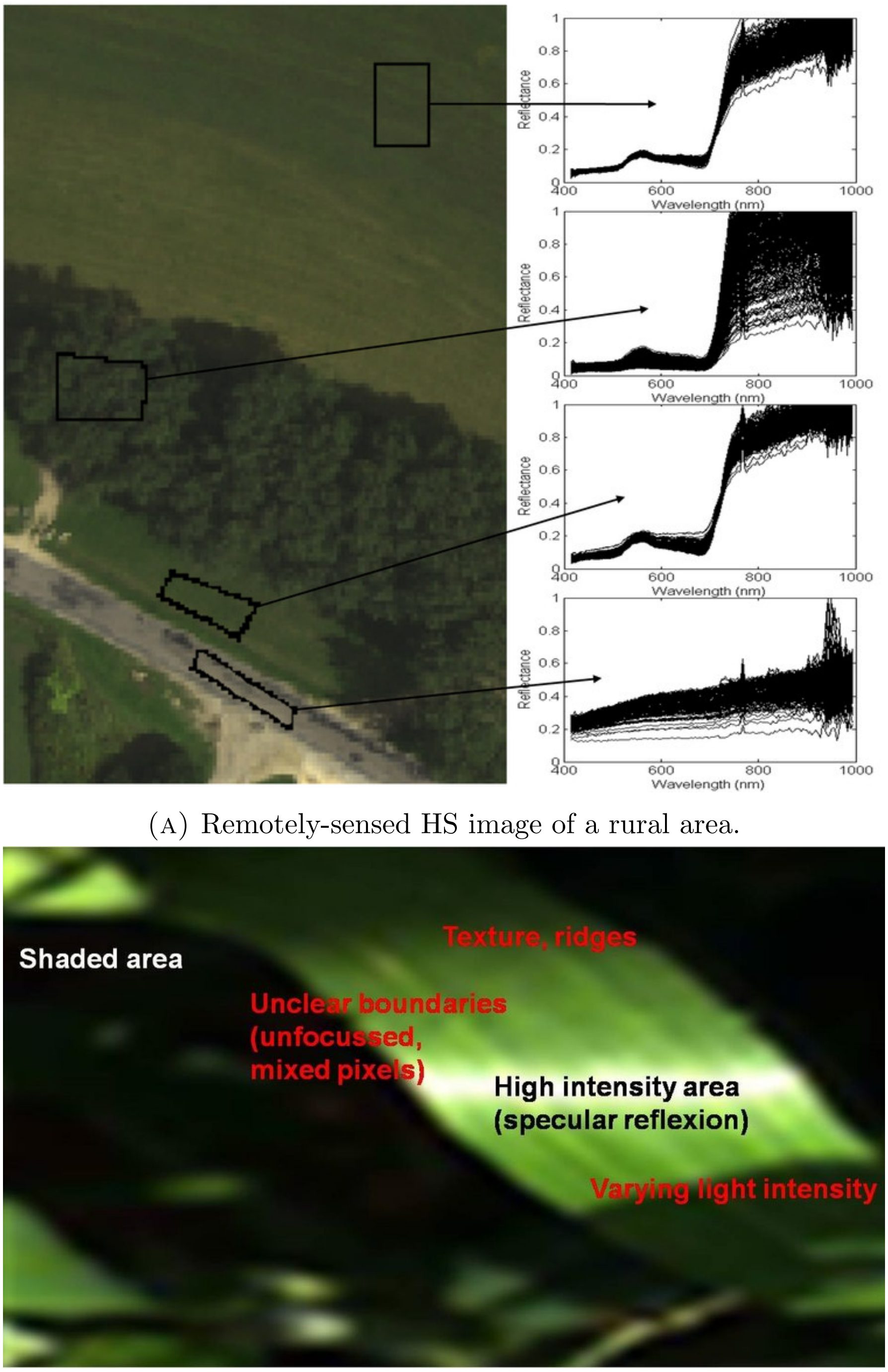
Until now we have described classification methods applied directly on spectral data. For instance, using a so-called pixel-based or spectral classifier only treats the HS data as a list of spectral measurement without considering spatial relations of adjacent pixels, thus discarding important information. However, the classification results could be improved by using the contextual spatial information provided in the HS data in addition to the spectral information. As illustrated in Figure 1.14, depending on the acquisition scale, different sources of spectral variability are present within objects, which could be managed through spatial information. To this end, from the famous Extraction and Classification of Homogeneous Objects (ECHO) method developed by Kettig and Landgrebe [1976], a great deal of research have been carried out to find effective spectral-spatial classifiers [Fauvel et al., 2013].

These methods, depending on what type of information is more discriminatory for the objects to classify, fall into three categories:
(1) If the objects to classify have strong spatial discriminatory features, these spatial features are extracted and then used to feed a classifier.
(2) If objects to classify have strong spectral and spatial discriminatory features, both are extracted and then used simultaneously in a classifier through kernel techniques.
(3) If objects to classify have strong spectral discriminatory features, spectral information is first processed and the spatial pixels neighboring information is then used to enhance the classification results.

The two first approaches are usually employed to discriminate classes with *a priori* information on objects shapes or textures, e.g., buildings, houses, roads, row fields. On the contrary, the third approach only assumes a certain homogeneity in the spatial neighborhoods of pixels. In the next chapter, some successfully developed spectral-spatial approaches of these three categories are reviewed .

1.3.3 Obtaining reflectance images

In the ideal scenario, each object to classify can be represented by its spectral signature. However, many uncontrollable variability sources such as the light



(A) Remotely-sensed HS image of a rural area.

(B) Color representation of a high resolution short-range HS image of a wheat leaf.

Figure 1.14: Illustration of the sources of spectral variability.

source angle, the direction of view, the atmospheric condition and a number of other variables substantially affect the measured spectral response [Barrett, 2013]. Three main correction or calibration stages are usually applied to the measured image to compensate these sources of variability (Figure 1.15):

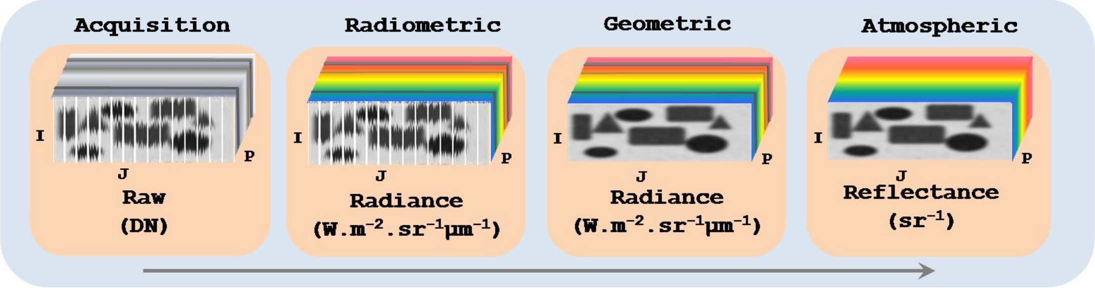


FIGURE 1.15: Main correction stages.

Radiometric calibration is a compulsory step before any further processing of the HS image. For each pixel, the recorded Digital Number (DN) obtained from the opto-electronic chain in the camera is converted into a physical measurement, i.e., radiance $(W.m^{-2}.sr^{-1}\mu m^{-1})$. The spectral calibration identifies the exact wavelength value associated with each band. Then, in order to quantify the exact amount of radiance, the transfer function of each pixel of the camera has to be evaluated [Gat, 2000]. For airborne or satellite imaging, HS cameras are usually calibrated in the laboratory using integrating spheres. Because of the cost of this procedure, cameras are often calibrated using the two-point techniques. In this case, the radiometric calibration is $L(\lambda) = A(\lambda) \cdot \big(DN(\lambda)–DC(\lambda)\big)$, where L is the radiance, A the pixel response , DN the recorder digital number and DC the dark current.

Geometric calibration mostly concerns images acquired with sensors that involved a scanning. These corrections focus on uncontrolled movement during the scanning, e.g., pitch, roll and yaw in airborne imaging and unequal speed for imaging using a conveyor belt. With staring systems, a registration between frames has to be performed if the objects were not perfectly still during the acquisition. In satellite and large field of view imaging, geometric distortion due to earth curvature also has to be taken into consideration.

Atmospheric and lighting correction is a prerequisite to every outside HS image analysis when the object surface reflectance has to be retrieved. For example, when spectra have to be compared with reference libraries or when a classification model calibrated on one image has to be used on other images. In order to be independent from the atmospheric conditions HS radiance images have to be transformed into reflectance images. The perfect atmospheric correction is an unsolved problem because of the complexity involved in modeling all possible interactions between light and atmospheric molecules. However, for practical applications two main strategies, reviewed in the following chapter, are usually employed: **Empirical corrections** that measure the received energy using a reference surface. **Modelisation** of the atmosphere radiative transfer that requires precise measures of the atmosphere at the acquisition time and solar spectrum estimation.

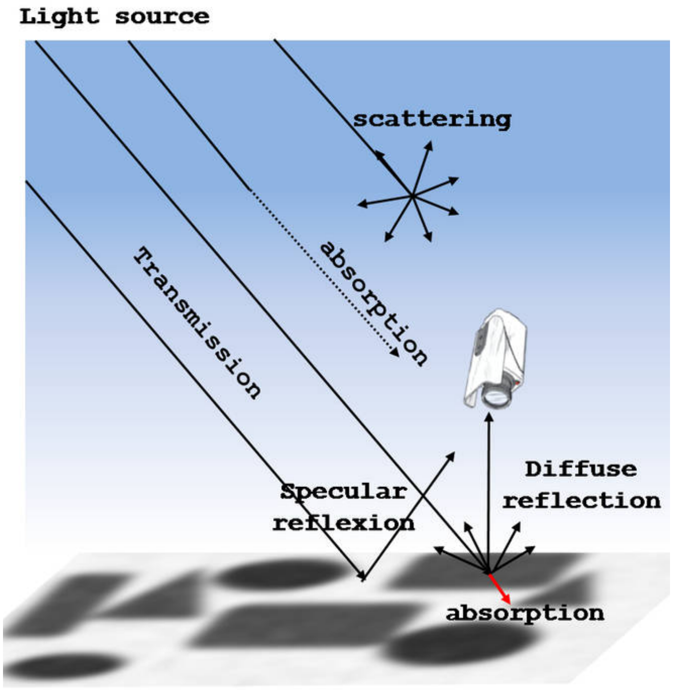


Figure 1.16: Schematic view of light interaction from source to sensor.

1.4 Conclusion

In this chapter, after having seen some main classification methods, we have detailed the main issues that are specific to classification of HS data. For instance:
1) The high dimensionality and collinearity of spectral data have to be dealt with to enable classification.
2) The spatial information should be used to help the spectral information in the classification process, especially because most of the variability is due to spatial inhomogeneities.
3) The acquired images have to be corrected from the atmosphere and lighting conditions in order to retrieve spectra that are only related to the object reflectance.

The remainder of this thesis is as follows. For each of these topics, a state-of-the-art is given in Chapter 3. The specific approaches developed in this thesis that address each of these topics are theoretically detailed in Chapter 4. Experimental results on real HS data as well as a detail discussion of the proposed approaches are given in Chapter 5. This thesis is concluded in Chapter 6 and future research directions are proposed.

Chapter 2

State-of-the-art

Contents

In the previous chapter, some general issues involving HS images classification were mentioned: dealing with high dimensionality of the data, including the spatial information in the classification process and being insensitive to illumination changes. In this chapter we review the available methods that were developed in the literature in order to deal with these issues. We first focus on spectral dimension reduction methods that can handle the high dimensionality and collinearity of the data. Second, we explore some of the available techniques that use both the spectral and spatial information for classification purposes. Third, after a quick background on the reflectance model, the main atmospheric correction methods are mentioned.

2.1 Introduction

Hyperspectral (HS) sensors can record images with a very detailed spectral information at each pixel that is related to the chemical properties of the targeted object. For classification purposes, differences in spectral responses are used to assign a label to each pixel of the HS image. In a supervised classification scheme, training samples with known labels are required to define the assignment rule. These training samples are manually assigned and necessitate the prior establishment of a ground truth. However, the processing of the obtained HS data is complex and thus requires adapted procedures.

This chapter gives the state-of-the-art regarding three main HS data classification issues, i.e., spectral dimension reduction, combination of spectral and spatial information and reflectance correction.

The first issue is due to the high dimensionality and collinearity of spectral data that make the supervised classification problem ill-posed. Furthermore, because of the often limited availability of training samples with respect to the data dimension, specific processing methods have to be used.

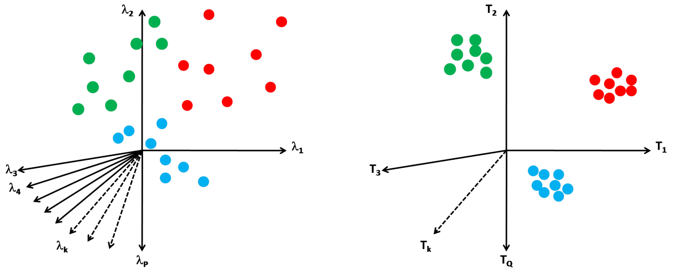
The second issue is to design effective ways to use the spatial information to increase the classification performances obtained using only spectral information.

The third issue is that, in order to be independent from the light source, HS radiance images first have to be transformed into reflectance images. In fact, in the general case, only a classification model calibrated with a reflectance image can be used to classify, other (also corrected) images.

The objective of the following sections is not to present an exhaustive survey of the available methods, but to give an overview of how researchers, from different communities (chemometrics, remote sensing and pattern recognition), have tackled these issues.

2.2 Dealing with the high-dimensionality of spectral data

Due to their ability to perform accurate and non-destructive measurements, hyperspectral imaging devices have been increasingly used in many scientific and industrial fields over the last decades. Spectral data acquired by these devices are often composed of more than a hundred narrow bands which make the classical classification techniques fail. In practice, because spectral variables are also highly correlated (which can be observed looking at the smoothness of the spectra observed as a function of the wavelength), their dimension can be reduced without loosing important information [Geladi, 2003, Wold et al., 2001]. Therefore, most methods include a dimension reduction as a first processing step, which is usually followed by a classical multivariate statistical method such as Multiple Linear Regression (MLR) when the responses are quantitative (concentrations) or Linear Discriminant Analysis when the responses are qualitative (classes) [Naes et al., 2002, Nocairi et al., 2005]. For classification, in the lower dimensional space, data are hoped to be well separated, i.e., small class spread and large distance between classes as represented in Figure 2.1.



(A) Original space $\mathbb{R}^P$ (B) Optimal space $\mathbb{R}^Q$

FIGURE 2.1: Illustration of the optimal dimension reduction method from a classification perspective.

Dimension reduction (DR) methods can be performed either in a supervised or unsupervised way. Unsupervised DR methods find a new set of variables (also called features or scores) only by analyzing the spectral matrix $\mathbf{X}$ of size $(N \times P)$ without any knowledge about the class membership of the spectrum. On the contrary, supervised methods also use the class membership matrix $\mathbf{Y}$ of size $(N \times C)$ to perform the reduction. The latter are thus usually preferred to find features that are relevant to classification [Indahl et al., 1999, Kemsley, 1996, Nocairi et al., 2005].

Dimension reduction (DR) methods are usually designated into two categories, i.e., feature selection (FS) and feature extraction (FE):
(1) FS methods find features by selecting a subset of variables from the spectra.
(2) FE methods project the data into a lower dimensional subspace whose axes are defined as linear or non-linear combination of the input variables.

For some applications, by removing non-informative or noisy wavelengths, FS proved to perform well [Xiaobo et al., 2010]. The predictive ability of models obtained using all the available wavelengths can in fact be greatly reduced if some parts of the spectra are corrupted by noise [Balabin and Smirnov, 2011]. An advantage of FS is that the extracted features are more easily understandable because they are actually related to the absorption properties of the studied media.

However, in the general cases, FE performs better. Moreover, similarly to classification methods, FE methods can benefit from the Kernel Trick if they only use dot products for computations. For example, Kernel PCA and Kernel LDA have been used in [Fauvel, 2007, Schölkopf et al., 1998] and [Baudat and Anouar, 2000]. However, linear FE techniques still generally outperform non-linear ones for real data [Van Der Maaten et al., 2009].

In the following we thus review both unsupervised and supervised linear FE techniques that are the most used with HS data.

The general idea behind linear FE approaches is to find the linear subspace that best summarizes the original data. Mathematically, this problem consists in the decomposition of $\mathbf{X}$ $(N \times P)$ such that:

$$\mathbf{S} = \mathbf{X}\mathbf{D} + \mathbf{E} \tag{2.1}$$

where X-scores $\mathbf{S}$ $(N \times Q \ll P)$ on X-loadings $\mathbf{D}$ $(P \times Q)$ minimize the reconstruction error $\mathbf{E}$ $(N \times P)$ in some sense. Different minimization choices lead to different methods.

2.2.1 Unsupervised approaches

Principal Component Analysis (PCA), also known as Karhunen–Loève transform and Hotelling transform, is undoubtedly the most common feature extraction method [Jolliffe, 2005]. PCA reduces the spectral dimension by keeping the principal components that best capture the data variability and the projected variables are de-correlated to one another [Wold et al., 1987].

The first PCA axis noted $\mathbf{d}_1$ maximizes the variance of data projection

$$\mathbf{d}_1 = \arg\max_{\|\mathbf{d}_1\|=1} \operatorname{var}\big(\mathbf{X}\mathbf{d}_1\big) = \arg\max_{\|\mathbf{d}_1\|=1} \operatorname{trace}\big(\mathbf{d}_1^\intercal \mathbf{X}^\intercal \mathbf{X}\mathbf{d}_1\big). \tag{2.2}$$

In the same way, the others axes are obtained by maximizing the captured variance and under some orthogonality constraints.

A solution is given by the eigenvectors of the symmetric[1] matrix $\mathbf{X}^\intercal\mathbf{X}$. In this context, this matrix is called total scatter matrix of $\mathbf{X}$ and is noted $\mathbf{T}(\mathbf{X})$.

The power of PCA is that this eigenvalue problem can be solved using the very efficient Singular Value Decomposition (SVD) [Bishop, 2007], which is defined as:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal \tag{2.3}$$

where $\mathbf{U}$ $(N \times P)$ and $\mathbf{V}$ $(P \times P)$ are orthogonal matrices and $\mathbf{S}$ $(P \times P)$ is a diagonal matrix that contains the singular values. Using this SVD decomposition on the scatter matrix gives

$$\mathbf{X}^\intercal\mathbf{X} = \big(\mathbf{U}\mathbf{S}\mathbf{V}^\intercal\big)^\intercal\big(\mathbf{U}\mathbf{S}\mathbf{V}^\intercal\big) = \mathbf{V}\mathbf{S}\mathbf{U}^\intercal\mathbf{U}\mathbf{S}\mathbf{V}^\intercal = \mathbf{V}\mathbf{S}^2\mathbf{V}^\intercal \tag{2.4}$$

where the eigenvectors are stored in $\mathbf{V}$ and the associated eigenvalues in $\mathbf{S}$.

When used for dimension reduction, only the $Q \ll P$ first principal components of $\mathbf{X}$ are retained. The Q first PCA scores are thus given by: $\mathbf{S} = \mathbf{X}\mathbf{D}$, where $\mathbf{D}$ contains the Q first PCA components. By maximizing the captured variance, for a given Q, the PCA obtains the minimum reconstruction error in the least square sense, i.e., $\| \mathbf{E} \|_2$.

Minimum Noise Fraction (MNF)

Green et al. [1988] proposed a method to find the most meaningful data variations

[1] The scatter matrix is symmetric and thus diagonalizable with orthogonal eigenvectors.

without taking into account noise-related variations. In fact, because PCA captures the variability information it also captures noise-related variability that is not relevant to summarize the useful information. MNF thus finds the linear subspace that maximizes the signal to noise ratio. The **X** matrix is first decomposed into

$$\mathbf{X} = \mathbf{X_S} + \mathbf{X_N} \tag{2.5}$$

where $\mathbf{X_S}$ and $\mathbf{X_N}$ respectively contain signal and noise variations from **X**. The source of variations are assumed to be uncorrelated and thus the covariance matrix $\Sigma = \Sigma_\mathbf{S} + \Sigma_\mathbf{N}$. MNF directions are found by maximizing the ratio

$$\arg\max_{\mathbf{d}} \frac{\mathbf{d}^\intercal \boldsymbol{\Sigma}_\mathbf{S} \mathbf{d}}{\mathbf{d}^\intercal \boldsymbol{\Sigma}_\mathbf{N} \mathbf{d}} = \arg\max_{\mathbf{d}} \frac{\mathbf{d}^\intercal \boldsymbol{\Sigma} \mathbf{d}}{\mathbf{d}^\intercal \boldsymbol{\Sigma}_\mathbf{N} \mathbf{d}} \tag{2.6}$$

which corresponds to the solution of the eigen-problem : $\boldsymbol{\Sigma}\boldsymbol{\Sigma}_N^{-1}\mathbf{d} = \lambda\mathbf{d}$.

MNF is therefore an interesting alternative to PCA when noisy wavelengths are present in the HS image. However, its practical use is often limited because of the difficulty involved in the estimation of the noise covariance matrix.

From unsupervised to supervised methods

Even though the previously described methods have been successfully implemented in many applications for classification purposes, they do not take into account the class information in the dimension reduction process. The optimal subspace is thus optimal in term of capturing the overall data variation but is not optimal from a classification point of view.

However, the aim of feature-reduction algorithms is not necessarily classification, but also representation. Several unsupervised algorithms are used to find a subspace to represent hyperspectral data, for visualization or processing.

PCA and MNF based methods produced a subspace in which each axis does not simply correspond to a single class but is generally a linear combination of spectral responses from several classes. To compensate this problem, Harsanyi and Chang [1994] proposed an approach using pure spectral signatures. Given these signatures, their method consists in removing the undesired effect by projecting the data in a subspace that is orthogonal to their variations. At the same time, Lee and Landgrebe [1993] proposed a feature extraction method that focuses on class boundaries in the feature space in order to avoid the high dimensionality problem. Their method, called Feature Extraction Based on Decision Boundaries

(DBFE), by modeling only the class boundary provided interesting results with fewer extracted features than methods that model the class probability densities.

Other approaches have received a lot of attention from the statistical pattern recognition and chemometric fields. Fisher Linear Discriminant Analysis (FDA) approaches tend to solve the computational issue of FDA in high dimensional space and Partial Least squares (PLS) based approaches use a covariance criteria in order to reduce the dimensionality.

2.2.2 PLS-like approaches

Original PLS

Partial Least Squares (PLS) has been designed by Wold [1966] to find a score subspace that takes into account the covariance between spectra and one or more responses (e.g., concentration of absorbing species). For this purpose, an iterative algorithm 'Non Linear Iterative Least Squares' was designed to give a set of axes called latent variables (LV) [Wold et al., 2001]. When the responses are discrete (e.g., classes), using Fisher Discriminant Analysis (FDA or LDA) on PLS scores (PLS-DA) has proven its efficacy for spectral discrimination [Barker and Rayens, 2003] and multivariate image analysis [Chevallier et al., 2006].

PLS uses the covariance between the matrix of input vectors $\mathbf{X}$ and their class $\mathbf{Y}$ instead of just using the variance of $\mathbf{X}$. The matrix $\mathbf{X}$ is of size $(N \times P)$, where N is the number of training samples and P the number of wavelengths in the digitalized spectrum; the matrix $\mathbf{Y}$ is of size $(N \times C)$ where each row codes the class membership (C classes) of the corresponding spectrum[2].

The aim of PLS is to transform the matrix $\mathbf{X}$ into a score matrix $\mathbf{S}$ $(N \times Q \ll P)$ using a weight matrix $\mathbf{D}$ (model) of size $(P \times Q)$ such as:

$$\mathbf{S} = \mathbf{XD}. \tag{2.7}$$

The capture of $\mathbf{X}$ variability is constrained by:

$$\mathbf{X} = \mathbf{SP}^{\intercal} + \mathbf{E}, \tag{2.8}$$

[2] e.g., $[0, 0, 1, 0]$ corresponds to the third class among four.

which means that the scores $\mathbf{S}$ summarize $\mathbf{X}$ by minimizing the residual reconstruction error $\mathbf{E}$. The decomposition is also constrained on $\mathbf{Y}$ by:

$$\mathbf{Y} = \mathbf{S}\mathbf{C}^{\intercal} + \mathbf{F}. \tag{2.9}$$

The scores $\mathbf{S}$ also have to summarize $\mathbf{Y}$ with a minimum reconstruction error $\mathbf{F}$. The matrices $\mathbf{P}$ and $\mathbf{C}$ are called the loadings of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

Orthogonal Projection to Latent Structure (O-PLS)

Wold et al. [1998] originally introduced Orthogonal Signal Correction (OSC) as a spectral pre-processing that removes systematic variations in $\mathbf{X}$ that are orthogonal to $\mathbf{Y}$. The solution proposed by the authors performed well but suffered from computational complexity issues and sometimes failed to converge. Fearn [2000] rapidly proposed an alternative approach that computes a similar correction by solving the eigenvalue problem of the matrix $\mathbf{M}\mathbf{X}^{\intercal}\mathbf{X}$ where:

$$\mathbf{M} = \mathbf{I} - \mathbf{X}^{\intercal}\mathbf{Y}\left(\mathbf{Y}^{\intercal}\mathbf{X}\mathbf{X}^{\intercal}\mathbf{Y}\right)^{-1}\mathbf{Y}^{\intercal}\mathbf{X}. \tag{2.10}$$

Finally, Trygg and Wold [2002] proposed an improved version of PLS that uses the power of OSC methods in order to clean the data before PLS.

From a classification point of view, OPLS-DA allows the separation of predictive from non-predictive variations as demonstrated in Bylesjö et al. [2006] and as illustrated in Figure 2.2. With OPLS-DA (right) the discriminatory direction t_{p1} is separated from the Y-orthogonal direction t_{o1}. The corresponding loading p_{p1} is thus easier to interpret and only one loading is necessary for perfect discrimination. With PLS-DA (left), discriminative information is shared between both loadings leading to a more complex interpretability.

2.2.3 FDA-like approaches

FDA background

In this section, we present some solutions to adapt Fisher's LDA (FDA) to high dimensional spaces. Because of the theoretical superiority of LDA- and QDA-based approaches for classification when classes are Normally distributed, Fisher-like dimension reduction methods that are based on a similar criterion have received particular attention in the pattern recognition field.

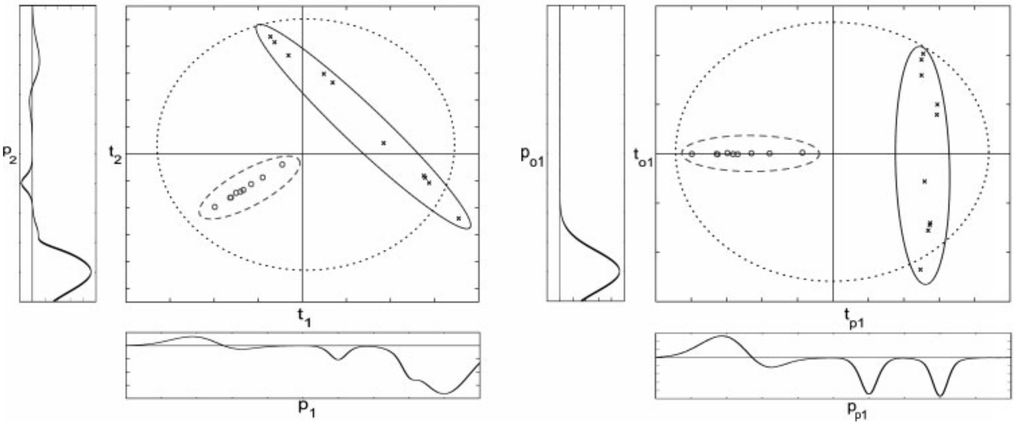


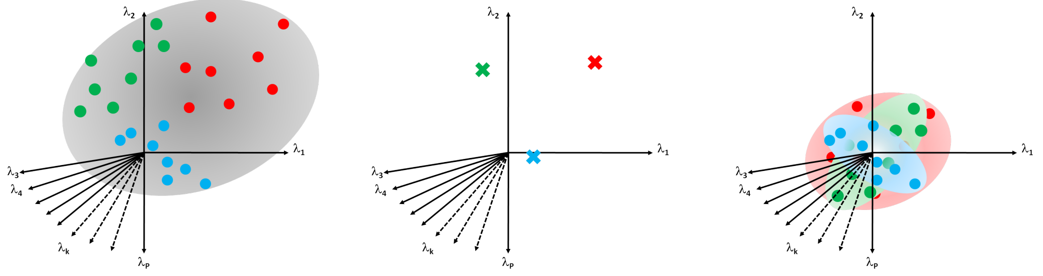
Figure 2.2: Comparison of (left) PLS-DA and (right) OPLS-DA on a two-class simulated data-set. (from Bylesjö et al. [2006]).

The total scatter matrix previously defined with PCA can be seen as the sum of the between- and within-class scatter matrices defined by

$$\mathbf{B}(\mathbf{X},\mathbf{Y}) = \mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X} \tag{2.11}$$

$$\mathbf{W}(\mathbf{X},\mathbf{Y}) = \mathbf{X}^\intercal\mathbf{X} - \mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X} \tag{2.12}$$

Note that when no confusion is possible these matrices are simply noted **T**, **B** and **W**. An illustration of this decomposition is given in Figure 2.3.



(a) Total (b) Between-class (c) Within-class

Figure 2.3: Space decomposition.

Originally, Fisher [1936] developed a method to find the optimal discriminant vector which maximizes the ratio of the between-class distance to the within-class distance for a two-class problem. Sammon [1970] generalized Fisher's idea by finding the optimal discriminant plane, and then Foley and Sammon [1975] proposed a complete set of discriminant vectors. Their method called Foley-Sammon transform (FST) uses the so-called Fisher criterion [Wilks, 1962], defined as:

$$\mathbf{D}_{FDA} = \arg \max_{\mathbf{D},\mathbf{D}^\intercal\mathbf{W}\mathbf{D}=\mathbf{I}} \operatorname{trace}\Big(\left(\mathbf{D}^\intercal\mathbf{W}\mathbf{D}\right)^{-1}\mathbf{D}^\intercal\mathbf{B}\mathbf{D}\Big). \tag{2.13}$$

where $\mathbf{D}^\intercal\mathbf{W}\mathbf{D} = \mathbf{I}$ corresponds to the orthogonality constraint. Although FST is considered as an optimum transformation, it suffers from many computational issues such as:
(1) The solution cannot be computed if the $\mathbf{W}$ matrix is singular.
(2) The number of optimal vectors is bounded by $\min(C-1, P)$.
(3) The transformation is sub-optimal.
(4) The transformation is not orthogonal with respect to $\mathbf{T}$.

These issues have attracted a lot of researchers' attention in the pattern recognition field and especially in the domain of Face recognition [Duin et al., 2006]. Some of the original solutions to these problems are explained in the following.

Modified criterion LDA: In order to overcome the singularity problem of the within-class scatter matrix, Cheng et al. [1992] proposed to use an alternative to the Fisher criterion that is defined by:

$$\mathbf{D}_{MCLDA} = \arg \max_{\mathbf{D}, \mathbf{D}^\intercal\mathbf{W}\mathbf{D}=\mathbf{I}} \mathrm{trace}\Big(\big(\mathbf{D}^\intercal\mathbf{T}\mathbf{D}\big)^{-1}\mathbf{D}^\intercal\mathbf{B}\mathbf{D}\Big), \tag{2.14}$$

which is proved to lead to the same discriminant vectors. This solution is only partial since it obviously requires the total scatter matrix to be non-singular, which is unfortunately not the case with many types of data and especially with spectral data.

Pseudo inverse LDA: Tian et al. [1988] proposed approximating the optimal Fisher criterion by replacing the inversion $\mathbf{W}^{-1}$ by its positive pseudo inverse $\mathbf{W}^{+}$. The positive pseudo inverse gives however only an approximation of the Fisher criterion.

Nullspace LDA: Chen et al. [2000] proposed computing the between-class maximization in the null space[3] of the within-class scatter matrix. Their method corresponds to solving the following problem:

$$\mathbf{D}_{NLDA} = \arg \max_{\mathbf{D}, \mathbf{D}^\intercal\mathbf{W}\mathbf{D}=0} \mathrm{trace}\big(\mathbf{D}^\intercal\mathbf{B}\mathbf{D}\big), \tag{2.15}$$

where the null space constraint is given by $\mathbf{D}^\intercal\mathbf{W}\mathbf{D} = 0$.

[3] The null space or Kernel of a matrix $\mathbf{A}$ is given by: $\{\mathbf{x} \in \mathbb{R}^P, \mathbf{A}\mathbf{x} = 0\}$ and its dimension is $P - \mathrm{rank}(\mathbf{A})$

The inversion problem is thus implicitly solved and interesting performance can be achieved if the null space contains enough discriminant information. It thus requires that the projection of $\mathbf{B}$ in the null space is non-zero which means that $\mathbf{B}$ and $\mathbf{W}$ eigenvectors are not collinear. Another problem is that the $\mathbf{W}$ null space is often quite large and that several dimensions do not help with the discrimination. Huang et al. [2002] used the relation $\mathbf{T} = \mathbf{B} + \mathbf{W}$ to reduce this dimensionality issue. Indeed, they showed that the null space of $\mathbf{T}$ was not helping with the discrimination. Therefore, they proposed to compute the discriminant vectors using the Chen et al. [2000] NLDA method but in a subspace excluding the null space of $\mathbf{T}$. Guo et al. [2006] proved that in cases of small number of samples ($N < P$) it is possible to find the $C - 1$ projecting directions in the null space of $\mathbf{W}$. This means that there exists a subspace in which no within-class variability is present as illustrated in Figure 2.4. However, in practice, the nullspace might not exists as N increases and it is not guaranteed that sufficient information remains for discrimination after projection in the nullspace.

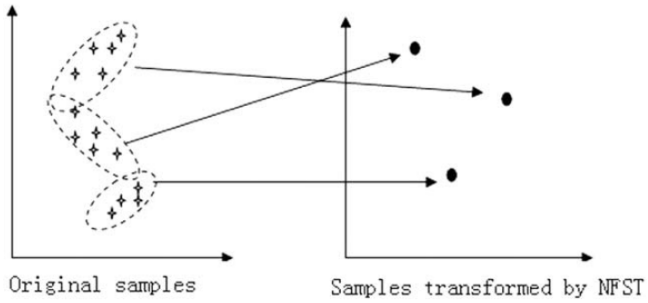


Figure 2.4: Geometric representation of Null Foley-Sammon Transform (from [Guo et al., 2006]).

Orthogonal LDA

Discriminant vectors found by FST are in general not orthogonal. The only cases in which the obtained basis is orthogonal is when $\mathbf{B}$ and $\mathbf{W}$ have the same set of eigenvectors [Hamamoto et al., 1993]. To solve this problem, Okada and Tomita [1985] proposed a method that is able to find up to $N - 1$ discriminant vectors that are orthogonal to one another. Their method, called Orthonormal Discriminant Vector (ODV), maximizes the Fisher criterion for each extracted feature under the constraint that features are orthogonal. This method was then proved to provide better results than conventional discriminant analysis in terms of the Fisher criterion [Hamamoto et al., 1993]. Another method proposed by Ye et al. [2005], called

Orthogonal LDA, uses the generalized LDA criterion in the optimization problem:

$$\mathbf{D}_{OLDA} = \arg\max_{\mathbf{D}^{\intercal}\mathbf{D}=\mathbf{I}} \operatorname{trace}\Big(\big(\mathbf{D}^{\intercal}\mathbf{T}\mathbf{D}\big)^{+}\mathbf{D}^{\intercal}\mathbf{B}\mathbf{D}\Big). \tag{2.16}$$

The orthogonal constraint on the discriminant vector is given by $\mathbf{D}^{\intercal}\mathbf{D} = \mathbf{I}$. The singularity problem is solved using the simultaneous diagonalization of the scatter matrices [Ye et al., 2005].

Regularized LDA: Hong and Yang [1991] used a regularization technique on the within-class matrix by adding a small perturbation to it. The regularization aims at increasing the rank of the singular matrix while keeping as much as possible the original information [Hastie et al., 1995, Witten and Tibshirani, 2011]. It corresponds to giving some penalty to excessively large value in the discriminant vectors obtained because of singularity. Krzanowski and Jonathan [1995] nicely worded this regularization process as: *If principal component analysis is viewed as providing the best r-dimensional approximation to a p-dimensional set of data, then our present objective can be seen as exactly the reverse, namely to provide the 'nearest' p-dimensional non-singular approximation to an r-dimensional singular set of data.* The resulting matrix thus becomes non-singular and can be inverted.

This is the basis of *ridge* regularization techniques [Zhang et al., 2010], which can be seen as the following optimization problem

$$\mathbf{D}_{RLDA} = \arg\max_{\mathbf{D}^{\intercal}\mathbf{D}=\mathbf{I}} \operatorname{trace}\Big(\big(\mathbf{D}^{\intercal}(\mathbf{W}+k\mathbf{I})\mathbf{D}\big)^{-1}\mathbf{D}^{\intercal}\mathbf{B}\mathbf{D}\Big). \tag{2.17}$$

Roger et al. [2005] proposed a continuum approach in order to solve this type of regularization problem for which they showed that the optimal solution was found in a space defined by union of the Kernels of: $\mathbf{B} - z\mathbf{T}$ where $z \in [0, 1]$.

Another way of regularizing is to use a hierarchical model as the one developed in [Brown et al., 1999], where the authors proposed estimating the covariance matrix in a Bayesian framework.

PCA-LDA: Pre-reducing the dimension using PCA before LDA is a commonly used technique for dimension reduction [Bertrand et al., 1990, Fearn, 2008, Naes et al., 2002]. Although PCA is not designed to help with the discrimination, it often works well in practice [Grahn and Geladi, 2007] and should at least be tried before using more complex methods [Fearn, 2011]. The advantage is that

PCA reduced variables are already orthogonal which helps with the further LDA processing. Some theoretical insights of the use of PCA plus LDA are given in Yang and Yang [2003] under the assumption of non-empty null space of $\mathbf{W}$.

2.3 Using spatial information: Spectral-spatial approaches

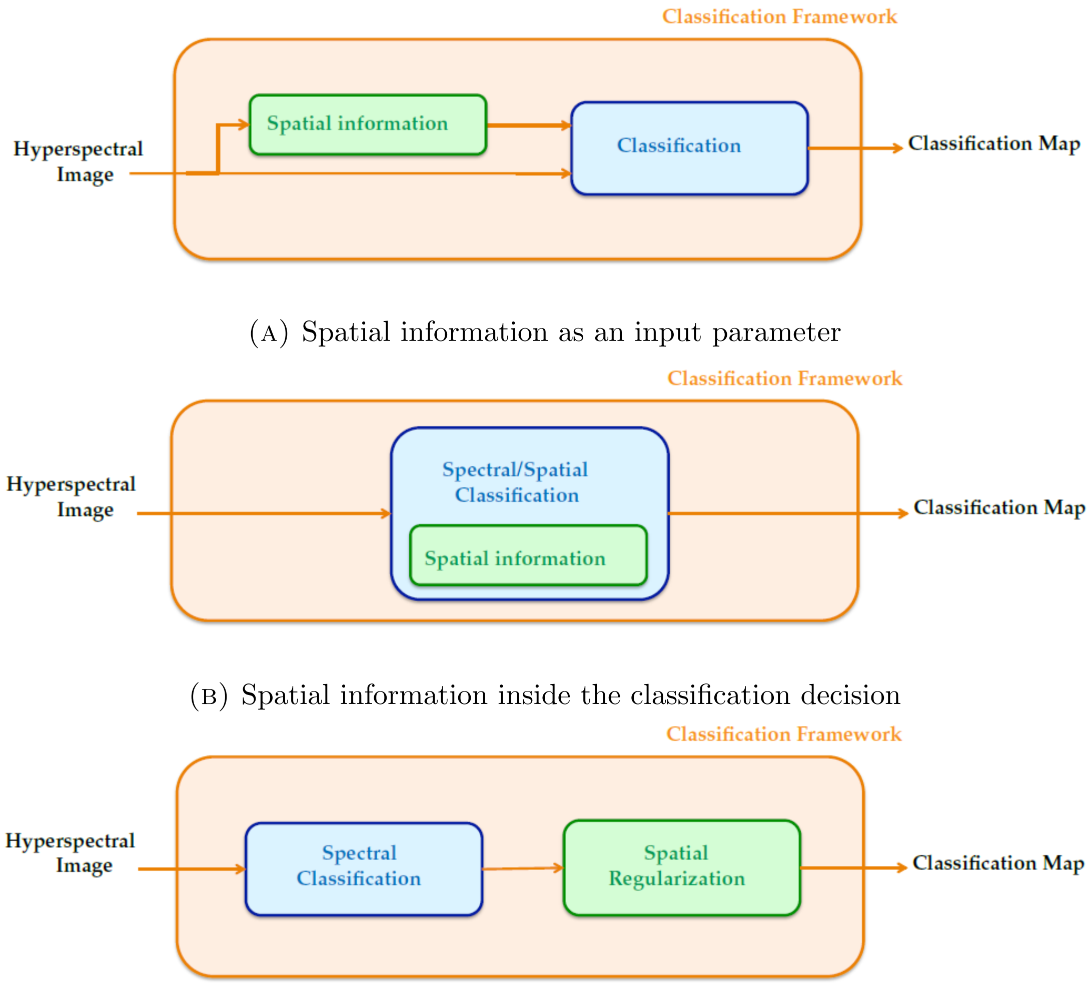
Every pixel-based classification method described in the first chapter usually performs well when the training set is representative enough and when classes to be discriminated are different enough in terms of spectral information. In other cases, in order to compensate for the lack of available spectral information, using spatial information provided by hyperspectral images has proved to be an important improvement [Dalla Mura et al., 2011, Gorretta et al., 2012, Tarabalka et al., 2010a].

Spectral-spatial methods for classification have had a short, but intense history and many papers have been published in the last decade, most of them being due to the remote sensing community [Bioucas-Dias et al., 2013, Fauvel et al., 2013]. These methods were originally classified into two families by Gorretta [2009] as:
(1) *Pixel-based classification with spatial constraints*
(2) *Extension of classical image processing techniques to HS image*: the main difficulty with this kind of method is to define a metric that makes sense in this high dimensional space and to create an ordering.

With the rapid development of new methods, this separation is now less obvious. It is preferred to define categories depending on the place where the spatial information is introduced in the classification chain, leading to three main categories [Bioucas-Dias et al., 2013]. A schematic view of these categories due to Valero [2011] is represented in Figure 2.5.

2.3.1 Spatial information as an input parameter

In this approach, a feature vector that contains spatial information is constructed for each pixel. It can contain any contextual information such as: shape, texture, orientation, size... These features are usually extracted from the image using



(A) Spatial information as an input parameter

(B) Spatial information inside the classification decision

(C) Spatial information as a post-processing stage

FIGURE 2.5: Pixel-based classification with spatial constraints. (from [Valero, 2011])

classical image processing techniques that have either been adapted to work in higher dimensions or applied on a spectrally reduced image.

Region growing segmentation

The first spectral-spatial classification method, originally developed for multispectral images, is the well-known ECHO (Extraction and classification of homogeneous objects) [Kettig and Landgrebe, 1976, Landgrebe, 1980]. With this method, the image is first segmented into homogeneous regions that are found using a recursive partitioning, i.e., 1) The image is partitioned into small rectangular regions of pre-defined sizes; 2) Adjacent regions that are similar enough according to an homogeneity criterion are merged 3) Step 2 is repeated until no more merging is possible. Each segmented region is finally classified using a classical maximum likelihood classifier.

Tilton [1998, 2010] adapted a sequential segmentation algorithm for hyperspectral data and proposed a more advanced hierarchical segmentation method (HSEG). Different similarity measures are used with spectral data. Successfully developed distances include
- Spectral Angle Mapper (SAM): $d_{SAM}(\mathbf{x},\mathbf{y}) = \cos^{-1}\left(\frac{\mathbf{x}^{\mathsf{T}}\mathbf{y}}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}\right)$
- Cross-entropy (Kullback-Leibler information measurement): $d_E(\mathbf{x} \parallel \mathbf{y}) = \sum_{i=1}^{P} a_i \log\left(\frac{a_i}{b_i}\right)$, where $a_i = \frac{x_i}{\sum_{l=1}^{P} x_l}$ and $b_i = \frac{y_i}{\sum_{l=1}^{P} y_l}$
- Spectral Information Divergence (SID): $d_{SID}(\mathbf{x},\mathbf{y}) = d_E(\mathbf{x} \parallel \mathbf{y}) + d_E(\mathbf{y} \parallel \mathbf{x})$

Mathematical morphology (MM) techniques were originally developed for binary image processing. Because of their potential , they have quickly been extended to work with grey-scaled and color images [Soille, 2003]. A detailed review of MM processing that involves HS images can be found in [Fauvel et al., 2013]. In the following, after a short description of the principal MM operators, two of the most important usages of MM in the case of HS image are given, i.e., watershed segmentation and morphological profiles.

Every MM technique needs the definition of a *structuring element*, noted B, of known shape and size (e.g., a disk of radius 5 pixels). In practice MM is very efficient for image processing because it is only based on the computation of minimum and maximum operations between the image I and this 'small' structuring element (SE).

The two basic MM operators are *dilatation* (noted δ_B) whose effect is to enlarge light areas compared to dark ones and *erosion* (noted ϵ_B) that corresponds to the dilatation of the negative of the image. For an image I, an erosion applied at pixel $\mathbf{x}$ is given by:

$$\epsilon_B\big(I(\mathbf{x})\big) = \min_{\mathbf{x}_i}\big(I(\mathbf{x}_i) \in B_{\mathbf{x}}\big) \tag{2.18}$$

and the dilatation is given by:

$$\delta_B\big(I(\mathbf{x})\big) = \max_{\mathbf{x}_i}\big(I(\mathbf{x}_i) \in B_{\mathbf{x}}\big) \tag{2.19}$$

where $B_{\mathbf{x}}$ is the structuring element centered at pixel $\mathbf{x}$. The other two most important MM processing operators are *opening* and *closing*. The opening operator

γ_B is defined by an erosion of I by B followed by a dilatation by B. On the contrary, the closing operator γ_B is defined by a dilatation of I by B followed by an erosion by B:

Because of the lack of ordering relation between vectors, the extension of these operators to HS image is challenging and no unique definition is available [Aptoula and Lefèvre, 2007].

Watershed segmentation uses a topological representation of a grey-scale image in which region boundaries are defined by the crest of the gradient image norm. The basic idea behind watershed segmentation is illustrated in Figure 2.6. The watershed algorithm defines the regions by simulating an elevation of the water level from the local minima in the image. When two neighboring regions are about to merge, a dam is added on top of the crest, the formed region are thus called catchment basins. The dams obtained at the end of the process correspond to the segmentation boundaries. Mathematically, watershed algorithm need the definition of a gradient image, which is not straightforward with HS images [Tarabalka, 2007]. The simplest solution is to consider a one-band image gradient computed from all bands such as the color morphological gradient (CMG) proposed by Evans and Liu [2006]. The CMG is computed as:

$$CMG_B(\mathbf{x_p}) = \max_{i,j \in \mathcal{X}} \left(\parallel \mathbf{x}_p^i - \mathbf{x}_p^j \parallel_2 \right) \tag{2.20}$$

where $\mathcal{X} = [\mathbf{x}_p^1, \cdots, \mathbf{x}_p^b]$ is the set of b vectors contained within the structuring element B. CMG thus corresponds to the maximum of the distances between all pairs of vectors in the set $\mathcal{X}$ [Fauvel et al., 2013]. The classical watershed [Soille, 2003] is then directly applied on this one-band gradient image to obtain a segmentation of the image. Some more advanced watershed techniques developed for HS images are found in Tarabalka et al. [2010a].

Morphological profile, similarly to granulometry, is a technique that sorts the object present in the image by their sizes. A series of morphological openings and closings with structuring elements of increasing sizes are applied to the image [Pesaresi and Benediktsson, 2001]. Closing thus suppresses small dark areas whereas opening suppresses small light areas.

A morphological profile thus results in a serie of images that contains objects of different sizes as illustrated in Figure 2.7, which can in turn be used as input of

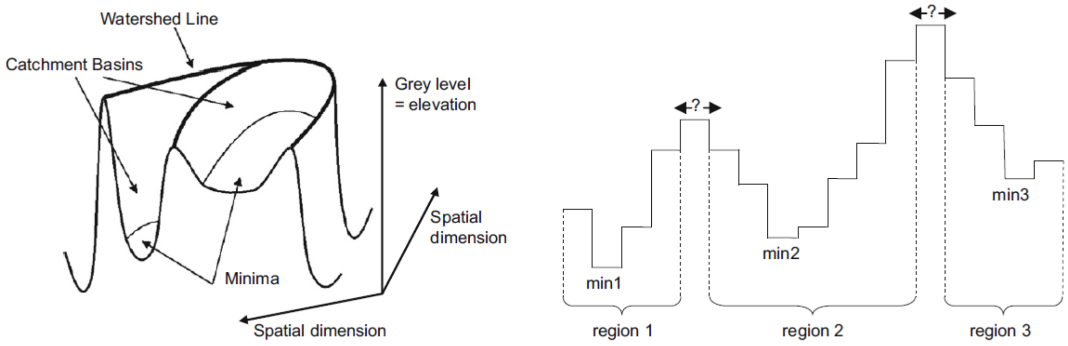


Figure 2.6: Representation of the watershed segmentation technique. (Left) topographic representation of a one-band image. (Right) Example of segmentation in one dimension. (from [Tarabalka et al., 2010a])

a classifier or combined with a spectral feature for enhanced processing [Fauvel et al., 2013]. Another recent approach was developed by Ghamisi et al. [2014] as an improvement of morphological profiles called *morphological attribute profile* (MAP). They combined in a classifier the output of the MAP with spectral features that were extracted using supervised FE techniques as those detailed in the previous section.

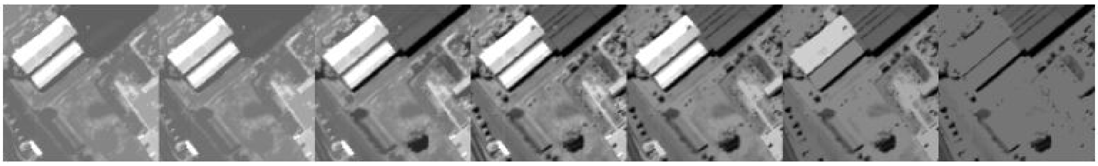


Figure 2.7: Morphological profile using a circular structuring element of size 2, 6 and 10. The left corresponds to the closings and the right to the openings. (from Fauvel et al. [2013])

Regularization

Spatial filters that are commonly used in image processing to enhance visual image quality and signal to noise ratio (SNR) can be adapted to HS images. It is well known that in order to keep objects spatial boundaries, using a classical low-pass filter is not relevant because of the blurring induced by this kind of method. Hence, Lennon et al. [2002] and Duarte-Carvajalino et al. [2006] proposed the use of a non-linear filtering method that preserves object borders. These kinds of transformations that fall under the name of *Edge Preserving Filtering* (EPF) have been commonly used in image and signal processing [Weickert, 1997, 1998]. One such EPF, due to Perona and Malik [1990], performs *anisotropic diffusion* in order to filter grey-scale images. Anisotropic diffusion filtering is a temporal process that mimics temperature diffusion in a physical medium. The diffusion process is written as:

$$\frac{\partial I(x,y,t)}{\partial t} = \mathrm{div}\Big(c\big(\mid \nabla I(x,y,t) \mid\big)\nabla I(x,y,t)\Big) \tag{2.21}$$

where div and ∇ are respectively the divergence and gradient operators. The condition at $t = 0$ corresponds to the initial image. When the function c is constant, the diffusion is isotropic and corresponds to a classical low-pass Gaussian filtering process [Lennon et al., 2002]. However, if the function depends on the local gradient in the gray-scale image, the diffusion process becomes anisotropic. With an adapted choice for c, the diffusion process can be stronger when there are low gradient and stop close to the border of the objects where there are high gradient values. Perona and Malik [1990] proposed a discrete version of this diffusion process:

$$I^{t+1}(i,j) = I^{t}(i,j) + 1/4\sum_{k=1}^{4} c_k^t(i,j)\nabla_k I^t(i,j) \tag{2.22}$$

where $c_k^t(i,j) = g(\mid \nabla_k I^t(i,j) \mid)$. The function has to be decreasing with respect to the gradient; they proposed:

$$g\big(\mid \nabla_k I^t(i,j) \mid\big) = \exp\big(-\frac{\mid \nabla_k I^t(i,j) \mid}{\eta}\big)^2 \tag{2.23}$$

where η is a Kernel width to be tuned. Perona and Malik [1990] implemented the discrete gradient using a 4-neighbors spatial connectivity [Gonzalez et al., 2009], the ∇_k being defined by:

$$\begin{aligned}
\nabla_1 I(i,j) &= I(i-1,j) - I(i,j)\\
\nabla_2 I(i,j) &= I(i+1,j) - I(i,j)\\
\nabla_3 I(i,j) &= I(i,j+1) - I(i,j)\\
\nabla_4 I(i,j) &= I(i,j-1) - I(i,j)
\end{aligned} \tag{2.24}$$

By definition of the diffusion described above, only scalar images can be filtered. For an HS image, each channel can be processed individually or a vector-valued diffusion has to be developed. This was first done by Whitaker and Gerig [1994] in the case of isotropic diffusion and was later extended to anisotropic diffusion by Weickert [1998]:

$$\frac{\partial I_i(x,y,t)}{\partial t} = \mathrm{div}\big(g(\theta)\nabla I_i(x,y,t)\big) \tag{2.25}$$

where θ corresponds to a vectorial measure of boundaries given by:

$$\theta = \sqrt{1/P \sum_{i=1}^{P} \parallel \nabla I_{\sigma,i}(x,y,t) \parallel^2} \tag{2.26}$$

where $I_{\sigma,i}$ corresponds to a low pass filtered version of I_i using a Gaussian Kernel of width σ.

In order to take into account the variability of spectral data due to illumination, Lennon et al. [2002] used a modified vector-valued gradient that uses a combination of Euclidean distance and spectral angle for the similarity measure. They also performed the spatial regularization on a MNF reduced image so that the noise related spectral variations were not taken into account.

Recently, Wang et al. [2010] used Weickert [1998] extensive work on tensor diffusion for vector-valued images in order to filter HS images. In the proposed anisotropic filtering scheme, the HS image is seen as a 3D image:

$$\frac{\partial I(x,y,z,t)}{\partial t} = \operatorname{div}\Big(\mathbf{D}^{*}\big(\nabla I_{\sigma}(x,y,z,t)\big)\nabla I(x,y,z,t)\Big) \tag{2.27}$$

where $\mathbf{D}^{*}$ corresponds to a 3×3 diffusion tensor. With this method, authors showed that SNR is greatly increased and visual inspection is thus facilitated. They also demonstrated an important improvement of classification results after regularization but their method suffers from a high complexity in parameter tuning.

2.3.2 Spatial information at the classification decision stage

Among spectral-spatial classification strategies that use simultaneously both sources of information, three main approaches have proved to be effective, i.e., Kernel methods, Markov Random Fields and Cross-analysis.

Kernel methods

As previously explained, Kernel methods map the input data from the original space into a higher dimensional space in which data can be linearly separable. An interest of Kernel methods is that under certain mathematical conditions [Fauvel, 2007], Kernels of different types can be merged. Spatial Kernels adapted to the HS image have thus been computed and merged to spectral ones in [Camps-Valls et al., 2006] and Fauvel [2007]. Indeed, since any linear combination of Kernels is still a Kernel [Camps-Valls et al., 2006], a convenient composite Kernel that weighs the relative importance of spectral versus spatial information can be computed as:

$$K(x, y) = \mu K_{spectral}(x, y) + (1 - \mu)K_{spatial}(x, y) \tag{2.28}$$

Camps-Valls et al. [2006] used a Gaussian spectral Kernel based on the Euclidean distance and a spatial Kernel based on the mean and standard deviation on a small square window around each pixel. Similarly, a spectral Kernel was proposed in [Mercier and Lennon, 2003] where the authors demonstrated that a Kernel based on a spectral angle can outperform a standard Kernel that is based on the Euclidean distances. Later, Fauvel [2007] proposed a spatial Kernel with a non-fixed neighborhood that uses morphological operators. Recently, Li et al. [2013a] proposed a general framework for mixing multiple Kernels through Multinomial Logistic Regression (MLR) and Extended Morphological Profiles (EMAPs). This framework is less restrictive than those previously developed in the literature since it relaxes the constraint on convexity of Kernels.

Markov Random Fields

When using Markov Random Fields (MRF) in images, the underlying assumption is usually that two neighboring pixels are likely to have the same class label. In particular a Markov Field is a Random Field that only depends on a neighborhood [Bishop, 2007]. In practice, the neighborhood is constrained to the 4- or 8- closest neighbors as illustrated in Figure 2.8.

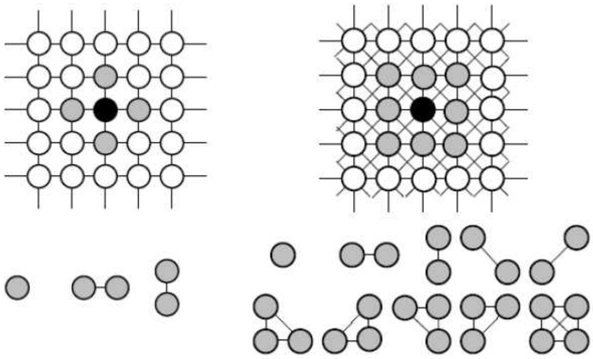


FIGURE 2.8: Illustration of the neighborhood using (left) a 4-connexity and (right) a 8-connexity. (from Gorretta [2009])

The assumed continuity of neighboring pixels is then exploited in a statistical sense and used for spatial modeling [Tarabalka et al., 2010b]. For instance, MRF can be used with a spectral classifier to encourage neighboring pixels to have the same label when using a probabilistic classifier [Bioucas-Dias et al., 2013, Li, 2011]. Note that another important use of MRF is to model textured classes as explained in [Rellier, 2002]. MRF despite its huge computational complexity is able to provide very good results in practice as demonstrated in [Bioucas-Dias et al., 2013] where in combination with a subspace MLR [Li et al., 2012] provided the best experimental results on a remote sensing data set.

Cross analysis

Recently, Gorretta et al. [2012] proposed a new framework for HS image segmentation using spectral and spatial information. In this framework, called *butterfly*, the analysis is done recursively by going ‘back and forth’ between the spectral and spatial representation of the data (see Figure 2.9). One of the interests of this approach comes from its flexibility since any dimension reduction (resp. segmentation) method can be plugged in the spectral (resp. spatial) analysis. Depending on the chosen methods, this framework even enables unsupervised segmentation of the HS image. Another advantage is that it leads to a more balanced use of these complementary sources of information as explained in [Gorretta, 2009].

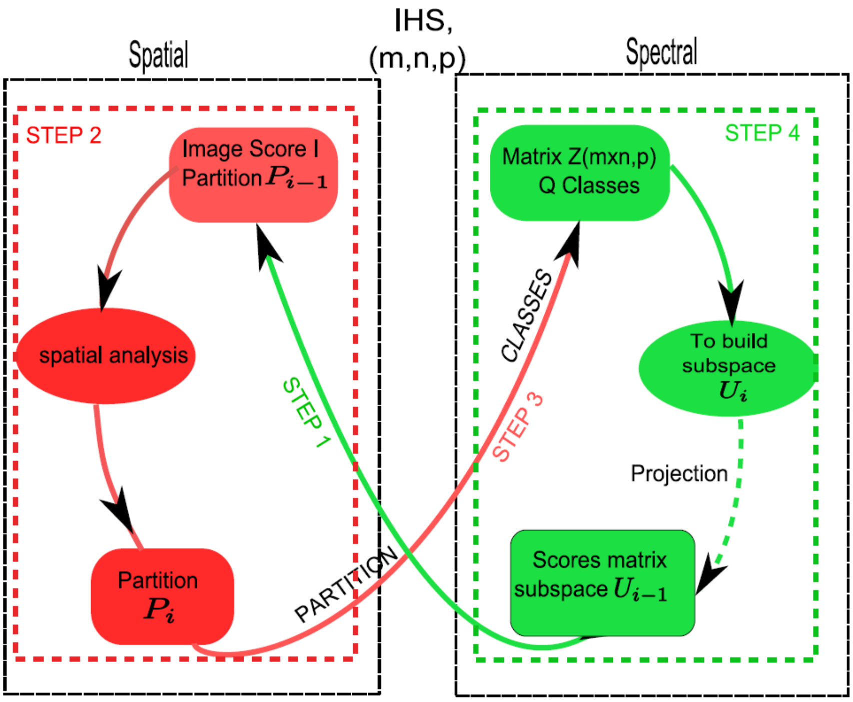


FIGURE 2.9: Framework of the butterfly approach proposed by Gorretta et al. [2012].

2.3.3 Spatial information as a post-processing stage

Using spatial information at a post-processing stage, in particular with a classification map, has received a lot of attention because of its simple implementation. Indeed, classification maps are single channel images and can thus be processed with any image processing technique. For classification, it usually results in a decrease of the *salt and pepper* aspect of the classification map. A simple approach consists of using classical morphological operators or median filtering to reduce this classification noise but more advanced approaches have been developed:

Segmentation

Tarabalka et al. [2010a] proposed to combine the output of a pixel-wise SVM classifier with a watershed segmented map. The combination was made using a majority of voting strategy between both maps as illustrated in figure 2.10.

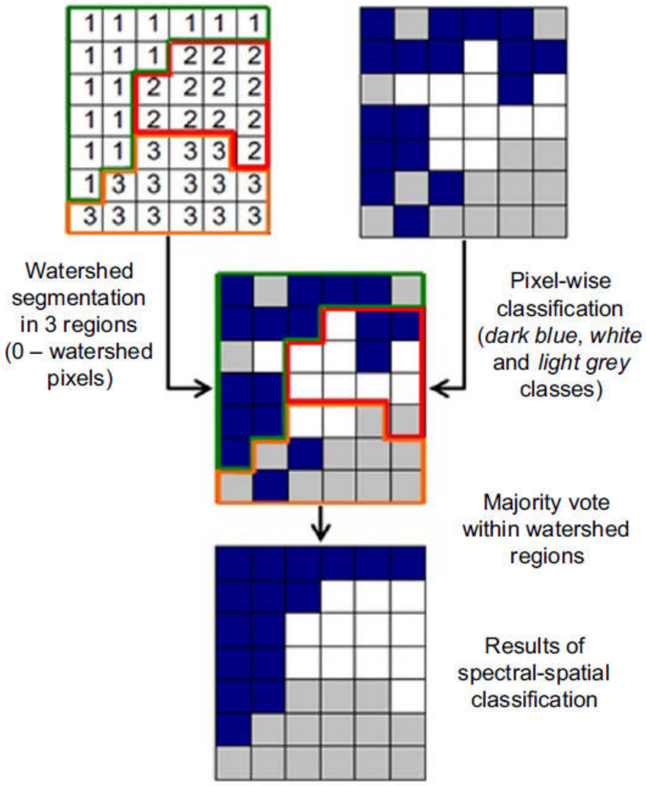

1 1 1 1 1 1
1 1 1 2 2 2
1 1 2 2 2 2
1 1 2 2 2 2
1 1 3 3 3 2
1 3 3 3 3 3
3 3 3 3 3 3
Watershed segmentation in 3 regions (0 – watershed pixels)
Pixel-wise classification (*dark blue*, *white* and *light grey* classes)
Majority vote within watershed regions
Results of spectral-spatial classification

FIGURE 2.10: Majority of voting between a pixel-wise classification and a segmented map. (modified from Tarabalka et al. [2010a])

Tarabalka [2007] then proposed a more advanced segmentation technique using a Hierarchical Segmentation approach (HSEG). Li [2011] proposed a segmentation of the probabilistic classification map they obtained from their multilevel logistic classification method.

Regularization

Edge Preserving Filtering (EPF) can also be used in order to regularize classification results at a post-processing stage. EPF has been implemented using for example Bilateral Filtering (BF) Tomasi and Manduchi [1998] or Anisotropic Regularization (AR) Perona and Malik [1990]. As explained before, EPF aims at regularizing gray-level or color images by smoothing spatially homogeneous regions while keeping their borders sharp. With AR, the regularization procedure is based on an image gradient computation whereas BF usually requires a guidance image.

In [Kang et al., 2014], authors have developed a regularization approach that uses BF in order to regularize a SVM classification map as illustrated in Figure 2.11. The guidance images needed for BF implementation were either a PCA score image or a RGB image reconstructed from the HS image. Their approach significantly decreased the classification noise, which proves the potential of EPF spatial regularization for such data in a classification context.

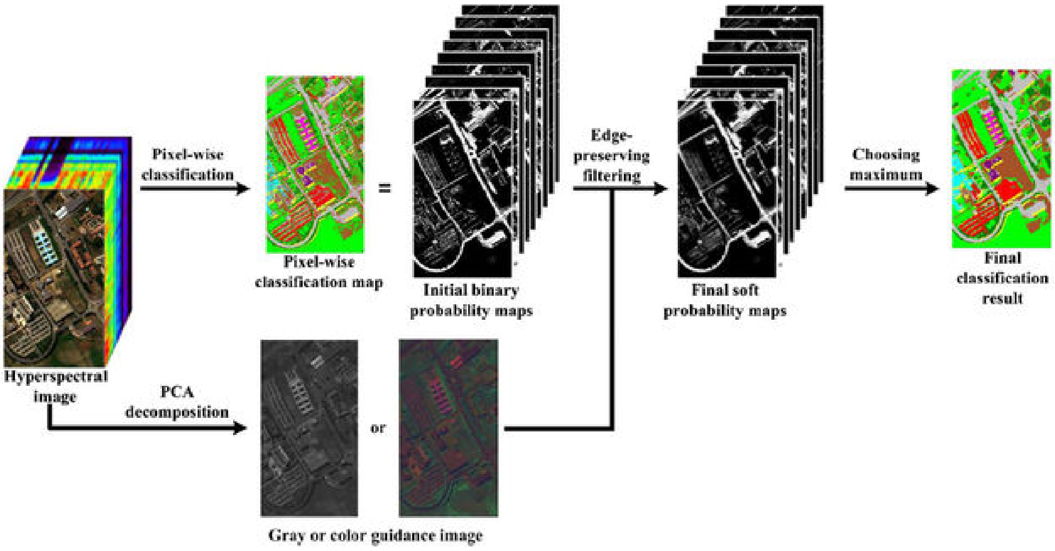

Pixel-wise classification
Pixel-wise classification map
=
Initial binary probability maps
Edge-preserving filtering
Final soft probability maps
Choosing maximum
Final classification result
Hyperspectral image
PCA decomposition
or
Gray or color guidance image

FIGURE 2.11: Spectral-spatial framework using EPF regularization and SVM pixel-wise classifier. (from Kang et al. [2014])

2.4 Reflectance correction

In the previous sections, HS images were assumed to be reflectance HS images, i.e., compensated/corrected from the incoming light. Depending on the circumstances with which the classification model is made, different scenarios can be seen [Hoffbeck and Landgrebe, 1994]:
(1) One HS image and one classification model: the correction is useful only for interpretation of spectral absorption bands but does not influence classification results.
(2) Different HS images and one model per image: similar to (1) and no interpretation of variation between images can be inferred.
(3) Different HS images, one model computed from spectra coming from one image: atmospheric correction is compulsory.

Reflectance correction is thus a prerequisite to most ‘real world’ HS data analysis and has therefore received a lot of attention over the years. The available methods are usually classified into physics-based (radiative transfer models), scene-based or image-based methods [Shaw and Burke, 2003]. Recent comprehensive and comparative reviews of the available methods can be found in [Gao et al., 2009, Griffin and Burke, 2003].

2.4.1 Background

When the solar light goes through the atmosphere its spectrum is changed because of absorption and scattering phenomena that are wavelength dependent. The light source seen by the target depends on atmospheric conditions, which creates problems for further spectral processing. The gases that are mainly responsible for spectral variations in the atmosphere and their spectral responses are represented in Figure 2.12. Shaw and Burke [2003] decomposes these atmospheric effects into four categories:
(1) Because of its composition, the atmosphere modulates the spectrum of the solar illumination before it reaches the ground (see Figure 2.13).
(2) A part of the solar radiation is scattered by the atmosphere directly in the field of view of the camera without even reaching the target (path-radiance).
(3) Shadowed objects receive the diffuse sky illumination that is different from the

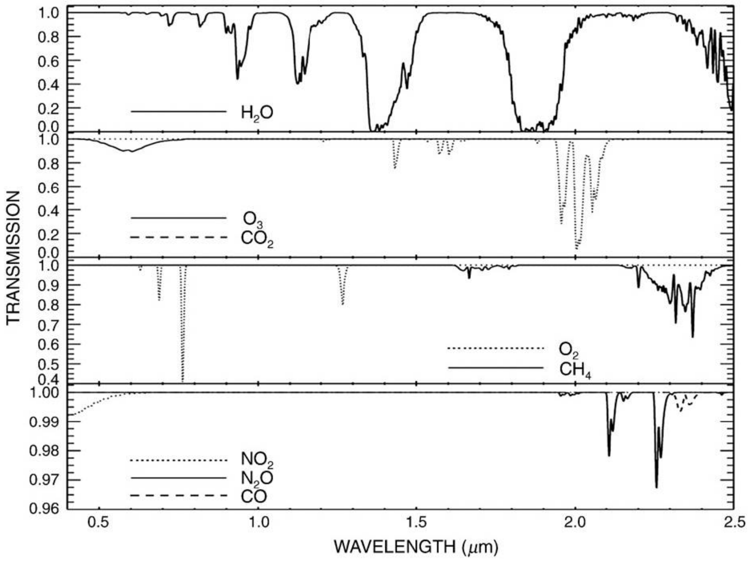


FIGURE 2.12: Simulated transmittance spectra of atmospheric water vapor, carbon dioxide, ozone, nitrous oxide, carbon monoxide, methane, oxygen, and nitrogen dioxide. (from Gao et al. [2009])

direct solar illumination.

(4) The light that leaves the target can still be absorbed by the atmosphere as it propagates toward the sensor thus changing its spectrum.

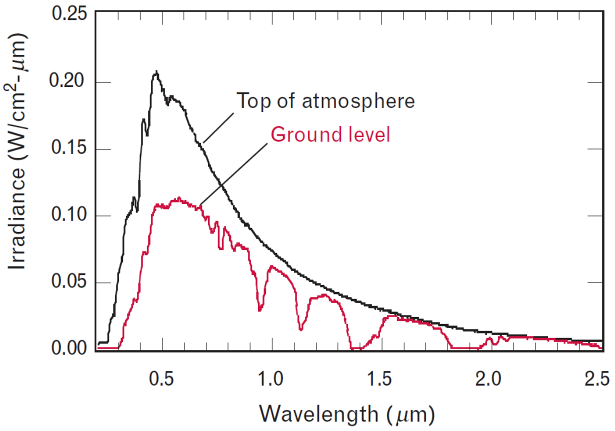


FIGURE 2.13: Difference between the solar spectral irradiance curves at the top of the atmosphere and at ground level. (from Shaw and Burke [2003])

Models

The general model/equation [Gao and Goetz, 1990, Hamm et al., 2012] from which all correction methods are based is given by:

$$L_{obs}(\lambda) = \Big(E_{\downarrow}(\lambda)T_{\downarrow}(\lambda)\cos\theta + L_{\downarrow}(\lambda)\Big)T_{\uparrow}(\lambda)\pi^{-1}\rho(\lambda) + L_{\uparrow}(\lambda), \tag{2.29}$$

where,

- $\rho(\lambda)$ = surface reflectance (what needs to be estimated)
- $L_{obs}(\lambda)$ =observed radiance at-sensor
- $L_{\uparrow}(\lambda)$ = Upwelling radiance along the target-sensor path (path-radiance) that is due to the atmosphere diffuse reflection toward the sensor
- $L_{\downarrow}(\lambda)$ = Downwelling irradiance (diffuse illumination)
- $E_{\downarrow}(\lambda)$ = Exo-atmospheric radiance onto the surface perpendicular to the incident beam
- θ = solar zenith angle relative to the surface
- $T_{\downarrow}(\lambda)$ = atmospheric transmission sun → target
- $T_{\uparrow}(\lambda)$ = atmospheric transmission target → sensor

This formulation can be written as a function of the available solar radiance on the target [Richter et al., 2002]:

$$L_{obs}(\lambda) = E(\lambda)T_{\uparrow}(\lambda)\pi^{-1}\rho(\lambda) + L_{\uparrow}(\lambda), \tag{2.30}$$

where $E(\lambda)$ corresponds to the available solar radiance on the scene:

$$E(\lambda) = E_{\downarrow}(\lambda)T_{\downarrow}(\lambda)\cos\theta + L_{\downarrow}(\lambda). \tag{2.31}$$

In both expressions, one can see that there is a linear relation between the observed radiance and the surface reflectance:

$$L_{obs}(\lambda) = a(\lambda)\rho(\lambda) + b(\lambda) \tag{2.32}$$

The goal of atmospheric correction methods is thus to give an accurate estimate of $a(\lambda)$ and $b(\lambda)$.

2.4.2 Physics-based transfer (model-based) correction

Radiative transfer based models simulate the solar irradiance spectrum, compute the scene radiance effects of solar position (using the acquisition date) and measure or estimate the amount of atmospheric absorption and scattering [Kruse, 2000]. The two most widely used corrections are ATmospheric REMOval (ATREM) [Gao et al., 1993] and Fast Line-of-Sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) [Adler-Golden et al., 1994, Cooley et al., 2002]. These correction methods have been compared in [Gao et al., 2009, Griffin and Burke, 2003, Kruse, 2000]. While a proper description of these methods is out of the scope of the review, the main steps are illustrated as a schematic flow chart in Figure 2.14a and Fig 2.14b.

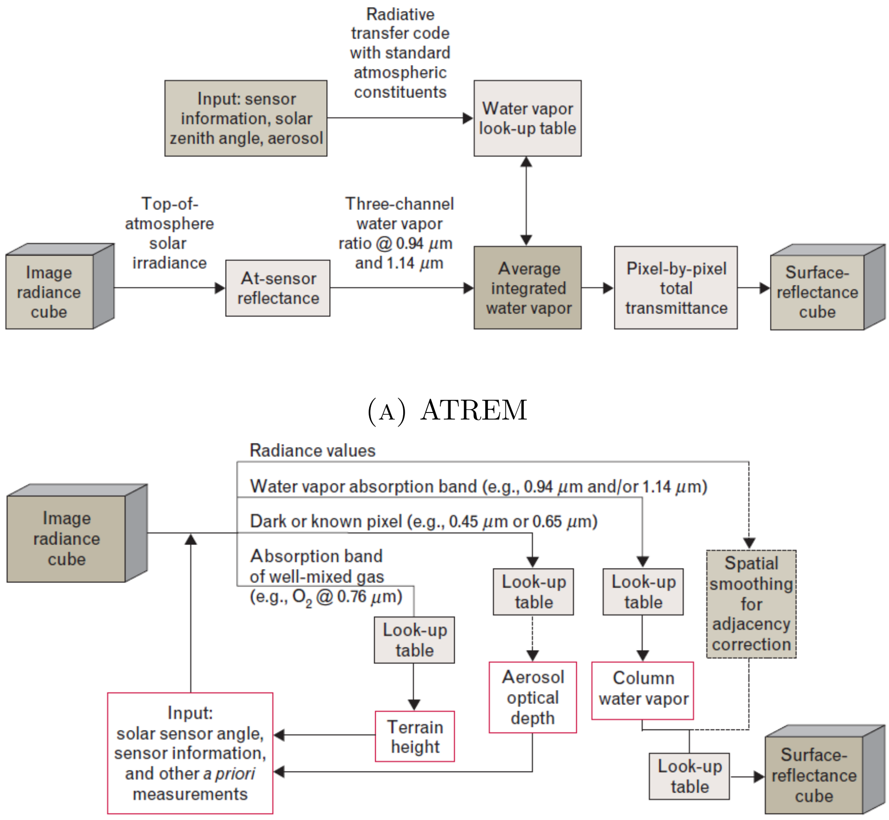
Both methods estimate mixed gases such as O_2, O_3 and CO_2 separately from water vapor. The former are indeed quite easy to estimate accurately while the latter is highly variable and more complex to estimate. The main difference between these methods is that ATREM does not model the influence of adjacent pixels scattering into the computation.

2.4.3 Scene-based correction

Scene-based correction methods, use extra sources of information in order to estimate empirically the additive and multiplicative terms of equation B.2.

The Empirical Line Method (ELM) is the simplest correction method to be used [Smith and Milton, 1999]. ELM consists in the estimation of $a(\lambda)$ and $b(\lambda)$ using classical linear regression between reflectance spectra measured in-field and the corresponding radiance spectra extracted from the HS image. The field reflectance spectra must be acquired on at least two surfaces that have a significantly different brightness. These surfaces also have to be homogeneous and large enough to cover at least a whole pixel in the HS data. These surfaces can be naturally present in the scene [Vain et al., 2009] (Low brightness surface are asphalt, tar or water and high brightness sand or concrete) or manually introduced into the field of view [Moran et al., 2001].

This correction has to be done for each wavelength and often requires a spectral resampling of the in-field spectrometer to match the HS sensor bands. An illustration of the ELM on three wavelengths of the HYDICE sensor is given in Figure 2.15.


Radiative transfer code with standard atmospheric constituents
Input: sensor information, solar zenith angle, aerosol
Water vapor look-up table
Top-of-atmosphere solar irradiance
Three-channel water vapor ratio @ 0.94 μm and 1.14 μm
Image radiance cube
At-sensor reflectance
Average integrated water vapor
Pixel-by-pixel total transmittance
Surface-reflectance cube
Radiance values
Water vapor absorption band (e.g., 0.94 μm and/or 1.14 μm)
Dark or known pixel (e.g., 0.45 μm or 0.65 μm)
Absorption band of well-mixed gas (e.g., O2 @ 0.76 μm)
Image radiance cube
Look-up table
Look-up table
Spatial smoothing for adjacency correction
Look-up table
Aerosol optical depth
Column water vapor
Input: solar sensor angle, sensor information, and other a priori measurements
Terrain height
Look-up table
Surface-reflectance cube

(A) ATREM

(B) FLAASH

FIGURE 2.14: Comparison of the two main physics-based atmospheric correction methods. (from [Griffin and Burke, 2003])

Irradiance Light Sensor correction is an interesting alternative which has been implemented in the CASI sensor [O'Neill et al., 2014]. It only requires that another sensor simultaneously records the solar light measurement Lennon et al. [2002]. The reflectance correction then simply corresponds to the ratio of the observed irradiance to the recorded sun light.

2.4.4 Image-based correction

Image-based correction methods only use information that can be retrieved from the image to perform the atmospheric correction. These corrections aim at estimating the available solar radiance on the scene $E(\lambda)$ (equation 2.30). With these corrections, the additive term $b(\lambda)$ is discarded and the atmospheric transmission variations are neglected because of their small influence at low altitude.

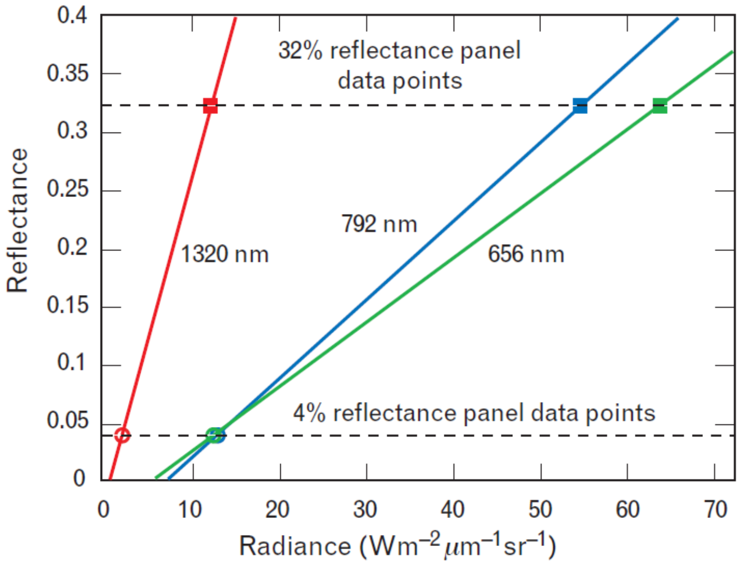


FIGURE 2.15: Illustration of the Empirical Line Method on three arbitrarily chosen wavelengths using low (4%) and high (32%) brightness reference surfaces. (from [Shaw and Burke, 2003])

Spectralon/ceramic correction

In the laboratory or in proximal detection a surface of known reflectivity is introduced into each image to perform the correction. The most commonly used surface is Spectralon (Lasphere, USA) which is a PTFE material that has a very flat and lambertian diffuse reflection [Geladi et al., 2004]. For more constrained environments, a calibrated surface of relatively flat reflectance can also be used to serve the same purpose. For instance, Vigneau [2010], Vigneau et al. [2011] calibrated a ceramic plate which is used for in-field HS image acquisition (see Figure 2.16). The correction thus become:

$$\rho(\lambda) = \frac{L_{obs}(\lambda)}{L_s(\lambda)}\rho_s(\lambda) \qquad (2.33)$$

where $\rho_s(\lambda)$ is known and $L_s(\lambda)$ is manually extracted from the image.

Flat field correction is the most widely used reflectance correction method in remote sensing. It requires that the image includes a uniform area that has a relatively flat spectral response. The spectral response also has to be relatively high so that the correction does not increase spectral noise. The mean spectrum of this area is then computed in order to increase the SNR. The entire scene is finally divided by this mean spectrum which leads to a ‘relative’ reflectance image. In practice, flat field reference is obtained with desert scenes, dry lake beds and human-made material such as concrete.

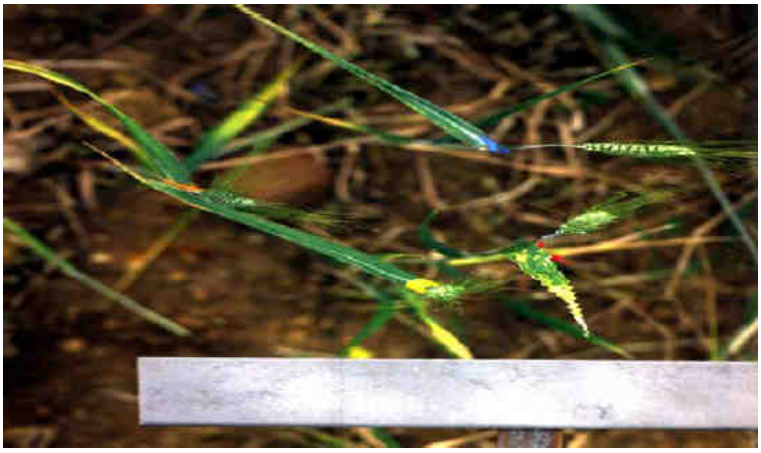


Figure 2.16: In-field reflectance correction using a calibrated ceramic plate. (from Vigneau et al. [2011])

Note that in the case of a specific absorption peak of the supposed flat surface, unexpected variation can happen at these wavelengths in the corrected image due to low signals.

Average Relative Reflectance correction divides each spectrum by the whole image mean spectrum. This method assumes that the scene possesses a lot of different materials in nearly constant proportions, so that the mean spectrum is quite stable from one image to another. In practice, this method is very difficult to apply because of a lack of such scenes and should be avoided in presence of vegetation spectra.

2.5 Conclusion

In this chapter, we have detailed some of the main approaches developed to tackle the issues with HS data classification.

For spectral dimension reduction, when the objective is classification, unsupervised methods lead to sub-optimal scores by not taking advantage of the class information during the reduction process. Among the supervised approaches, PLS-like methods tend to model the class structure of the data by maximizing the capture of covariance between the variables and the classes in order to build the reduced scores.
These approaches are thus naturally prone to overfitting and their parameters have to be tuned with a cross-validation procedure. When not carefully made, these

cross-validations can lead to over optimistic results and find a class structure when there is none.

On the other hand, Fisher-based methods tend to solve the within-class matrix inversion problem by using mathematical tricks such as pseudo inverse or inversion of the total scatter matrix instead of the within-class one. Another simple approach consists in using a PCA as a first step in order to obtain fewer variables on which a LDA can be performed. This is however sub-optimal, since the first step selects components that are not related to the class differences, and usually tend to select too many components for a given problem. Finally, Nullspace LDA is mathematically very promising by perfectly responding to the LDA paradigm. However, it requires that the nullspace of the within-class scatter matrix exists, which is only the case when the number of variables is greater than the number of observations. This means that in case of new observations acquired to enhance a model, NLDA cannot be used anymore and thus has few practical applications in hyperspectral classification. Also, in this nullspace, because the projections is orthogonal to all the within-class directions, the left-over information is very small and leads to noisy discriminant vectors.

In this thesis, we will propose an approach in which, contrary to NLDA, the removal of the within class variability is controlled and which also allows to preserve explicitly the most important discriminant axes.

Concerning spectral-spatial approaches, methods that use edge preserving filtering appear to be well adapted to HS image classification. In particular, being able to reduce the variability within classes by such filtering seems very interesting to complement the spectral within-class variability reduction. However, among the proposed approaches in the literature, this spatial regularization is only performed on either the original HS images or on score images obtained in an unsupervised way.

In both cases, the natural variability within each class can lead to very textured images. Thus, when using edge preserving filtering, edges are also preserved within the class that needs to be homogenized.

Therefore, we will propose in this thesis to use the EPF in a slightly different way: it is applied on a score image, which has been obtained from a supervised dimension reduction method. In fact, with the supervised method, within-class variability tends to be reduced and between class distance increased, which help the spatial regularization to find the edges only at class borders.

Finally, we have seen that in order to be independent from the light source and atmospheric conditions, radiance values in the HS images had to be transformed into reflectance values. This transformation requires the lighting to be known for each acquired image. As we have detailed, the light measurement on the scene can be performed by different means depending on the situation. However, this procedure often requires ground truth measurement, which is not always possible. In this thesis, we will propose an automatic procedure to compensate for the lighting conditions that is adapted to the classification paradigm. Providing that objects to discriminate are Lambertian, we show that, after a logarithm transformation, the difference in lighting corresponds to the same additive effect for all the pixels in the image. We then propose a method that estimates this translation even when there are missing classes.

Chapter 3

Proposed approaches

Contents

The objective of this chapter is to propose new approaches to deal with some hyperspectral classification issues. We thus propose a new supervised dimension reduction method that can handle the high dimensionality and high collinearity of spectral data and provides score images. We then propose an approach to combine spectral and spatial information through supervised spectral score images and spatial regularization before classification. Finally, using the supervised scores, we propose an approach that avoids reflectance correction through class densities registration. Although these approaches can be used separately to tackle specific HS classification issues, we also propose in this chapter a general classification framework that combines all of them.

3.1 Introduction

When using hyperspectral (HS) data for classification purposes, differences in spectral responses are used to assign a label to each pixel of the HS image. Then if the classification is supervised, training samples with known labels are required in order to create the classification model. However, specific issues are raised when a reliable classification model has to be created with such complex data. In this chapter, we present three approaches in order to tackle some of these main issues, i.e, spectral dimension reduction, spectral/spatial combination and light source variability correction.

The high dimensionality and collinearity of spectral data requires a dimension reduction to be performed beforehand. In this context, methods that mimic Fisher LDA have been proposed in the literature. The ones that tend to invert the within-class scatter matrix cannot be performed with spectral data because of collinearity issues with such data. Other approaches, mostly used for face-recognition applications, require that the nullspace is non-empty which limits it practicability as the number of available sample increases. We tackle this problem as well by proposing an original spectral dimension reduction method, called Dimension Reduction by Orthogonal Projection for Discrimination (DROP-D), that uses orthogonal projections. On the contrary to previous approaches, the method does not try to invert the within-class scatter matrix, but projects the data orthogonal to its main directions. DROP-D is supervised because it uses the class information in order to extract the reduced variables. Therefore, the obtained reduced variables (scores) tend to minimize the within-class scatter and to preserve the between-class scatter. One main advantage of DROP-D is that, by not attempting to model the class structure as is done with PLS-like approaches, overfitting can be prevented without the need for cross-validation.

The second issue is to define effective ways to use the spatial information in order to increase the classification performances obtained using only spectral information.
The core idea of the approach proposed in this thesis is to use a spatial regularization on score image channels obtained from a supervised dimension reduction method such as DROP-D or PLS. Because these channels are built to enhance differences between classes and to reduce the background variability, edges in the spatial domain correspond to actual class borders. Therefore, applying an edge-preserving spatial regularization on the channels of this score image reduces the remaining within-class variability due to the background and thus leads to an easier class decision.

The third issue is due to the dependency of radiance HS images with respect to lighting conditions. Thus, in order to be independent from the light source, HS radiance images first have to be transformed into reflectance images. In fact, only a classification model calibrated with a reflectance image can be used to classify other images. However, the classical reflectance correction technique implies that a surface of known reflectivity is introduced in the scene, and thus requires human intervention.
We propose an approach that avoids prior reflectance correction of HS radiance images before classification. Under the assumption that classes have Lambertian reflectance, we show that, after log-transformation, the difference in lighting corresponds to a translation in the spectral space. Then, using a linear dimension reduction, this translation in the spectral space corresponds to a translation in the score space. Due to the use of a supervised dimension reduction such as DROP-D or PLS, classes form clusters in the low dimensional score space. Using these clusters, we propose a method to estimate the translation that is robust against unbalanced number of samples and missing classes between images.

These three approaches have been designed independently, but can be combined in a complete classification framework. This framework is illustrated in Figure 3.1. In the following, each approach theory is detailed.

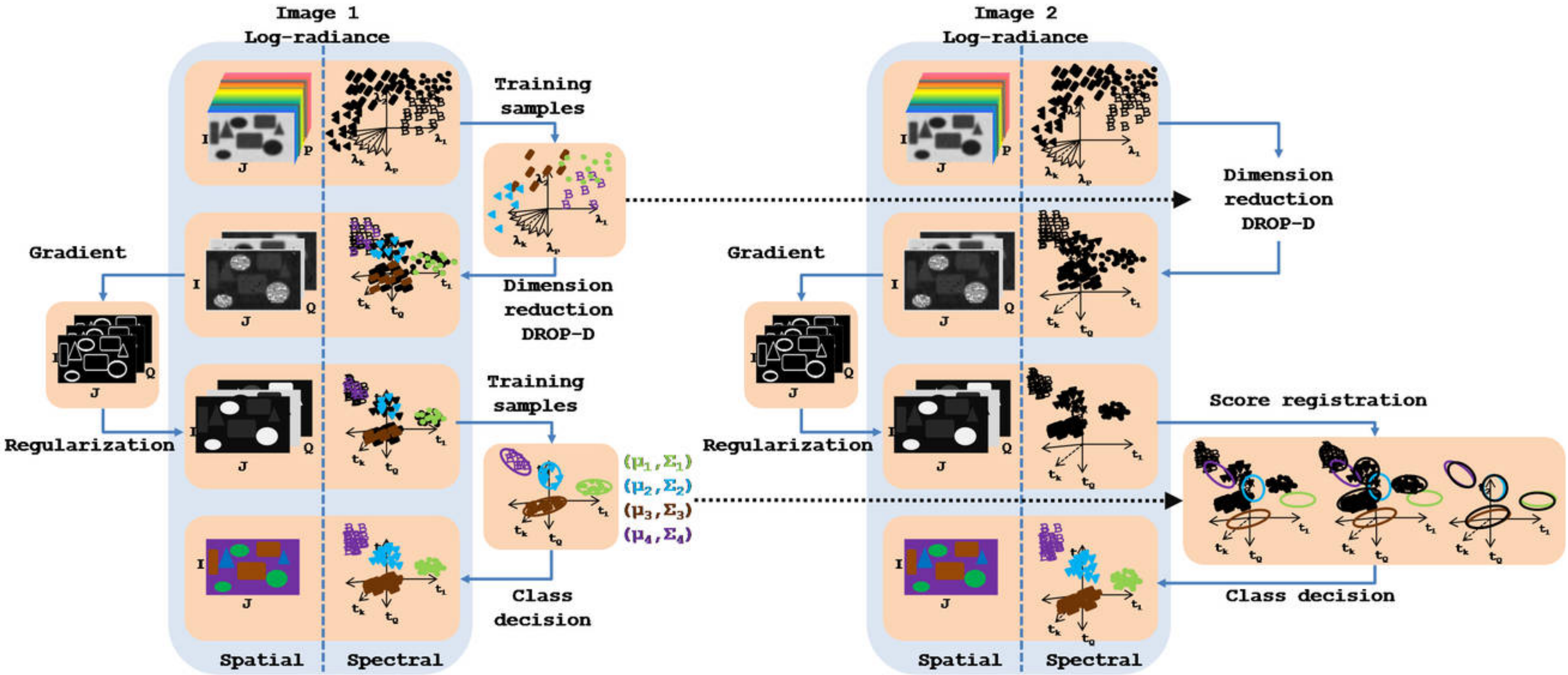


Figure 3.1: Illustration of the approaches presented in this thesis. (left) The spectral-spatial approach using first a supervised dimension reduction and then a spatial regularization. (left to right) The supervised model calibrated on the log-radiance image 1 applied to the log-radiance image 2. (right) The difference in radiance between images (translation in the log subspace) estimated by the score registration approach.

3.2 Dimension reduction

3.2.1 Prerequisites

In the following we call the *individual space*, the N-dimensional space (one axis per observation) in which we can represent the variables (wavelengths) as vectors. Conversely, the *variable space* is the P-dimensional space (one axis per variable) in which we can represent the observations as vectors.

Recall on orthogonal projection

For any column vector $\mathbf{v}$, and for any subspace defined by its basis $\mathbf{P}$, the orthogonal projection of $\mathbf{v}$ on $\mathbf{P}$ is given by

$$P^{\perp}_{\mathbf{P}}(\mathbf{v}) = \mathbf{P}\big(\mathbf{P}^{\intercal}\mathbf{P}\big)^{-1}\mathbf{P}^{\intercal}\mathbf{v}^{\intercal} \tag{3.1}$$

3.2.2 Subspace decomposition: problem statement

Supervised classification consists, using a data matrix $\mathbf{X}$ and a class matrix $\mathbf{Y}$ based on training samples, in finding a model that is capable of predicting the class of any observation $\mathbf{x}$ using its P descriptors. With spectral data, classification is often done in two steps:
(1) projection of the observation in a lower-dimensional subspace;
(2) affectation of the individual to a class.

The efficacy of the second step is highly influenced by the first one. Hence, we are looking for a subspace in which class centers are well separated and classes spread around their center are small. From a mathematical point of vue, it corresponds to finding the loadings $\mathbf{D}$ $\big(P \times Q\big)$ such that the projection of $\mathbf{X}$ on $\mathbf{D}$: (1) maximizes the between-class scatter given by

$$\mathbf{B}\big(\mathbf{XD},\mathbf{Y}\big) = \big(\mathbf{XD}\big)^{\intercal}\mathbf{Y}\big(\mathbf{Y}^{\intercal}\mathbf{Y}\big)^{-1}\mathbf{Y}^{\intercal}\big(\mathbf{XD}\big) \tag{3.2}$$

and (2) minimizes the within-class scatter given by

$$\mathbf{W}\big(\mathbf{XD},\mathbf{Y}\big) = \big(\mathbf{XD}\big)^{\intercal}\big(\mathbf{XD}\big) - \big(\mathbf{XD}\big)^{\intercal}\mathbf{Y}\big(\mathbf{Y}^{\intercal}\mathbf{Y}\big)^{-1}\mathbf{Y}^{\intercal}\big(\mathbf{XD}\big). \tag{3.3}$$

In addition, we are looking for a subspace of reduced dimensions, i.e., Q minimal. These three constraints on the way to build the set of axes $\min(Q)$, $\max(\mathbf{B})$ and $\min(\mathbf{W})$ are illustrated in Figure 3.2. The general approach consists in minimizing the ratio of within- to total-class scatter given by the Wilk's Lambda

$$\Lambda_{Wilks} = \frac{\mid \mathbf{W} \mid}{\mid \mathbf{T} \mid} = \frac{\mid \mathbf{W} \mid}{\mid \mathbf{B} + \mathbf{W} \mid}. \tag{3.4}$$

In cases where the data are well conditioned, a solution is given by the Fisher Linear Discriminant Analysis paradigm, which can be expressed as:

$$\mathbf{D} = \arg\max_{\mathbf{D}} \Big(\mathrm{trace}\big(\mathbf{D}^{\intercal}\mathbf{W}^{-1}\mathbf{B}\mathbf{D}\big)\Big) = \mathbf{E}_Q\big(\mathbf{W}^{-1}\mathbf{B}\big) \tag{3.5}$$

where for any diagonalizable square matrix $\mathbf{A}$, the notation $\mathbf{E}_Q\big(\mathbf{A}\big)$ corresponds to the Q eigenvectors associated to its Q largest eigenvalues. However, for ill-conditioned data, the inversion of $\mathbf{W}$ is problematic. Thus, LDA is known to be unable to deal with spectral data and several solutions have been proposed in the literature to overcome this problem (see Chapter 2).

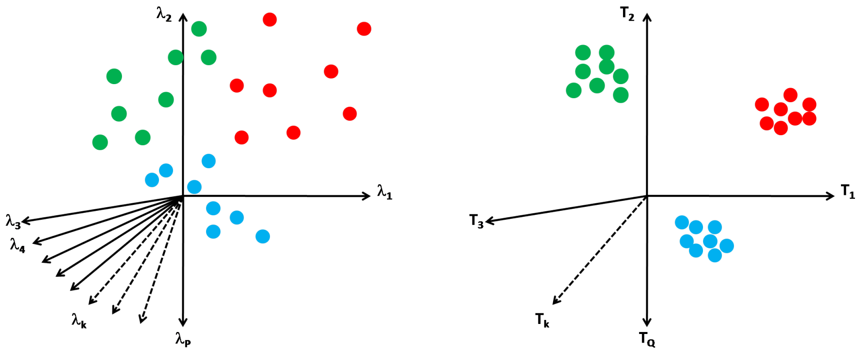
Nevertheless, the construction of a classification model corresponds to find a subspace of the variable space that 'copies' the class structure observed in the individual space of the sample set. The Fisher LDA does this by contracting the subspace carried by the within-class scatter and by focusing on the one carried by between-class scatter.

The method proposed in this thesis offers another way to realize this copy. This idea is to use the between- and within-class scatter to decompose the variable space into different subspaces so that one of them carries a large part of between-class scatter and a small part of within-class scatter.

3.2.3 Variability decomposition in $\mathbb{R}^N$ and $\mathbb{R}^P$

Suppose we have a matrix $\mathbf{X}$ containing N spectra of P variables from C classes coded using with dummy variables and stored in a matrix $\mathbf{Y}$. We can then define a mean per class using the matrix operation:

$$\mathbf{X}_B = \mathbf{Y}\big(\mathbf{Y}^{\intercal}\mathbf{Y}\big)^{-1}\mathbf{Y}^{\intercal}\mathbf{X}. \tag{3.6}$$



(A) $\mathbb{R}^p$ (B) $\mathbb{R}^q$

FIGURE 3.2: Dimension reduction for classification purposes, i.e., fewer axes ($Q \leq P$), a small within-class scatter and a large distance between class centroids

Thanks to the dummy variable coding, each of $\mathbf{X}_B$ row contains, instead of the original spectrum, the mean spectrum of its own class. The operation $\mathbf{X} \mapsto \mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}$ thus defines a new sample in which each observation is replaced by the mean of its class (centroid).

We can also define $\mathbf{X}_W$ as:

$$\begin{aligned}\mathbf{X}_W &= \mathbf{X} - \mathbf{X}_B \\ &= \left(\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\right)\mathbf{X}. \end{aligned} \tag{3.7}$$

The matrix $\mathbf{X}_W$ thus contains the observations centered on their class centroid.

What happens in the individual space ($\mathbb{R}^N$)?

In this space, we can represent the $\mathbf{Y}$ matrix, i.e., each column of $\mathbf{Y}$ corresponds to one vertex of the unit N-dimensional hypercube.

In this condition, the operation $\mathbf{X} \mapsto \mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X} = \mathbf{X}_B$ projects the columns of $\mathbf{X}$ (individuals) on the subspace defined by $\mathbf{Y}$. The removed part corresponds to $\mathbf{X}_W$ which is also an orthogonal projection, but on the orthogonal complement of $\mathbf{Y}$.

$$\mathbf{X}_B = \mathcal{P}_{\mathbf{Y}}(\mathbf{X}) = \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X} \tag{3.8}$$

$$\mathbf{X}_W = \mathcal{P}^{\perp}_{\mathbf{Y}}(\mathbf{X}) = \left(\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\right)\mathbf{X} \tag{3.9}$$

The subspaces spanned by these matrices, $\mathcal{E}_B = \mathcal{R}(\mathbf{X}_B)$ and $\mathcal{E}_W = \mathcal{R}(\mathbf{X}_W)$, are orthogonal and complementary subspaces of $\mathcal{E}_T$ in $\mathbb{R}^N$ (equation 3.10).

$$\mathcal{E}_T = \mathcal{E}_B \oplus \mathcal{E}_W \subseteq \mathbb{R}^N \tag{3.10}$$

In this space ($\mathbb{R}^N$), thanks to the orthogonality, we can thus completely eliminate $\mathbf{X}_W$ without affecting $\mathbf{X}_B$:

Proof. $\mathcal{P}^{\perp}_{\mathbf{X}_W}(\mathbf{X}) = \mathcal{P}^{\perp}_{\mathbf{I}_N - \mathbf{X}_B}(\mathbf{X}) = \mathcal{P}_{\mathbf{X}_B}(\mathbf{X})$ □

However, we are in the individual space, which means that this operation can be applied to vectors expressed as combinations of individuals and only modify the spectral values of the N observations of the training set, i.e., it is not applicable to any incoming spectrum.

What happens in the variable space ($\mathbb{R}^P$)?
In this space, the operation $\mathbf{X} \mapsto \mathbf{X}_B$ defined by a matrix $N \times N$ cannot be applied to a unique vector (spectrum). It is thus not a linear application.

However, in this space, we can use the subspaces spanned by $\mathbf{X}_B$ and $\mathbf{X}_W$. We can show that the between- and within-class scatter matrices define an orthogonal basis for these subspaces and can therefore be used to copy the class structure from $\mathbb{R}^N$ to $\mathbb{R}^P$. We thus have the following equations:

$$\mathbf{T}(\mathbf{X}_B) = \mathbf{B}(\mathbf{X}, \mathbf{Y}) \tag{3.11}$$

$$\mathbf{T}(\mathbf{X}_W) = \mathbf{W}(\mathbf{X}, \mathbf{Y}) \tag{3.12}$$

, i.e., the total scatter of $\mathbf{X}_B$ and $\mathbf{X}_W$ correspond to $\mathbf{B}$ and $\mathbf{W}$ respectively.

Proof.

$$\begin{aligned}\mathbf{T}(\mathbf{X}_B) &\triangleq \left(\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}\right)^\intercal\left(\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}\right)\\ &= \left(\mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\right)\left(\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X}\right)\\ &= \mathbf{X}^\intercal\mathbf{Y}\left(\mathbf{Y}^\intercal\mathbf{Y}\right)^{-1}\mathbf{Y}^\intercal\mathbf{X} \qquad (3.13)\\ &\triangleq \mathbf{B}(\mathbf{X},\mathbf{Y})\end{aligned}$$

$$\begin{aligned}\mathbf{T}(\mathbf{X}_W) &\triangleq \left((\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal)\mathbf{X}\right)^\intercal\left((\mathbf{I}_N - \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal)\mathbf{X}\right)\\ &= \left(\mathbf{X}^\intercal - \mathbf{X}^\intercal\mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{Y}\right)\left(\mathbf{X} - \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X}\right)\\ &= \mathbf{X}^\intercal\mathbf{X} - 2\mathbf{X}^\intercal\mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X} + \mathbf{X}^\intercal\mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X}\\ &= \mathbf{X}^\intercal\mathbf{X} - \mathbf{X}^\intercal\mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X} \qquad (3.14)\\ &\triangleq \mathbf{W}(\mathbf{X},\mathbf{Y})\end{aligned}$$

Noting that $\left((\mathbf{Y}^\intercal\mathbf{Y})^{-1}\right)^\intercal = (\mathbf{Y}^\intercal\mathbf{Y})^{-1}$ because $(\mathbf{Y}^\intercal\mathbf{Y})^{-1}$ is symmetric □

Hence, the subspace spanned by $\mathbf{X}_B$ is containing the between-class scatter while the subspace spanned by $\mathbf{X}_W$ is containing the within-class scatter.

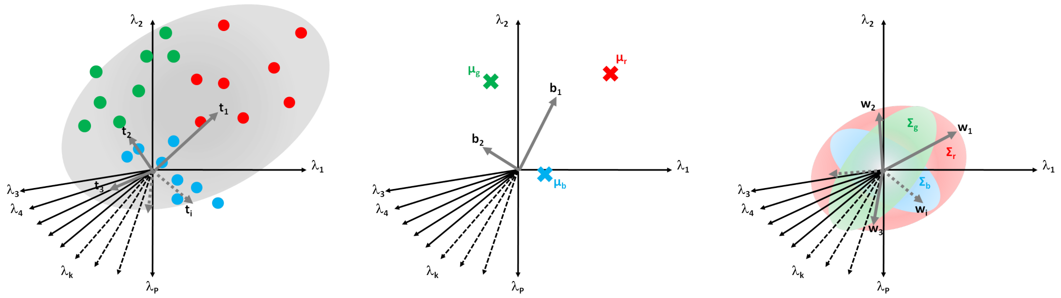
In the variable space $\mathbb{R}^P$, let us define these two subspaces $\mathcal{F}_B = \mathcal{R}(\mathbf{X}_B^\intercal)$ and $\mathcal{F}_W = \mathcal{R}(\mathbf{X}_W^\intercal)$. Using the range property $\mathcal{R}(\mathbf{A}^\intercal) = \mathcal{R}(\mathbf{A}^\intercal\mathbf{A})$, these subspaces are expressed as:

$$\begin{aligned}\mathcal{F}_T &= \mathcal{R}(\mathbf{X}^\intercal) = \mathcal{R}(\mathbf{X}^\intercal\mathbf{X}) = \mathcal{R}(\mathbf{T}) &(3.15)\\ \mathcal{F}_B &= \mathcal{R}(\mathbf{X}_B^\intercal) = \mathcal{R}(\mathbf{X}_B^\intercal\mathbf{X}_B) = \mathcal{R}(\mathbf{B}) &(3.16)\\ \mathcal{F}_W &= \mathcal{R}(\mathbf{X}_W^\intercal) = \mathcal{R}(\mathbf{X}_W^\intercal\mathbf{X}_W) = \mathcal{R}(\mathbf{W}) &(3.17)\\ & &(3.18)\end{aligned}$$

where we have by construction $\mathbf{T} = \mathbf{B} + \mathbf{W}$. Figure 3.3 illustrates this decomposition in the feature space:

- The total subspace $\mathcal{F}_\mathbf{T}$, whose dimension is bounded by $\dim\left(\mathcal{F}_\mathbf{T}\right) \leq \min(N,P)$ represents the overall data variability in the variable space without considering classes
- The between-class subspace $\mathcal{F}_\mathbf{B}$ is defined by the class centroids spread in the variable space. Its dimension is thus bounded by $\dim\left(\mathcal{F}_\mathbf{B}\right) \leq \min(C-1,P)$

- The within-class subspace $\mathcal{F}_{\mathbf{W}}$ corresponds to the overall spread of the data removed of class centroids. Its dimension is bounded by $\dim\left(\mathcal{F}_{\mathbf{W}}\right) \leq \min(N, P)$



(A) Total (B) Between-class (C) Within-class

FIGURE 3.3: Decomposition in the feature space $\mathbf{T} = \mathbf{B}+\mathbf{W}$. Note the possible collinearity between $\mathbf{b}_i$ and $\mathbf{w}_j$.

Then, because the subspace dimension and matrix rank are linked by the fundamental relation

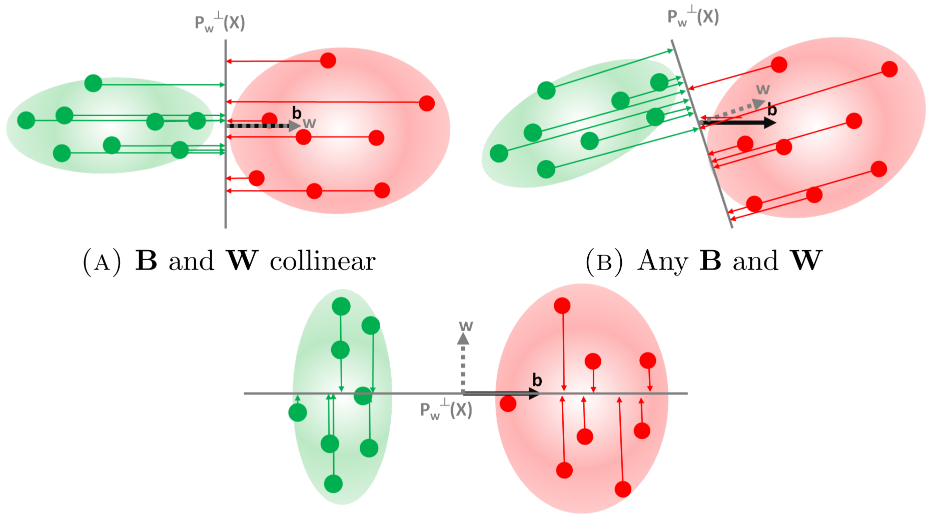
$$dim\ \mathcal{R}\left(\mathbf{A}\right) = rank\left(\mathbf{A}\right) = rank\left(\mathbf{A}^\intercal\right) = dim\ \mathcal{R}\left(\mathbf{A}^\intercal\right) \tag{3.19}$$

and since, $dim\ \mathcal{F}_{\mathbf{T}} = dim\ \mathcal{E}_{\mathbf{T}}$, $dim\ \mathcal{F}_{\mathbf{B}} = dim\ \mathcal{E}_{\mathbf{B}}$ and $dim\ \mathcal{F}_{\mathbf{W}} = dim\ \mathcal{E}_{\mathbf{W}}$, therefore $\mathcal{F}_{\mathbf{B}}$ and $\mathcal{F}_{\mathbf{W}}$ define two subspaces of $\mathcal{F}_{\mathbf{T}}$ in the variable space such that

$$\mathcal{F}_{\mathbf{T}} = \mathcal{F}_{\mathbf{B}} + \mathcal{F}_{\mathbf{W}} \subseteq \mathbb{R}^p \tag{3.20}$$

These subspaces $\mathcal{F}_{\mathbf{B}}$ and $\mathcal{F}_{\mathbf{W}}$ are however not orthogonal in $\mathbb{R}^P$ and their intersection is not necessarily empty. The separation of the between- and within-class scatter is therefore less obvious than in the individual space. Hence, depending on the class configuration, removing within-class variability does not necessarily improves the separability as illustrated with Figure 3.4.

In the following (section 3.2.4), we propose a method, called DROP-D, that enables a controlled removal of the within-class scatter, i.e., by preserving its axes collinear with $\mathcal{F}_{\mathbf{B}}$.



(A) **B** and **W** collinear

(B) Any **B** and **W**

(C) **B** and **W** orthogonal

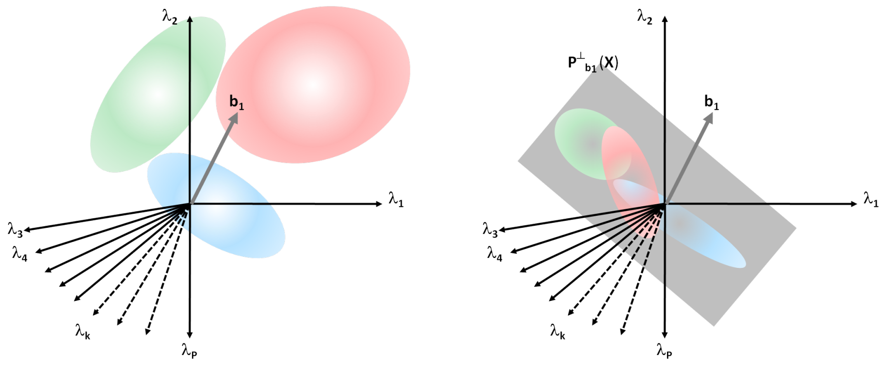
FIGURE 3.4: Effect of removing the within-class axis with different class configurations in $\mathbb{R}^P$

3.2.4 DROP-D

Dimension Reduction by Orthogonal Projection for Discrimination method (DROP-D) is in three steps.

The first step consists in removing from $\mathbf{X}$ the b principal directions of the between-class scatter, as expressed in equation 3.21.

$$\mathbf{X}_b^{\perp} = P_{\mathbf{B},b}^{\perp}(\mathbf{X}) \tag{3.21}$$



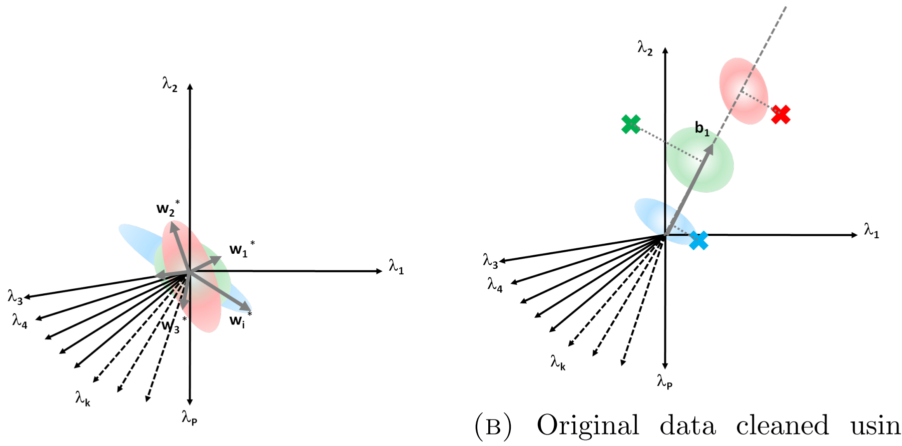
(A) Initial classes and the first between-class axis $\mathbf{b}_1$

(B) Initial classes projected orthogonal to $\mathbf{b}_1$

FIGURE 3.5: DROP-D first step

In a second step, the within-class scatter matrix is computed with $\left(\mathbf{X}_b^{\perp}, \mathbf{Y}\right)$. Then, the w principal directions linked to this within-class scatter ($\mathbf{W}^*$) are eliminated according to the equation 3.22.

$$\mathbf{X}_{clean} = P^{\perp}_{\mathbf{W}^*\left(\mathbf{X}_b^{\perp}, \mathbf{Y}\right), w}\left(\mathbf{X}\right) \tag{3.22}$$



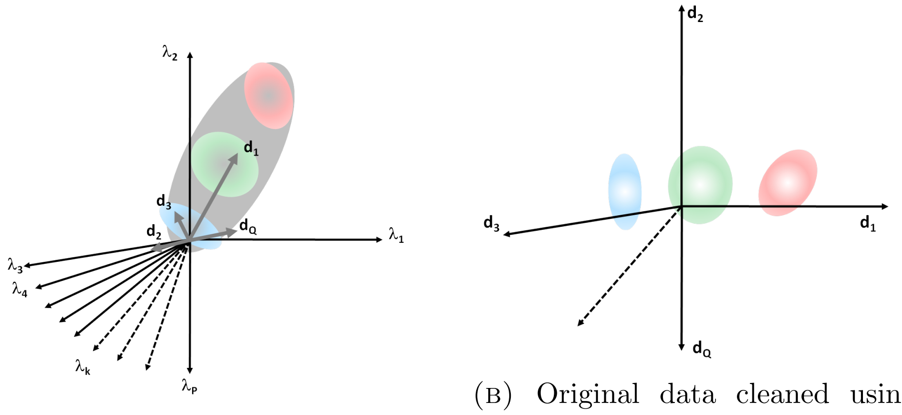
(A) Within-class scatter principal axes in the space orthogonal to $\mathbf{b}_1$

(B) Original data cleaned using $\mathbf{W}^*$. Note that the direction $\mathbf{b}_1$ is untouched.

FIGURE 3.6: DROP-D second step.

The third step is to extract the Q principal directions of $\mathbf{X}_{clean}$ which are given by:

$$\mathbf{D} = \mathbf{E}_Q\Big(\mathbf{T}\big(\mathbf{X}_{clean}\big)\Big). \tag{3.23}$$



(A) Within-class scatter principal axes in the space orthogonal to $\mathbf{b}_1$

(B) Original data cleaned using $\mathbf{W}^*$. Note that the direction $\mathbf{b}_1$ is untouched.

FIGURE 3.7: DROP-D third step.

To summarize, DROP-D defines three subspaces $\mathcal{F}_B$, $\mathcal{F}_{W^*}$ and $\mathcal{F}_D$ of $\mathbb{R}^P$, such that:

- $\mathcal{F}_B$ is linked to the b principal directions of the between-class scatter
- $\mathcal{F}_{W^*}$ contains the w principal directions of the within-class variance that are orthogonal to $\mathcal{F}_B$
- $\mathcal{F}_D$ contains the Q directions that include the b principal directions of the between-class scatter and the $Q - b$ principal directions that are orthogonal to the within-class scatter.

In doing so, DROP-D eliminates the principal directions of the within-class scatter while preserving the most important directions of the between-class scatter. A rough projection orthogonal to $\mathbf{W}$ would bring the risk of removing important axes of $\mathbf{B}$, because $\mathcal{F}_B$ and $\mathcal{F}_W$ can have a collinear part. In that sense, the step 1 of DROP-D guarantees to preserve at least the most important b axes of $\mathcal{F}_B$. In addition, axes of $\mathcal{F}_B$ that were not included in step 1 preservation, but that are orthogonal to $\mathcal{F}_W$, are preserved as well.

3.2.5 DROP-D algorithm

If the data is not already centered: $\mathbf{xm} \leftarrow mean(\mathbf{X})$ and $\mathbf{X} \leftarrow center(\mathbf{X}, \mathbf{xm})$.

DROP-D algorithm is as follows:

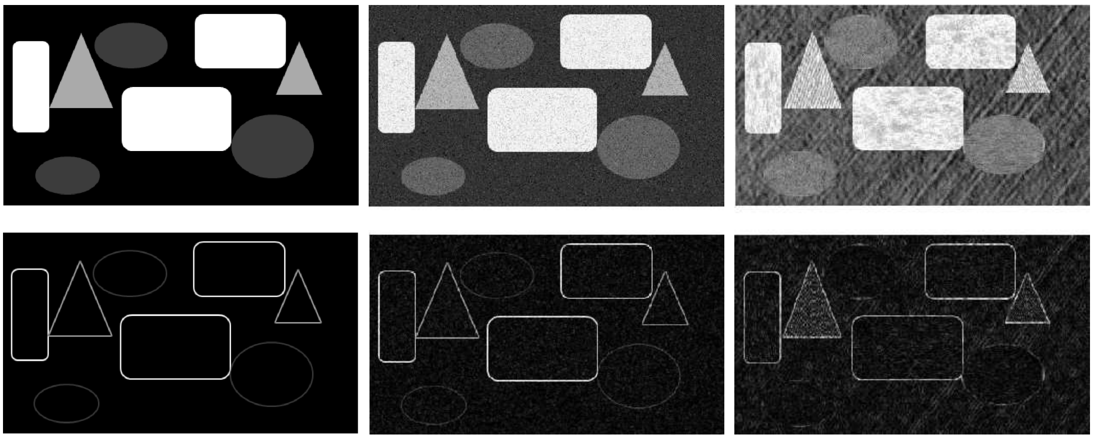
Algorithm 1: DROP-D

```
$\mathbf{B} \leftarrow \mathbf{X}^{\intercal}\mathbf{Y}(\mathbf{Y}^{\intercal}\mathbf{Y})^{-1}\mathbf{Y}^{\intercal}\mathbf{X}$ ; // Compute the between-class scatter
$\mathbf{B}_b \leftarrow \mathbf{E}_b(\mathbf{B})$ ; // Extract the $b$ principal eigenvectors of $\mathbf{B}$ (via SVD($\mathbf{B}$))
$\mathbf{X}_b^{\perp} \leftarrow \mathbf{X}\big(\mathbf{I}_P - \mathbf{B}_b(\mathbf{B}_b^{\intercal}\mathbf{B}_b)^{-1}\mathbf{B}_b^{\intercal}\big)$ ; // Remove from $\mathbf{X}$ these $b$ directions
$\mathbf{W}^{*} \leftarrow \mathbf{X}_b^{\perp\intercal}\mathbf{X}_b^{\perp} - \mathbf{X}_b^{\perp\intercal}\mathbf{Y}(\mathbf{Y}^{\intercal}\mathbf{Y})^{-1}\mathbf{Y}^{\intercal}\mathbf{X}_b^{\perp}$ ; // Compute the within-class scatter with $\mathbf{X}_b^{\perp\intercal}$ and $\mathbf{Y}$
$\mathbf{W}_w^{*} \leftarrow \mathbf{E}_w(\mathbf{W}^{*})$ ; // Extract the $w$ principal eigenvectors of $\mathbf{W}^{*}$ (via SVD($\mathbf{W}^{*}$)). These $w$ directions are assured to be at least orthogonal to the $b$ previously removed directions
$\mathbf{X}_{clean} \leftarrow \mathbf{X}\big(\mathbf{I}_P - \mathbf{W}_w^{*}(\mathbf{W}_w^{*\intercal}\mathbf{W}_w^{*})^{-1}\mathbf{W}_w^{*\intercal}\big)$ ; // Remove from THE ORIGINAL $\mathbf{X}$ these $w$ directions
$\mathbf{T}^{*} \leftarrow \mathbf{X}_{clean}^{\intercal}\mathbf{X}_{clean}$ ; // Compute the principal directions of $\mathbf{X}_{clean}$
$\mathbf{D} \leftarrow \mathbf{E}_Q(\mathbf{T})$ ; // Extract the $Q$ principal eigenvectors of $\mathbf{T}$ via SVD($\mathbf{T}$)
Optimize $b, w$ and $Q$;
```

Any new vector $\mathbf{x}$ is projected on this new basis by computing $\mathbf{s} = \big(\mathbf{x} - \mathbf{xm}\big)\mathbf{D}^{\intercal}$.

3.3 Spectral-spatial

In the previous chapter, we have seen several spectral-spatial approaches aiming at improving classification performances. Among them, Edge Preserving Filtering (EPF) proved its efficacy with different studies. In this thesis, we also propose a spectral-spatial approach that uses EPF spatial regularization in order to improve the pixel-wise classification results. The assumption made when using EPF spatial regularization to improve classification results is that edges are expected to be found only at the class borders and not within classes. However, with real images, edges are also found elsewhere than at class borders because of background variability caused by texture, non homogeneity of color and illumination within similar classes, etc. Therefore, applying EPF directly to a HS image preserves background edges and thus fails to reduce its variability. Applying it to a score image obtained by a non-supervised feature extraction method similarly fails because the extracted features also include the background as illustrated in Figure 3.8.



(A) Ideal (B) Ideal + noise (C) Ideal + noise + background

Figure 3.8: Single-channel images (top) and their gradients (bottom).

To overcome this issue, we thus propose an approach in which the spatial regularization is applied on a score image obtained by a supervised dimension reduction method (such as DROP-D). The core idea is that, since the score image already describes the classes to be discriminated by minimizing the background variability, edges mostly correspond to class borders and the spatial regularization process is more effective.

In the following, we assume that a linear supervised dimension reduction model, i.e., a matrix $\mathbf{D}$ of size $P \times Q$, is available.

3.3.1 Construction of the score image

Any hyperspectral image $\mathcal{H}$ of size $I \times J \times P$, i.e., I rows, J columns and P wavelengths, can always be unfolded into a data matrix $\mathbf{H}$ of size $M \times P$ where $M = I \cdot J$. The notation $\mathcal{H}_i$ then refers to the i^{th} channel of the HS image.

The reduced score image $\mathcal{S}$ of size $I \times J \times Q$ is similarly obtained by re-folding the scores matrix $\mathbf{S}$ of size $M \times Q$ given by:

$$\mathbf{S} = \mathbf{HD}. \tag{3.24}$$

Each channel $\mathcal{S}_i$ of the score image thus corresponds to the i^{th} score.

As DROP-D scores are computed from a PCA, both loadings and scores are orthogonal to one another. This way, the different channels of the score image are considered to be uncorrelated.

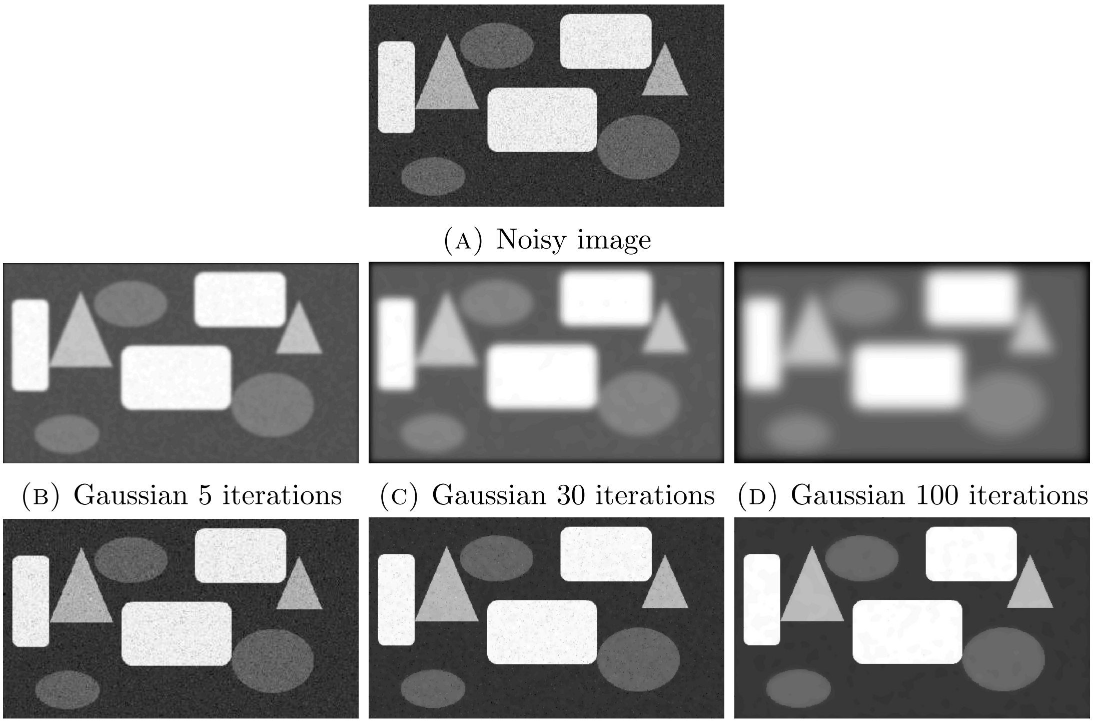
3.3.2 Anisotropic regularization

We implement our approach using the anisotropic diffusion method from Perona and Malik [1990] to enhance the within region homogeneity while keeping intact the borders between adjacent regions. This method was developed for de-noising gray-scale images by smoothing the image without removing the main edges.

The Perona and Malik [1990] method is an iterative process in which, at each iteration, the amount of smoothing is weighted by the intensity of the local gradient value. Considering a single channel image $\mathcal{I}$ (supposed continuous), the evolution equation

$$\frac{\partial \mathcal{I}(x, y, t)}{\partial t} = div\big(\nabla \mathcal{I}(x, y, t)\big) = \triangle\ I(x, y, t) \tag{3.25}$$

corresponds to the heat equation, where div and ∇ are respectively the divergence operator and the gradient operator with respect to the space variables, and where t is the time used to define the evolution of the diffusion process.



(A) Noisy image

(B) Gaussian 5 iterations (C) Gaussian 30 iterations (D) Gaussian 100 iterations

(E) AR 5 iterations (F) AR 30 iterations (G) AR 100 iterations

FIGURE 3.9: Effect of the regularization of a noisy image (top) using: classical Gaussian filter (middle) and anisotropic regularization (AR) (bottom).

The solution of this equation corresponds to a temporal Gaussian filtering, whose variance is $\sigma^2 = 2t$, given by

$$I(x,y,t) = I(x,y,t_0) * G(x,y,t) \text{ where } G(x,y,t) = \frac{1}{4\pi t} \exp\Big(-\frac{x^2+y^2}{4t}\Big) \quad (3.26)$$

However, this Gaussian filtering can reduce noise in images, but it operates in an identical way in every direction and thus does not preserve the image discontinuities due to object transitions (see Figure 3.9b to 3.9d).

In order to provide sharp edges while smoothing within regions, Perona and Malik [1990] proposed to modulate the gradient as following:

$$\frac{\partial \mathcal{I}(x,y,t)}{\partial t} = div\left[g(\parallel \nabla \mathcal{I}(x,y,t) \parallel)\nabla \mathcal{I}(x,y,t)\right] \quad (3.27)$$

where the function g has to be decreasing with respect to the local image gradient norm $\alpha = \parallel \nabla \mathcal{I} \parallel$. In [Perona and Malik, 1990], the authors have used a Gaussian function only determined by one parameter, which corresponds to a smoothing kernel width η. This function is given by:

$$g(\alpha) = \exp\big(- \big(\alpha/\eta\big)^2\big). \tag{3.28}$$

This diffusion process is then anisotropic which allows the conservation of main edges as represented in Figure 3.9e to 3.9g.

3.3.3 Score image regularization

Depending on the obtained scores, two regularization schemes can be considered: With non-orthogonal scores, a multidimensional regularization scheme is preferred to avoid outliers as described in chapter 2.
With orthogonal scores, we can find homogeneous regions on one score while there is a class transition on another one.

In this latter case, each channel of the score image can be processed independently, leading to a very simple and parallel method. The diffusion process is thus applied individually to every channel $\mathcal{S}_i$ of the score image $i \in [\![1, \cdots, Q]\!]$. The process is initialized as $\mathcal{S}_{i,0} = \mathcal{S}_i$. Then, at the iteration $k+1$, the diffusion process numerically applied to the channel $\mathcal{S}_{i,k}$ is written as:

$$\mathcal{S}_{i,k+1} = \mathcal{S}_{i,k} + \epsilon \cdot div\left[g\big(\parallel \nabla\mathcal{S}_{i,k} \parallel \big) \cdot \nabla\mathcal{S}_{i,k}\right] \tag{3.29}$$

where ϵ tunes the amount of change at each iteration of the diffusion process.

3.4 Reflectance correction

In this section we propose an automatic method for reflectance correction which overcomes the use of a reference measurement in the case of classification.

3.4.1 Hypotheses

As a first hypothesis, we consider that a radiometric correction has been applied to all the images, i.e., the HS images provide radiance spectra. Also, the materials to be discriminated are supposed to be Lambertian. Then, the calibration set has to be representative of potential material in other images, i.e. every possible class has to be represented and class variability has to be large enough to include every potential future outcome. A ground truth is supposed to be available for the calibration image as the model is supervised.

3.4.2 Lambertian hypothesis

A hyperspectral image is a measure of the radiation emitted or reflected from a scene in a large number of contiguous spectral bands. The quantity measured by an hyperspectral sensor is, after radiometric correction, a spectral radiance $L(\lambda)$, i.e., an irradiance measured in a specific direction (in $W.sr^{-1}.m^{-2}.nm^{-1}$). Note that in the following we only consider the case of Lambertian surfaces which means that the reflectance is independent from both the angle between the camera and the source and from the light incidence angle. With Lambertian materials, for a pixel i, j, the measured radiance is:

$$L_{i,j}(\lambda) = r_{i,j}(\lambda)E(\lambda) \tag{3.30}$$

where $r_{i,j}(\lambda)$ is the reflectance in radiance (also called remote sensing reflectance) and $E(\lambda)$ is the descending irradiance (in $W.m^{-2}.nm^{-1}$) supposed identical for each pixel.

The usual reflectance correction method to retrieve $r_{i,j}(\lambda)$ from $L_{i,j}(\lambda)$ consists in estimating $E(\lambda)$. Then,

$$r_{i,j}(\lambda) \simeq \hat{r}_{i,j}(\lambda) = L_{i,j}(\lambda)\frac{E(\lambda)}{\hat{E}(\lambda)} \tag{3.31}$$

where $\hat{E}(\lambda)$ is either directly measured or retrieved using a reference surface in each image by $\hat{E}(\lambda) = \dfrac{L_{ref}(\lambda)}{r_{ref}(\lambda)}$.

The reference surface is usually a Spectralon ® $(r_{ref}(\lambda) \simeq 1)$ or a calibrated surface of known reflectivity. $L_{ref}(\lambda)$ can then be (manually) extracted in each image and the correction computed on every pixel.

3.4.3 Discrimination model hypothesis

Let us consider a matrix $\mathbf{X}$ of dimensions $(N \times P)$ that corresponds to N spectra of P wavelengths extracted from an hyperspectral image. Let us also define a matrix $\mathbf{Y}$ of dimension $(N \times C)$ that codes the class belonging for each spectra of $\mathbf{X}$.

Let us consider a classification method that first computes a set of reduced variables (such as DROP-D), and then use a discrimination rule on the obtained scores. Recalling that projections from high- to low-dimensional spaces tend to increase Normality of the distribution[1], in the following, we use the Bayes classifier on the reduced scores. We thus have the following discrimination model:

(1) The linear dimension reduction model $\mathbf{D}$ decomposes the spectral matrix into a score matrix $\mathbf{S}$:

$$\mathbf{S} = \mathbf{X}\mathbf{D} \tag{3.32}$$

of dimension $N \times Q$, with $Q \ll P$.

(2) For each class $c \in [1, \cdots, C]$ an estimate of the class mean $\hat{\boldsymbol{\mu}}_c$ and class covariance matrix $\widehat{\boldsymbol{\Sigma}}_c$ is computed using the samples available in the training set.

(3) The class decision for a new observation (spectrum $\mathbf{x}$) $\mathbf{s} = \mathbf{x}\mathbf{D}^{\mathsf{T}}$ is made using Bayes classifier defined as:

$$\hat{c} = \arg\max_{c} \frac{1}{(2\pi)^{\frac{Q}{2}} |\widehat{\boldsymbol{\Sigma}}_c|^{\frac{1}{2}}} \exp^{\left(\frac{1}{2}\left(\boldsymbol{s} - \hat{\boldsymbol{\mu}}_c\right)^{\mathsf{T}} \widehat{\boldsymbol{\Sigma}}_c^{-1} \left(\boldsymbol{s} - \hat{\boldsymbol{\mu}}_c\right)\right)} \tag{3.33}$$

[1] (see Section 1.3.1)

3.4.4 Problem statement

We have seen that reflectance and radiance are linked by a multiplicative term, constant for each pixel in a given image, $E(\lambda)$. The bulk of our method is to transform equation 3.30 using the logarithm:

$$log(L_{i,j}(\lambda)) = log\big(r_{i,j}(\lambda)E(\lambda)\big) = log\big(r_{i,j}(\lambda)\big) + log\big(E(\lambda)\big). \tag{3.34}$$

In the following of this section, we analyze which differences occur in the model when using the log-reflectance and the log-radiance matrix to build the model.

Consider a matrix $\mathbf{X}^{(R)}$ containing log-reflectance spectra and the matrix $\mathbf{X}^{(L)}$ containing the corresponding, but uncorrected, log-radiance spectra. In both cases, N spectra of P wavelengths have known labels and can be use to calibrate a supervised classification model.

Effect on the dimension reduction model

Let us consider the dimension reduction models $\mathbf{D}^{(L)}$ and $\mathbf{D}^{(R)}$ calibrated using $\mathbf{X}^{(L)}$ and $\mathbf{X}^{(R)}$ respectively. Under the hypothesis of similar lighting at each pixel, the mean centered spectral matrices $\widetilde{\mathbf{X}^{(L)}}$ and $\widetilde{\mathbf{X}^{(R)}}$ are the same:

Proof.

$$\begin{aligned}
\widetilde{\mathbf{X}^{(L)}} &= \mathbf{X}^{(L)} - \mathbf{1}_N \widehat{\boldsymbol{\mu}_{\mathbf{X}^{(L)}}}^{\intercal} \\
&= \left(\mathbf{X}^{(R)} + \mathbf{X}^{(E)}\right) - \left(\mathbf{1}_N \widehat{\boldsymbol{\mu}_{\mathbf{X}^{(R)}}}^{\intercal} + \mathbf{1}_N \widehat{\boldsymbol{\mu}_{\mathbf{X}^{(E)}}}^{\intercal}\right) \qquad (3.35) \\
&= \widetilde{\mathbf{X}^{(R)}} \qquad (3.36)
\end{aligned}$$

The last part of this equality holds because the lighting is supposed identical for each pixels and thus $\mathbf{X}^{(E)} = \mathbf{1}_N \widehat{\boldsymbol{\mu}_{\mathbf{X}^{(E)}}}^{\intercal}$. □

Because the first step of (most) dimension reduction methods is to center these matrices ($\mathbf{X}^{(L)}$ and $\mathbf{X}^{(R)}$), the models obtained from either a centered log-radiance image or a centered log-reflectance image are the same: $\mathbf{D}^{(L)} = \mathbf{D}^{(R)} = \mathbf{D}$

Effect of the reflectance correction on the reduced scores

The scores obtained from the log-radiance spectra matrix $\mathbf{X}^{(L)}$ using $\mathbf{D}$ can thus be decomposed as:

$$\begin{aligned}\mathbf{S}^{(L)} &= \mathbf{X}^{(L)}\mathbf{D}\\ &= \big(\mathbf{X}^{(R)} + \mathbf{X}^{(E)}\big)\mathbf{D}\\ &= \mathbf{X}^{(R)}\mathbf{D} + \mathbf{X}^{(E)}\mathbf{D}\\ &= \mathbf{S}^{(R)} + \mathbf{S}^{(E)}\\ &= \mathbf{S}^{(R)} + \mathbf{1}_N \widehat{\boldsymbol{\mu}_{\mathbf{S}^{(E)}}}^{\intercal} \qquad (3.37)\end{aligned}$$

which means that using a log-radiance image, the scores obtained through a linear model are only translated versions of the ones that are obtained with a log-reflectance image.

Effect on the classification decision

The score translation does not change the class separability but has a direct effect on the classifier in terms of class decision. For instance, in the following we show that from the class parameters estimated from the training samples, only the class mean is changed.

Let us define $\hat{\boldsymbol{\mu}}_c$ and $\widehat{\boldsymbol{\Sigma}}_c$ the mean vector and covariance matrix estimated using the training samples for class c, where $c \in [1, \cdots, C]$. Then, the class decision is computed as:

$$\begin{aligned}\widehat{\boldsymbol{\mu}_c^{(L)}} &\triangleq \frac{1}{N_c}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \mathbf{S}_i^{(L)}\\ &= \frac{1}{N_c}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \mathbf{S}_i^{(R)} + \frac{1}{N_c}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \mathbf{S}_i^{(E)}\\ &\triangleq \widehat{\boldsymbol{\mu}_c^{(R)}} + \widehat{\boldsymbol{\mu}^{(E)}} \qquad (3.38)\end{aligned}$$

and,

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}_c^{(L)}} &\triangleq \frac{1}{N_c - 1}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \big(\mathbf{S}_i^{(L)} - \widehat{\boldsymbol{\mu}_c^{(L)}}\big)\big(\mathbf{S}_i^{(L)} - \widehat{\boldsymbol{\mu}_c^{(L)}}\big)^{\intercal}\\ &= \frac{1}{N_c - 1}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \big(\mathbf{S}_i^{(R)} + \mathbf{S}_i^{(E)} - \widehat{\boldsymbol{\mu}_c^{(R)}} - \widehat{\boldsymbol{\mu}^{(E)}}\big)\big(\mathbf{S}_i^{(R)} + \mathbf{S}_i^{(E)} - \widehat{\boldsymbol{\mu}_c^{(R)}} - \widehat{\boldsymbol{\mu}^{(E)}}\big)^{\intercal}\\ &= \frac{1}{N_c - 1}\sum_{\mathbf{S}_i \in \mathcal{Y}_c} \big(\mathbf{S}_i^{(R)} - \widehat{\boldsymbol{\mu}_c^{(R)}}\big)\big(\mathbf{S}_i^{(R)} - \widehat{\boldsymbol{\mu}_c^{(R)}}\big)^{\intercal}\\ &\triangleq \widehat{\boldsymbol{\Sigma}_c^{(R)}} \qquad (3.39)\end{aligned}$$

Which proves that the only parameters that change in the Bayes classifier are the class mean vectors. Therefore, from a classification point of view, the reflectance correction can be avoided if we know the translation $\widehat{\boldsymbol{\mu}^{(E)}}$ and by changing the decision rule considering $\hat{\boldsymbol{\mu}}_c + \widehat{\boldsymbol{\mu}^{(E)}}$ instead of $\hat{\boldsymbol{\mu}}_c$.

By simple extension, represented in Figure 3.10, this approach can be applied to transfer the decision rule between two images that are corrected in reflectance or not.

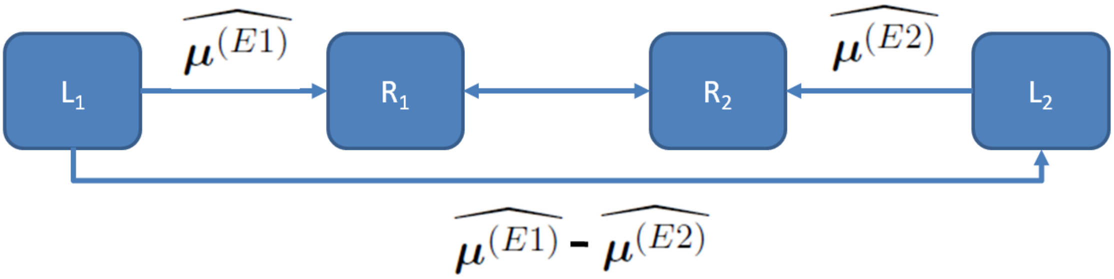

$\widehat{\boldsymbol{\mu}^{(E1)}}$
$\widehat{\boldsymbol{\mu}^{(E2)}}$
L_1
R_1
R_2
L_2
$\widehat{\boldsymbol{\mu}^{(E1)}} - \widehat{\boldsymbol{\mu}^{(E2)}}$

FIGURE 3.10: Translation estimation scheme between two radiance images L_1 and L_2.

3.4.5 Translation estimation

In this section we present a method to retrieve the translation parameter that is robust against missing classes and number of samples in each class. A graphical framework of the method is given in Figure 3.11.

In the general case, the class information is known in the one image used for training the model and unknown for the other images. While the problem of finding a translation is obvious when the class information is known in both sets, it becomes more complex when it is not.

This problem is known as *registration* and several methods have been developed in the field of image processing [Zitová and Flusser, 2003]. When the transformation is simple such as a rigid transformation, a convenient tool is the normalized cross-correlation. Cross-correlation uses the grey level information between two images as a measure of matching. One of the images is transformed (translated, rotated, ...) until the best matching is found.

In our score registration case, things are quite different because:
1) we are in a Q-dimensional space,
2) we do not have pixels and thus no grey-level to assess the matching.

Hence, before registration, a Q-dimensional image is created using the Q-dimensional scores. The image creation step is illustrated in Figure 3.11 using labeled training data, unlabeled data and unlabeled data with a missing class. It is based on the estimation of Gaussian distributions of classes in the Q space for every HS image.

In the following, the subscript 1 is used for the training data with a known class label and the subscript 2 for the unknown data.

In the first image, from which the labeled samples are known, the estimation of class parameters (which characterize their Gaussian distributions)

$$\hat{\boldsymbol{\theta}}_1 = \{\hat{\boldsymbol{\mu}}_{11}, \hat{\boldsymbol{\mu}}_{12} \cdots, \hat{\boldsymbol{\mu}}_{1C}, \widehat{\boldsymbol{\Sigma}}_{11}, \widehat{\boldsymbol{\Sigma}}_{12}, \cdots, \widehat{\boldsymbol{\Sigma}}_{1C}\} \tag{3.40}$$

is straightforward. In the other images, the probability density function (pdf) of each class has to be estimated in the score space without the knowledge of the class. In the case of Normally distributed classes, as assumed here, a powerful tool is the Expectation Maximization (EM) algorithm [Moon, 1996].

Using the EM algorithm, the parameters of the estimated Gaussian mixture are in our case[2] noted

$$\hat{\boldsymbol{\theta}}_2 = \{\hat{\boldsymbol{\mu}}_{21'}, \hat{\boldsymbol{\mu}}_{22'} \cdots, \hat{\boldsymbol{\mu}}_{2C'}, \widehat{\boldsymbol{\Sigma}}_{21'}, \widehat{\boldsymbol{\Sigma}}_{22'}, \cdots, \widehat{\boldsymbol{\Sigma}}_{2C'}\} \tag{3.41}$$

The problem is that there is no direct correspondence between the class indices $c \in \{1, \cdots, C\}$ in the training set $\hat{\boldsymbol{\theta}}_1$ and the indices $c' \in \{1, \cdots, C'\}$ issued from EM in $\hat{\boldsymbol{\theta}}_2$. To overcome this problem, we chose to represent the pdf of the scores as Q-dimensional images and then to match them using a cross-correlation registration. For this purpose, we partition the Q dimensional space in a set of R pixels (of dimension Q). For a given set of parameters $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2 \cdots, \hat{\boldsymbol{\mu}}_C, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \cdots, \widehat{\boldsymbol{\Sigma}}_C\}$, the image intensity at a position $\mathbf{w}$, where $\mathbf{w}$ is a vector of Q components, is given by the mixture of Gaussians:

$$g(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \sum_{c=1}^{C} f(\mathbf{w}, \hat{\boldsymbol{\mu}}_c, \widehat{\boldsymbol{\Sigma}}_c). \tag{3.42}$$

[2] Note that in our case, the mixture weights π_c are omitted to enable classes having a different number of samples between two images to be dealt with in the following step.

where $f(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the value at $\mathbf{w}$ of the multivariate Gaussian probability function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Using these equations and averaging over the R pixels $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_R\}$, a global mismatch error between images is computed as:

$$Err(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \sum_{\mathbf{w} \in R} \big(g(\mathbf{w}, \hat{\boldsymbol{\theta}}_1) - g(\mathbf{w}, \hat{\boldsymbol{\theta}}_2)\big)^2 \tag{3.43}$$

In practice, the sum is bounded by plausible boundaries given by min/max scores on each axis. If Q is large, a subset of R of random vectors $\mathbf{w}$ can be selected.

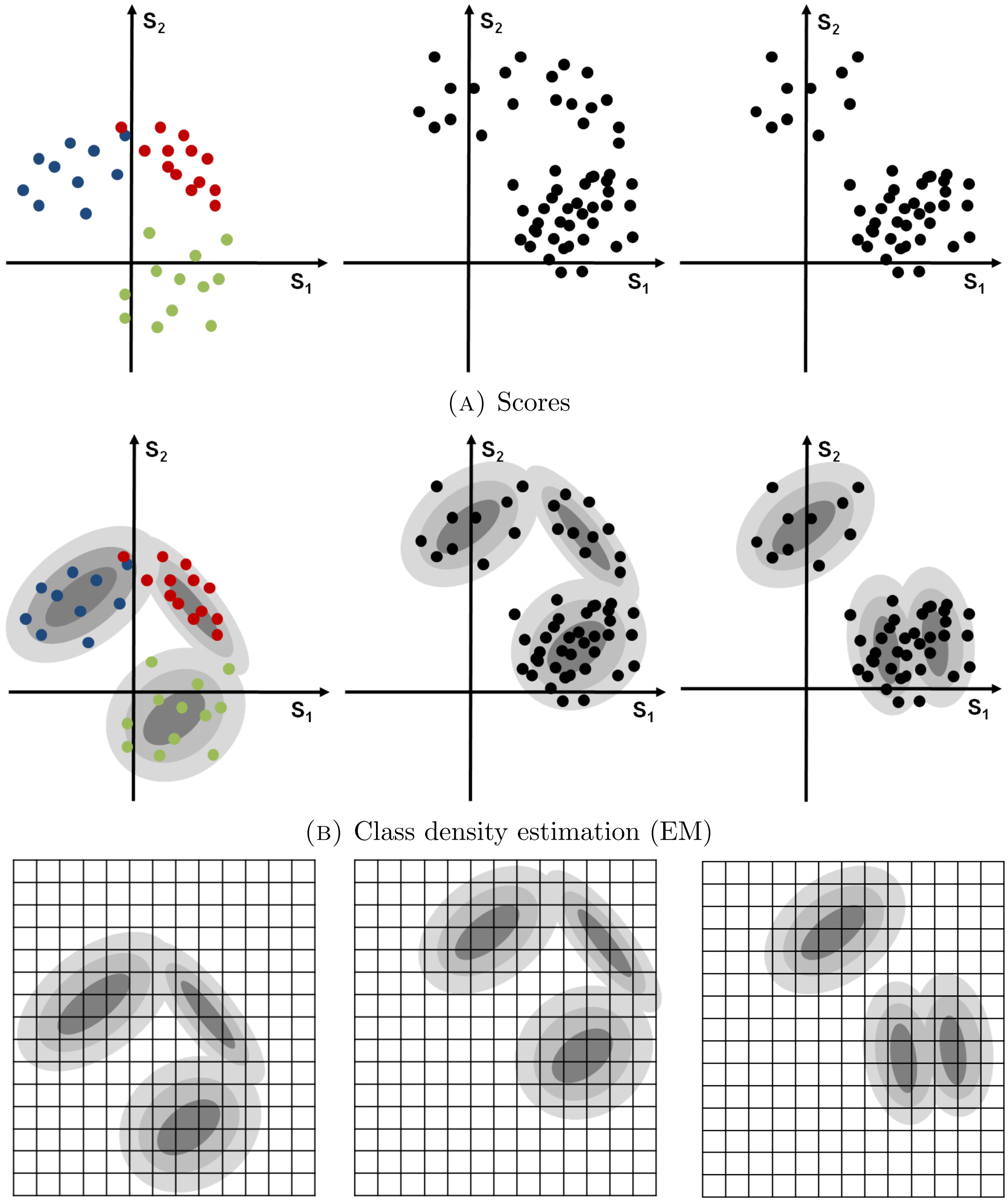
Let us define a translation operator $\mathcal{T}$ on $\hat{\boldsymbol{\theta}}$:

$$\mathcal{T}(\hat{\boldsymbol{\theta}}, \mathbf{t}) = \{\hat{\boldsymbol{\mu}}_1 + \mathbf{t}, \hat{\boldsymbol{\mu}}_2 + \mathbf{t}, \cdots, \hat{\boldsymbol{\mu}}_C + \mathbf{t}, \widehat{\boldsymbol{\Sigma}}_1, \widehat{\boldsymbol{\Sigma}}_2, \cdots, \widehat{\boldsymbol{\Sigma}}_C\} \tag{3.44}$$

The final registration step only consists in the estimation of a translation in a Q-dimensional spaces and is done by:

$$\hat{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}} Err\big(\hat{\boldsymbol{\theta}}_1, \mathcal{T}(\hat{\boldsymbol{\theta}}_2, \mathbf{t})\big). \tag{3.45}$$

for which an important initial guess is given by the difference of positions in each score image of the average observations.



(A) Scores

(B) Class density estimation (EM)

(C) Image sampling

FIGURE 3.11: Creation of a 2-dimensional image from the score of the training set (left), unknown labels (middle) and unknown labels with one missing class (right).

Chapter 4

Experimental Results

Contents

The objective of this chapter is to show the relevance of the developed approaches using hyperspectral images acquired in real outdoor conditions. Two data sets (called A and B) are used in order to illustrate the main steps of each part of the approaches and to give a better understanding of how the three approaches proposed in the previous chapter can be implemented together. The data sets that contain both reflectance and radiance images are described in the first section. In the second and third section, both DROP-D and the spectral-spatial approach are detailed using reflectance images. Finally, in the last section, we show that using the score registration on the log-radiance images provide similar results as when using the classical reflectance correction.

4.1 Data sets

For the purpose of this study, we will consider two data sets (called A and B). These data sets, are represented in Figure 4.1 and 4.4 respectively.

In the remainder of this chapter, the notation used to describe the data sets are:
1^{st} character: letter R for reflectance and L for radiance (luminance);
2^{nd} character: letter A or B for the data set;
3^{rd} character: subset number within the data set.

Note that in every case, the models are calibrated using data 1.

4.1.1 Data set A: Proximal detection

For this data set, illustrated in Figure 4.1, short-range hyperspectral images were acquired using a Hyspex V-NIR 1600 camera (Norks Elektro Optikk, Norway).

The images were recorded in-field using a translation stage mounted on a tractor (see illustration in Figure 4.2). The acquisition device was formerly developed by Irstea to map the nitrogen content in a wheat crop (see [Vigneau, 2010] for details) and with a very high spatial resolution. The images were thus captured at 1 meter above the ground, which led to a spatial resolution of 0.2 mm/pixel.

Using a calibrated reference surface in each image (gray plate seen Figure 4.1), radiance images (LXX) were transformed into reflectance images (RXX).

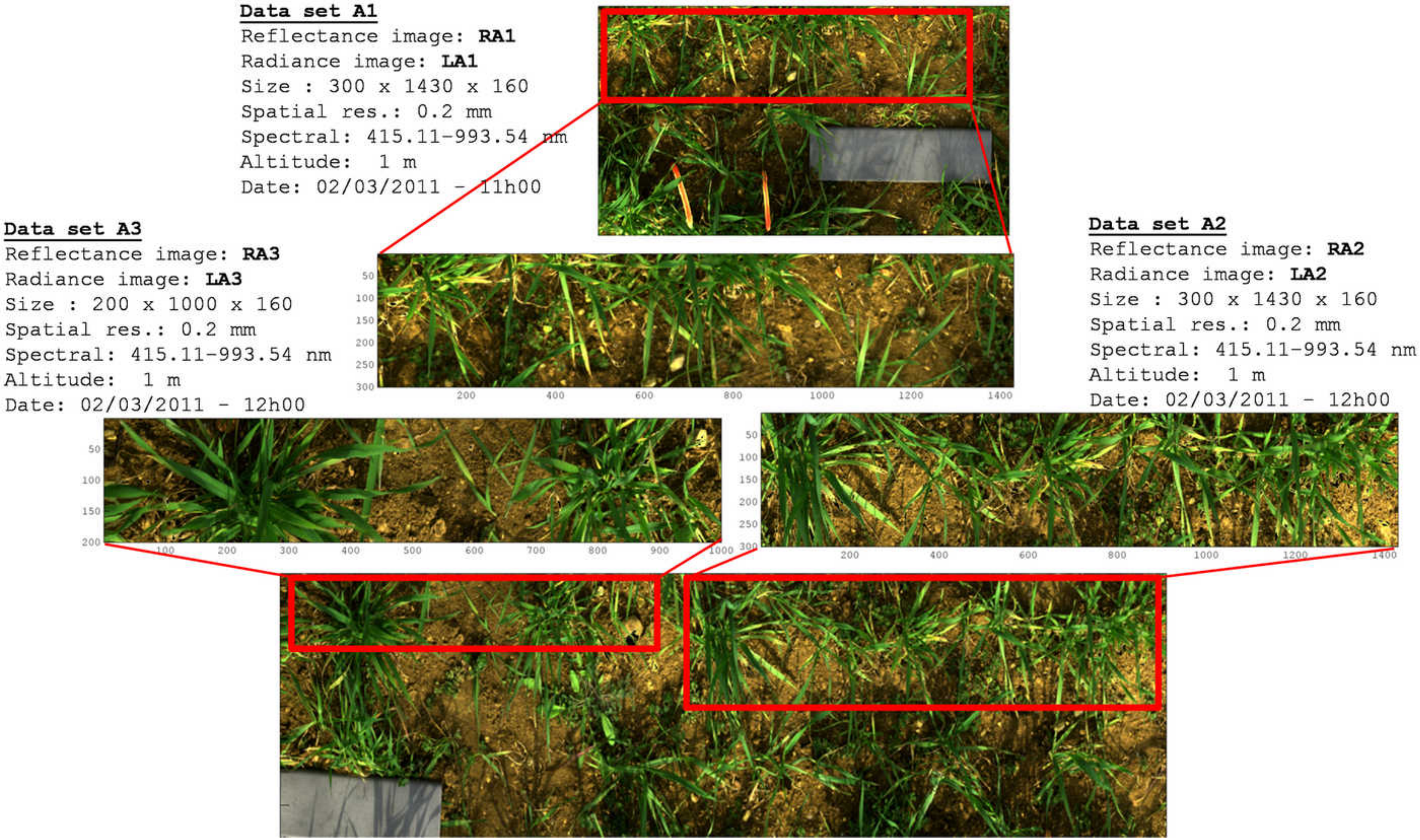


FIGURE 4.1: Presentation of the data set A: proximal detection

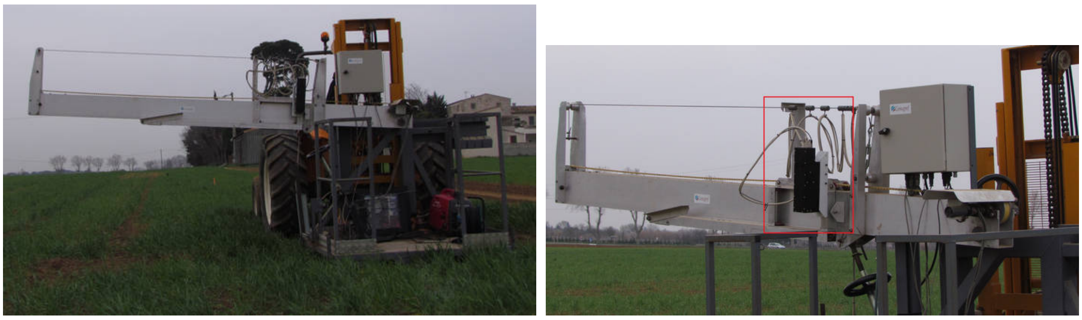


FIGURE 4.2: Illustration of Irstea acquisition setup for in-field measurements. The Hyspex camera is in the red rectangle on the right figure.

The acquired spectra were composed of 160 spectral bands ranging from 415.11 nm to 993.54 nm. Due to low sensitivity of the sensor in the NIR, the 20 spectral bands above 920.78 nm were discarded. Then, because of the high absorption of oxygen at around 750 nm, bands 93 to 96 were discarded as well.

Two acquisitions were performed at 1 hour interval at different places in the same field. From these two HS images, three sub-data sets were created: A1, A2 and A3.

Data set A1 was used to calibrate the models as detailed in Figure 4.3. The ground truth map is presented in Figure 4.3a and its associated class labels in Figure 4.3b. The three classes to discriminate are *wheat*, *weed* and *soil*. These classes are represented in every figure with the color defined in Figure 4.3b.

For training the models, 100 spectra per class (300 in total) were randomly extracted from the available ground truth map, which corresponds to approximately 0.6% of the available data. The amount of available data for training versus validating the model are represented as a principal component scatter plot in Figure 4.3c.

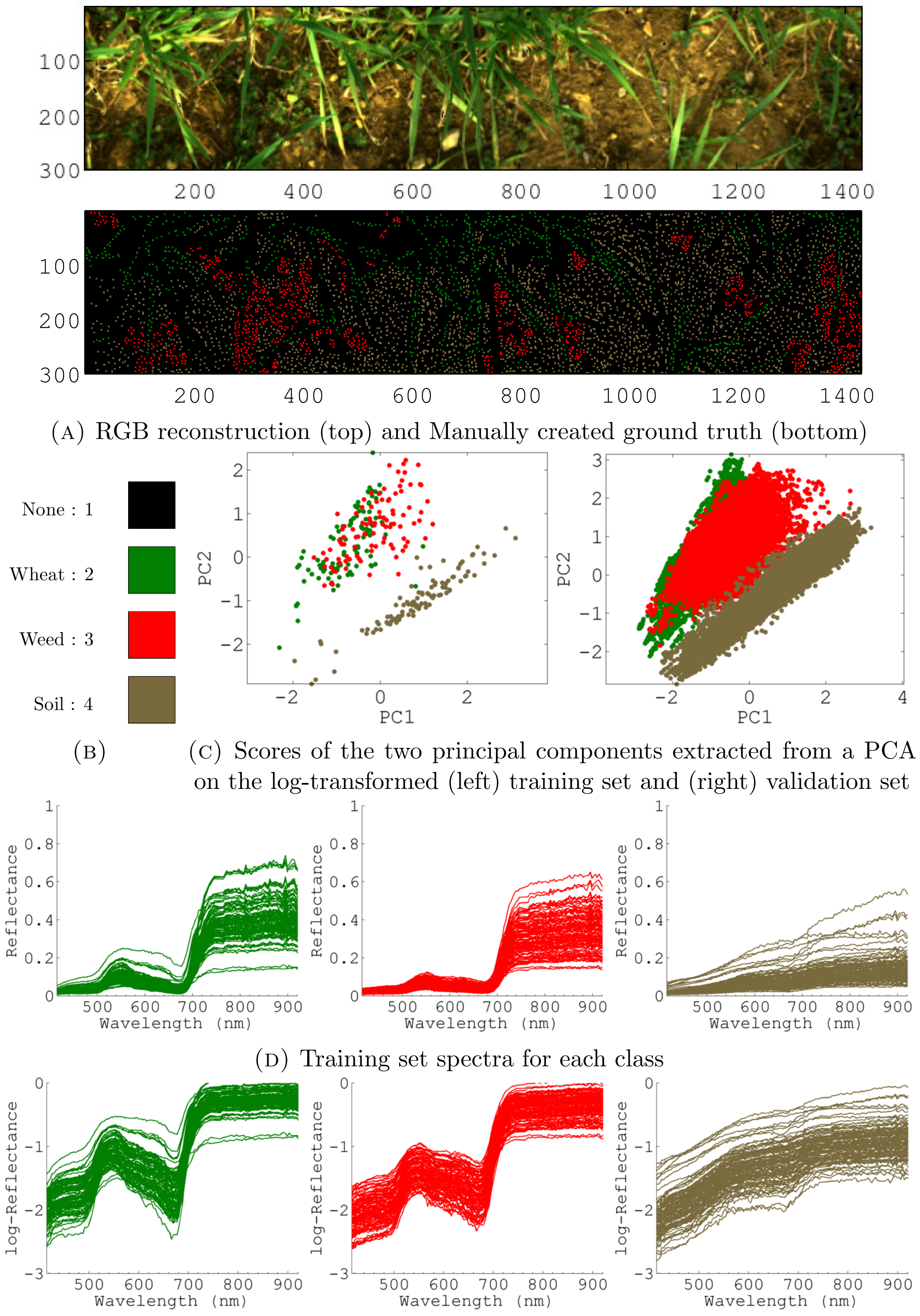
These training spectra for each class are represented in Figure 4.3d. For consistency of the results with the final section, in which a logarithm transformation has to be performed, we used log-transformed spectra throughout the whole chapter. The log-transformed spectra are presented in Figure 4.3e.

Data set A2 corresponds to the same field as data set A1 but acquired 1 hour later. A similar ground truth was manually created but is not represented here. This data set is used in the final section to illustrate the translation problem occurring when using log-radiance images.

Data set A3 is extracted from the same image as data set A2. It corresponds to an area in which only two classes (wheat and soil) are represented. This data set is also only used in the final section to assess the robustness of the registration step when there is a missing class.

4.1.2 Data set B: Remote-sensing

This data set, illustrated in Figure 4.4, contains remotely-sensed hyperspectral images acquired with a camera (Hyspex V-NIR 1600, Norks Elektro Optikk, Norway) embedded in a plane Piper Seneca II PA 34. The data used for this study were extracted from a field measurement campaign carried out by Actimar within the exploratory research and innovation project named HypLitt over the Quiberon peninsulas, France. Further information of the measurement campaign that contains 133 images and that maps 404 linear km are found in [Smet et al., 2010].



(A) RGB reconstruction (top) and Manually created ground truth (bottom)

(B)

(C) Scores of the two principal components extracted from a PCA on the log-transformed (left) training set and (right) validation set

(D) Training set spectra for each class

(E) Log-transformed training set spectra for each class

FIGURE 4.3: Presentation of the dataset RA1

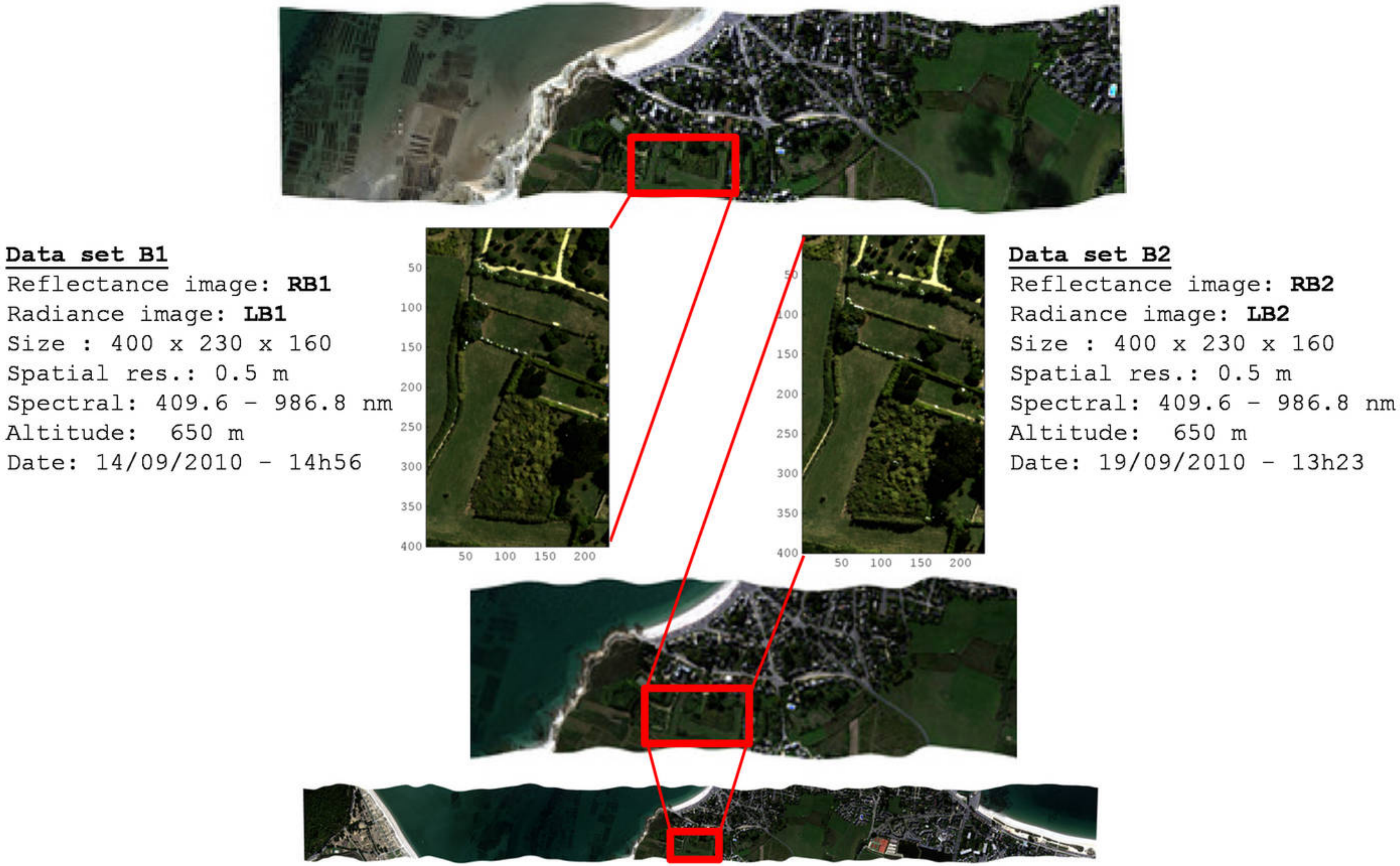


FIGURE 4.4: Presentation of the data set B: remote-sensing

For our purpose we only used two images (B1 and B2) acquired at 650 m above the ground level with a spatial resolution of 0.5 m. We chose these images, represented in Figure 4.4, because they include a common region not affected by any cloud shadow for which a ground truth could be manually created. To help with the ground truth creation, another image acquired at 500 m above the ground (spatial resolution of 0.4 m) was used.

Reflectance images were obtained through the atmospheric model ATCOR and then adjusted using spectroradiometric measurement on the ground using reference surfaces [Smet et al., 2010]. Because of low signal values below 442.3 nm, the ten first spectral bands were discarded. Then, because of saturation of vegetation spectra above 841.6 nm, spectral bands from 121 to 160 were discarded as well. Finally, the oxygen absorption bands at around 750 nm, which corresponds to spectral bands 84 to 90 were also removed.

Note that the acquisition of B1 and B2 were performed at 6 days of interval at 14h56 and 13h23 respectively.

Data set B1 was used to calibrate the models and is represented in Figure 4.5. The ground truth map is presented in Figure 4.5a and its associated class labels in Figure 4.5b. The four classes to discriminate are *grass*, *deciduous*, *conifer* and *sand*. These classes are represented in every figure with the color defined in Figure 4.5b.

For training the models, 100 spectra per class (400 in total) were randomly extracted from the available ground truth map, which corresponds to approximately 4.2% of the available data. The amount of available data for training and for validating the model are represented as a principal component scatter plot in Figure 4.5c.

These training spectra for each class are represented in Figure 4.5d. For consistency of the results with the final section in which a logarithm transformation has to be performed, we used log-transformed spectra throughout the whole chapter. The log-transformed spectra are presented in Figure 4.5e.

Data set B2 corresponds to the same area as data set B1 but acquired 6 days later. A similar ground truth was manually created but is not represented here. This data set is used in the final section to illustrate the translation problem when models are calibrated with data set B1.

4.1.3 Performance measurements

In order to provide numerical results, different measures were used.

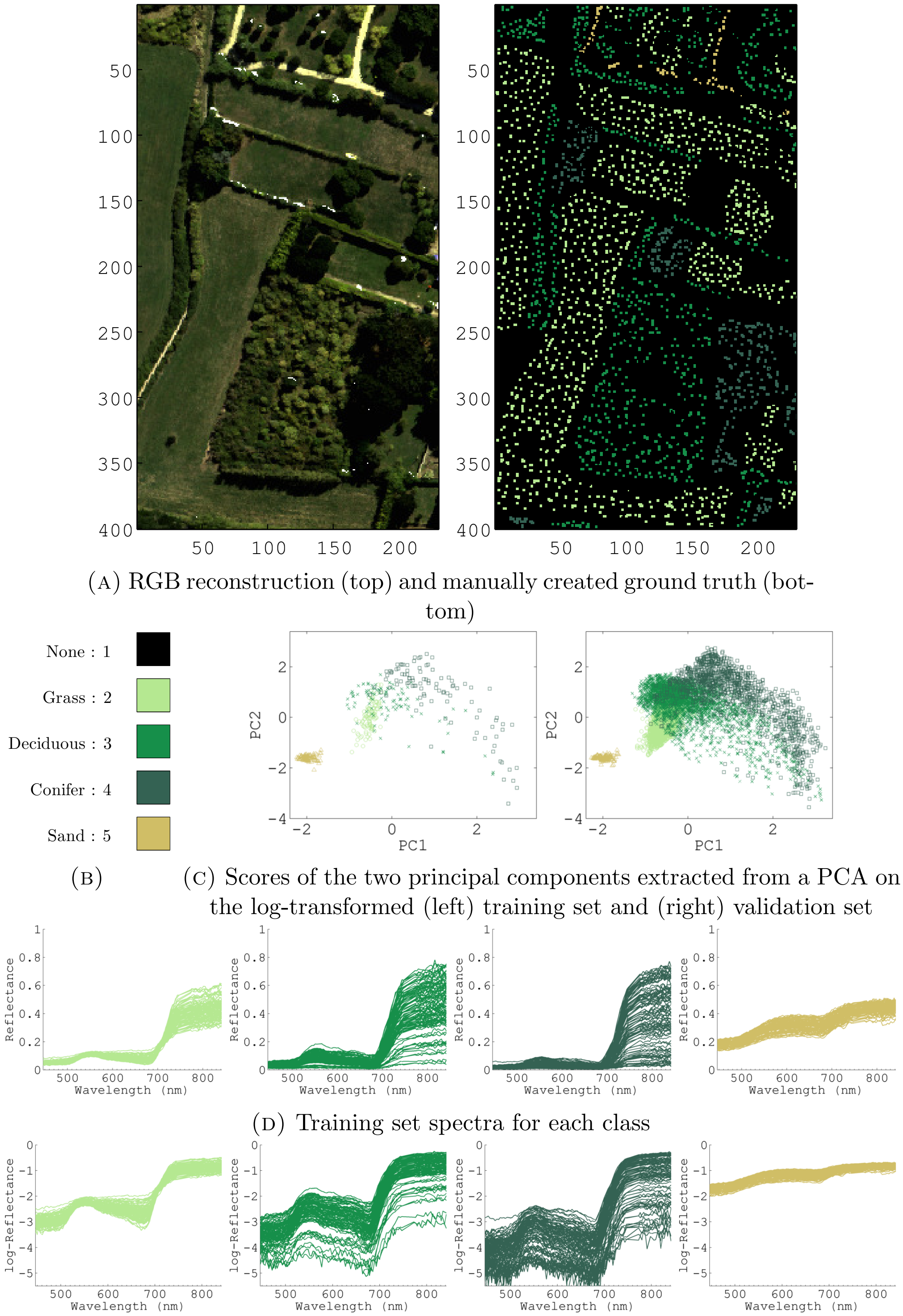
Classification error corresponds to the ratio of pixels incorrectly classified expressed in percent.
Wilk's Lambda is a measure of the class separability. It is given by the following determinant ratio:

$$\Lambda_{Wilks} = \frac{\mid \mathbf{W} \mid}{\mid \mathbf{T} \mid} = \frac{\mid \mathbf{W} \mid}{\mid \mathbf{B} + \mathbf{W} \mid}. \tag{4.1}$$

It can take values between 0 (perfect discrimination) to 1 (no discrimination).

Cross-validation results are used to tune different methods parameters. In such cases, we used a 10-fold procedure on the training data [Esbensen and Geladi, 2010]. For validation, results are presented on the overall ground truth minus the training data.

In the following, each part of the proposed approaches are detailed using these data sets.



(A) RGB reconstruction (top) and manually created ground truth (bottom)

(B)

(C) Scores of the two principal components extracted from a PCA on the log-transformed (left) training set and (right) validation set

(D) Training set spectra for each class

(E) Log-transformed training set spectra for each class

FIGURE 4.5: Presentation of the dataset RB1

4.2 Dimension reduction

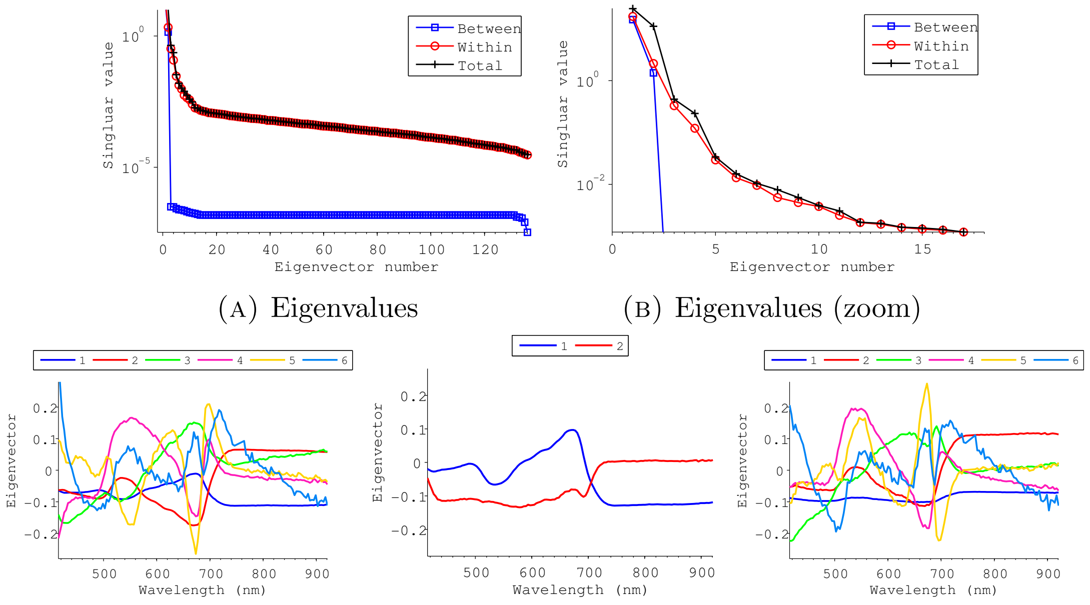
In this section we present different aspects of the dimension reduction method DROP-D using data sets RA1 and RB1. We first illustrate the collinearity of the scatter matrices eigenvectors in the variable space. Secondly, we show DROP-D in action step by step and illustrate the effect of removing the within-class variability on the class separability with data set RA1. Thirdly, we show, using data set RA1, that the number of within-class axes to remove can be tuned without cross-validation by comparing calibration and cross validation results. Then, we show, using data set RB2, that only calibration data can be used to select every DROP-D parameters. We also illustrate, using an artificial data set, that by removing information DROP-D cannot learn a class structure when there is none. Finally, we compare DROP-D classification performances with PCA-LDA, Nullspace LDA (NLDA) and PLS-LDA.

4.2.1 Collinearity in $\mathbb{R}^P$

In order to assess the collinearity issue in the variable space, let us first have a look at the eigenstructure of the total, between- and within- class scatter matrices for both data sets. In Figure 4.6 and Figure 4.7 are represented the eigenvalue plot and the main eigenvectors for data set RA1 and RB1 respectively. The angle between each combination of these eigenvectors is also given is Table 4.1 and Table 4.2.

The eigenvalue plot illustrates that the maximum number of eigenvectors for $\mathbf{B}$ is $C-1$, where C is the number of different classes. For RA1, in which three classes have to be discriminated, there are thus only 2 non-zero eigenvalues (Figure 4.6). Similarly, for RB1 there are only 3 non-zero eigenvalues (Figure 4.7).

For $\mathbf{W}$ and $\mathbf{T}$, the number of non-zero eigenvalue is at maximum $\min(P-1, N)$, where P is the number of variables and N the number of observations in the training set. In our case, for both data sets, we have more observations than variables. Therefore, these matrices ranks are numerically full. However, they cannot be inverted because of their bad conditioning (ratio of the maximum to minimum eigenvalue). The actual rank is thus only around 10 to 15, the remaining is only due to observation noise.



(A) Eigenvalues

(B) Eigenvalues (zoom)

(C) **T**

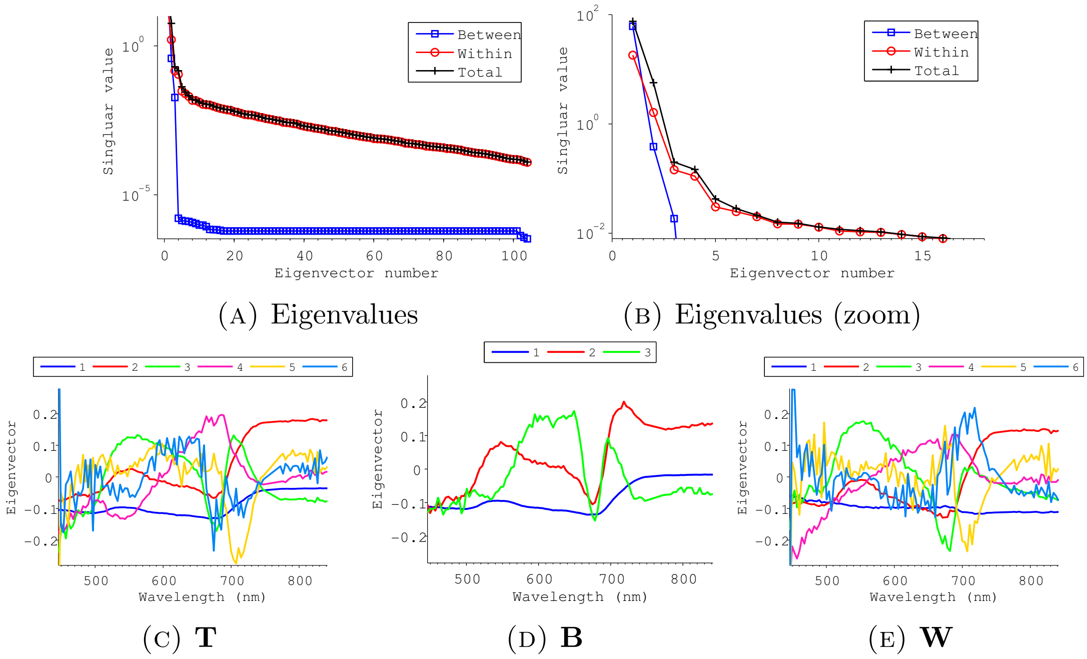
(D) **B**

(E) **W**

FIGURE 4.6: Dataset RA1: Eigenvalues and the principal eigenvectors of the scatter matrices **T**, **B** and **W**

		$E_b(\mathbf{B})$		$E_Q(\mathbf{T})$					
		$b=1$	$b=2$	$Q=1$	$Q=2$	$Q=3$	$Q=4$	$Q=5$	$Q=6$
$E_w(\mathbf{W})$	$w=1$	63	**30**	**29**	62	88	89	90	90
	$w=2$	**34**	61	62	**30**	80	90	88	90
	$w=3$	90	90	85	83	**28**	64	88	89
	$w=4$	89	88	87	87	65	**27**	89	82
	$w=5$	89	89	89	88	88	90	**8**	89
	$w=6$	89	89	90	90	87	86	89	**23**
	$E_b(\mathbf{B})$	$b=1$		**34**	56	88	89	90	90
		$b=2$		58	**38**	77	77	89	88

TABLE 4.1: Dataset RA1: angle (in degree) between the principal eigenvectors of the scatter matrices **T**, **B** and **W**



(A) Eigenvalues

(B) Eigenvalues (zoom)

(C) **T**

(D) **B**

(E) **W**

FIGURE 4.7: Dataset RB1: Eigenvalues and the principal eigenvectors of the scatter matrices **T**, **B** and **W**

		$E_b(\mathbf{B})$			$E_Q(\mathbf{T})$					
		$b=1$	$b=2$	$b=3$	$Q=1$	$Q=2$	$Q=3$	$Q=4$	$Q=5$	$Q=6$
$E_w(\mathbf{W})$	$w=1$	**28**	63	86	**22**	68	89	90	90	90
	$w=2$	66	**44**	88	68	**24**	81	88	89	90
	$w=3$	81	65	77	87	84	**28**	64	87	88
	$w=4$	90	90	88	88	85	65	**30**	78	89
	$w=5$	90	89	90	90	90	87	81	**26**	74
	$w=6$	90	89	83	90	90	88	90	81	**31**
			$E_b(\mathbf{B})$	$b=1$	**6**	84	90	90	90	90
				$b=2$	85	**30**	81	64	81	86
				$b=3$	88	68	53	54	87	76

TABLE 4.2: Dataset RB1: angle (in degree) between the principal eigenvectors of the scatter matrices **T**, **B** and **W**

The eigenvector plots on the Figure 4.6 and Figure 4.7 illustrate well the noise captured by these eigenvectors. For RA1, the fifth eigenvector is already noisy and for RB1, noise is perceived starting from the fourth eigenvector.

Observing the shape of the eigenvectors plotted as curves gives a first hint on the non-orthogonality of the eigenvectors: some of them are nearly identical between the different sets. This intuition is confirmed by the angles reported in Table: 4.1 and Table 4.2.

With RA1, some eigenvectors are clearly non-orthogonal (angle $<$ 90 degrees). In this particular case, most non-orthogonal eigenvectors have an angle of approximately 30 degrees. As we will see in the following, with this type of data, a compromise has to be made on removing or keeping these within-class eigenvectors using an orthogonal projection because of a potential loss in discrimination power.

With data set RB1, the non-orthogonality is still large for the within and total scatter matrices with approximately 25 degrees. However, for the between-class scatter matrix, only its two first eigenvectors are non-orthogonal to those of the within-class matrix. This third eigenvector is in addition, slightly collinear to the third to total scatter matrix eigenvector, which shows that it corresponds to an important direction for this data set. In the following we will show using DROP-D that preserving the two first between-class principal axes and removing one within-class axis proves to be the optimal combination for this data set. Note that this is just an observation and this table in itself cannot be used to tuned DROP-D parameters. In particular, at every projection the angle between each eigenvector is changed and a new table would have to be analyzed the same way until an optimal combination is found. For example, in Table 4.3 we show the angle of the between- to within-class scatter eigenvector after removing the first between-class axis. In this space, the first between-class axis thus become perfectly orthogonal to all within-class axes. We also observe that the two most collinear vectors are the combination $(w^*, b) = (1, 2)$ and $(w^*, b) = (3, 3)$. Then, removing the second axis of the between-class (see Table 4.4), the third axis of the between-class scatter matrix becomes the most collinear to the first within-class axis.

Although this approach, which consists in observing the angle between eigenvectors is interesting to understand the class structure of the dataset, tuning DROP-D parameters looking at these tables is unmanageable as the number of classes

			$E_b(\mathbf{B})$	
		$b=1$	$b=2$	$b=3$
$E^*_w(\mathbf{W}^*)$	$w^*=1$	90	**26**	74
	$w^*=2$	90	79	88
	$w^*=3$	90	72	**41**
	$w^*=4$	90	79	78
	$w^*=5$	90	84	78
	$w^*=6$	90	88	79

Table 4.3: Dataset RB1: angle between the principal eigenvectors of the scatter matrices $\mathbf{B}$ and $\mathbf{W}^*$ computed with $E_1(\mathbf{B})$ removed

			$E_b(\mathbf{B})$	
		$b=1$	$b=2$	$b=3$
$E^*_w(\mathbf{W}^*)$	$w^*=1$	90	90	**52**
	$w^*=2$	90	90	70
	$w^*=3$	90	90	69
	$w^*=4$	90	90	71
	$w^*=5$	90	90	78
	$w^*=6$	90	90	78

Table 4.4: Dataset RB1: angle between the principal eigenvectors of the scatter matrices $\mathbf{B}$ and $\mathbf{W}^*$ computed with $E_2(\mathbf{B})$ removed

increases. Therefore, in the following, we adopt a classical parameter tuning approach.

Final remarks concern the ‘shape’ of the obtained eigenvectors. Indeed, for both data sets, the discrimination includes mostly vegetation spectra. As it was observed in the data sets presentation (Figure 4.3 and Figure 4.4), vegetation spectra have a strong reflectance feature at around 700 nm. This transition, situated at the edge of the red and infrared part of the electromagnetic spectrum is very distinctive for vegetation spectra and is often referred to as the red-edge. This red-edge is mostly due to a strong absorption of the chlorophyll within the vegetation and is thus characteristic of the type of plant. The structure of the red edge (position, slope) is thus naturally found as a discriminative feature by classifiers. At the end of the red-edge, another characteristic feature of the vegetation spectra is the NIR plateau. Finally, the greenness of the vegetation is also often discriminative between vegetation types and specific features are thus often found in the 500 to 600 nm range.

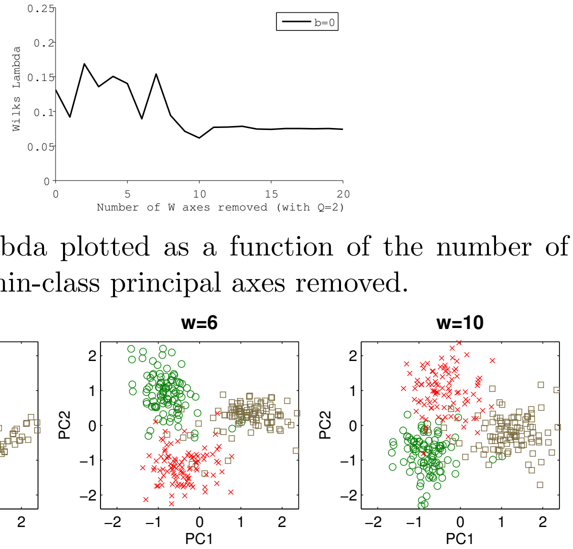
4.2.2 Effect of removing W on the class separability

As we have seen in the previous section, the between- and within-class scatter matrices can have some non-orthogonal principal directions. Also, in order to decrease the Wilk's Lambda and thus to increase class separability, a possible approach would be to suppress the within-class variability by removing the principal axis of the within-class scatter. However, because of this non-orthogonality, removing these directions may affect the class separability as well. The idea of DROP-D, as presented in the previous chapter, is thus to prevent the suppression of too much between-class scatter. In the following we present, using the data set RA1, this effect step by step for different numbers (w) of within-class axes removed and by preserving different numbers (b) of between-class directions.

With $b = 0$, no between-class direction is a priori preserved. Figure 4.8 show the Wilk's Lambda of the training data plotted as a function of the number of within-class directions removed. We also represent the obtained scatter plot for three specified values. At $w = 0$, the discriminant vectors correspond to the ones of the Principal Component Analysis (PCA) and we can observe the same scatter plot as in the data presentation of Figure 4.1. The two vegetation classes (wheat and weed) are poorly separated but distinct from the third class (soil). Then, until $w = 9$, the 'noisy' aspect of the obtained curve is due to the non-orthogonality of $\mathbf{W}$ and $\mathbf{T}$. The discriminant vectors keep changing due to the removal of within-class directions. Then, at around 9 or 10 removed axes, a clearer minimum is obtained. Removing more axis only degrades the class separability (even for the training set).

With $b = 1$, the principal direction of the between-class scatter is preserved. As observed in Figure 4.9, a similar 'noisy' pattern is obtained, but only until $w = 5$. In this case only the second axis of the between-class is affected by the successive cleaning of the within-class directions. The clear minimum obtained for $w = 5$ is stable until $w = 8$. Then, removing more directions starts affecting the class separability as well, e.g., see the Figure 4.9 at $w = 12$ where the wheat (green) and weed (red) starts to cluster.

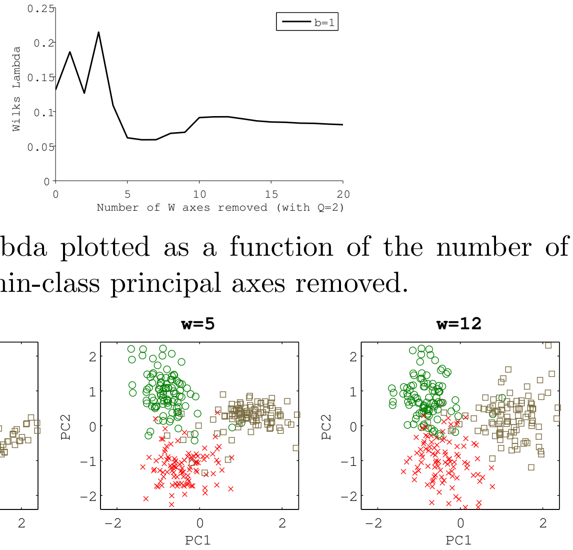
With $b = 2$, every between-class direction is preserved. With three classes, it corresponds to the limit case of DROP-D in which any cleaning does not change the class separability unless more discriminant axes (Q) are kept (results not shown).



(A) Wilk's Lambda plotted as a function of the number of within-class principal axes removed.

(B) Scores on principal components 1 and 2 for the training (left) and validation (rigth) sets

Figure 4.8: Dataset RA1: Class separability as a function of the number of within-class principal directions removed. ($b = 0$)



(B) Wilk's Lambda plotted as a function of the number of within-class principal axes removed.

(B) Scores on principal components 1 and 2 for the training (left) and validation (right) sets

Figure 4.9: Dataset RA1: Class separability as a function of the number of within-class principal directions removed. ($b = 1$)

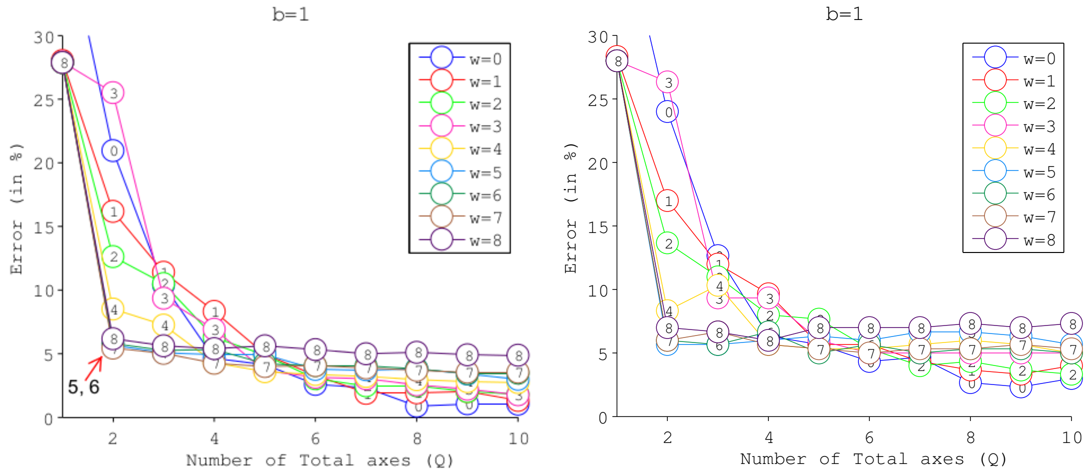
These results on class separability were all performed with $Q = 2$ in order to provide these two-dimensional scatter plots. In the general case, Q is another parameter to tune, which corresponds to the final number of discriminant vectors to be used. It actually corresponds to a PCA on the cleaned spectral matrix as explained in the previous chapter. Usually, with PCA, the correct number of components to retain is always subject to discussion because the error decreases only slowly and an optimal threshold is difficult to estimate. The rule of thumb in such cases is 'less is better'. Fortunately, we will see in the following figures that when cleaning the spectral matrix with DROP-D, this threshold appears to be easier to find.

4.2.3 Model calibration

As we have seen in the previous section, with a careful selection of the between- and within-class principal axes to keep or to remove, various class separabilities can be obtained. Also, we have seen that owing to the DROP-D approach, which consists in removing information (contrary to PLS-LDA, which learns the class structure by modeling $\mathbf{B}$), overfitting can be spotted directly on the training set by observing the class separability. In the following, we show that a similar 'behaviour' is obtained with the classification performance.

In figure 4.10, we show the classification error obtained with the training set (calibration error) and using a 10-fold cross-validation on the training set. This graph presents the classification error as a function of the number of final discriminant axes (Q) for different numbers of within-class axes removed (number inside the circle). With $w = 0$ (which corresponds to a classical PCA) both calibration and cross-validation error smoothly decrease without any clear minimum. Then, from $w = 1$ to $w = 4$, we obtain the similar noisy aspect, but in terms of classification performance. From $w = 5$ (optimal) to $w = 7$, the same classification error is obtained. In addition, a clear optimal value for Q emerged ($Q = 2$). Then, as we observed with the class separability, when removing one more axis, the error starts increasing. Therefore, $w = 5$ is chosen as an optimal value since it corresponds to the smallest value for which the optimal results are obtained. Also note a similar behavior obtained for both calibration and cross-validation curves.

Let us assess the optimal parameters b, w and Q for the data set RB1 using only the calibration error. Figure 4.11 shows the four sets of curves that correspond to

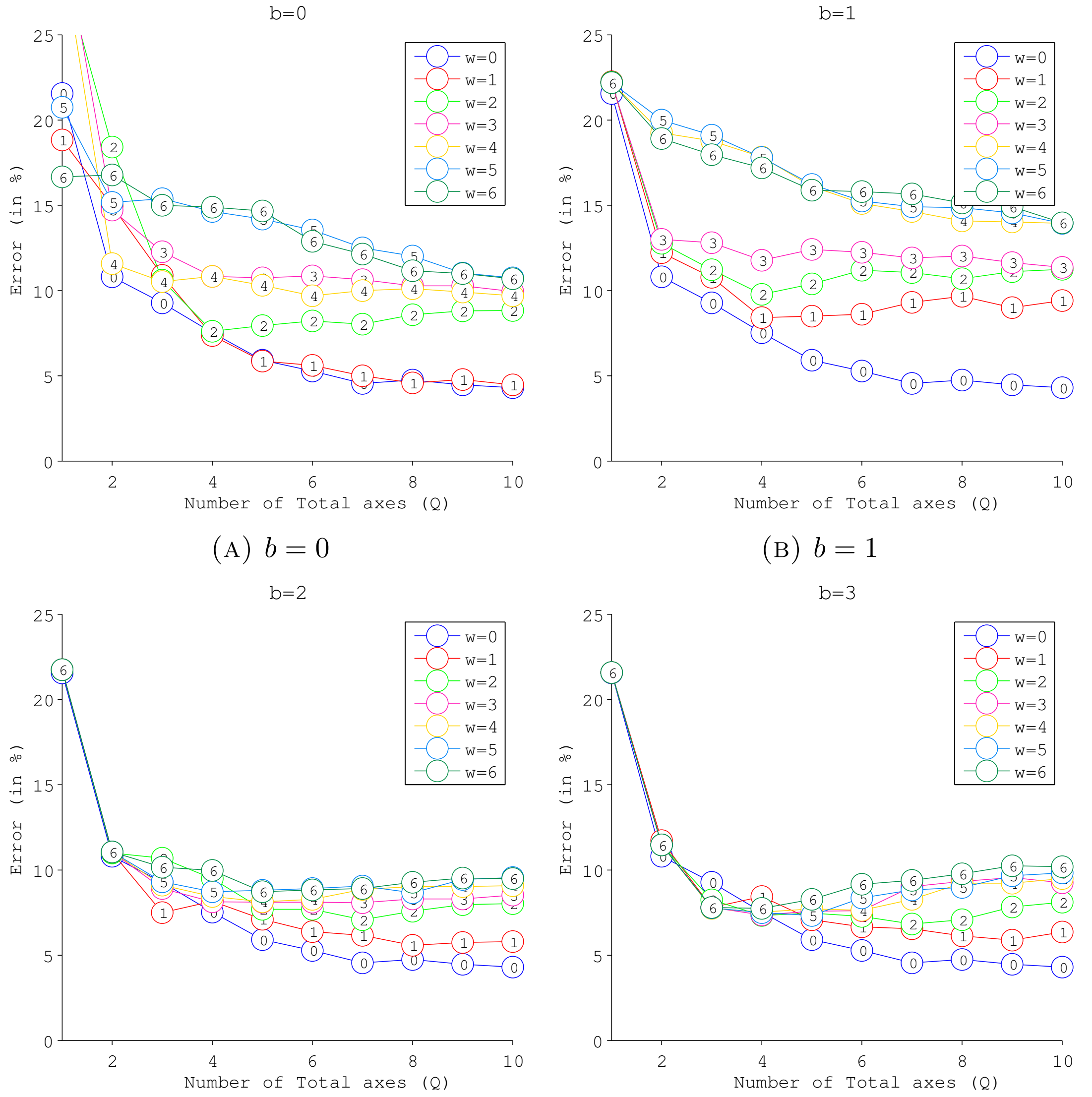


(A) Calibration error (B) Cross-validation error

Figure 4.10: Dataset RA1: Representation of the classification error of calibration (A) and cross-validation (B) for different parameters for w and Q with $b = 1$.

every possible values for b. Without preserving the two first between-class axes, removing w always leads to worse results (Figure 4.11a and 4.11b). Note that in these cases, deciding for an optimal value for Q is not an easy task as explained before. When $b = 2$ (Figure 4.11c), a clear optimum is reached by removing only one within-class axis. In addition, the optimal value for Q also becomes more obvious to choose. In particular, removing more w or increasing Q both lead to worse results. Finally, when preserving the last possible between-class direction, classification results become slightly worse. The optimal parameters for this data set are thus $b = 2$, $w = 1$ and $Q = 3$. These values actually correspond to the one obtained with the 10-fold cross-validation (not represented here).

Finally, to illustrate that DROP-D cannot learn a class structure when there is none, we show in Figure 4.12 the classification error obtained with data set RA1 in which the class matrix has been randomly shuffled. For any number of removed within-class axis, no structure can be extracted and the classification results remain the same. On the other hand, with a PLS-LDA model trained on the same data, a class structure can be learned and is thus prone to overfitting. This main difference comes from the fact that DROP-D removes **W** while PLS-LDA learns a class structure by modeling **B** [Barker and Rayens, 2003]. Therefore, because of the high dimensionality, a class structure can always be learned, especially with a small training set. On the contrary, when removing information with DROP-D, if the information was useful for discrimination, even the training data is affected



(A) $b = 0$ (B) $b = 1$

(C) $b = 2$ (D) $b = 3$

FIGURE 4.11: Dataset RB1: representation of the training error of classification for different parameters b,w and Q.

by the loss.

4.2.4 Classification performances

As we have seen, a major interest of DROP-D is to provide a method relatively robust to overfitting. Let us now have a look at the classification performance that can provide this method on our data sets. We also compare the results with the most classically used dimension reduction methods, i.e., PCA-LDA, NLDA (setting w=15) and PLS-LDA. For all methods the class decision is made using

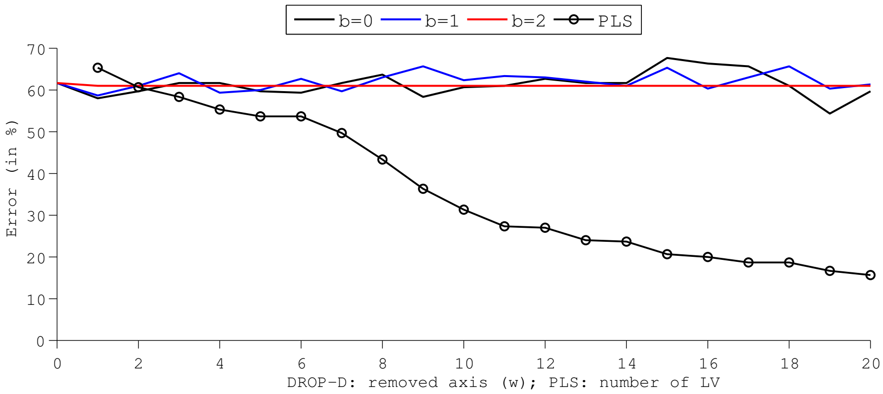


FIGURE 4.12: DROP-D classification error on the training set with a random class matrix.

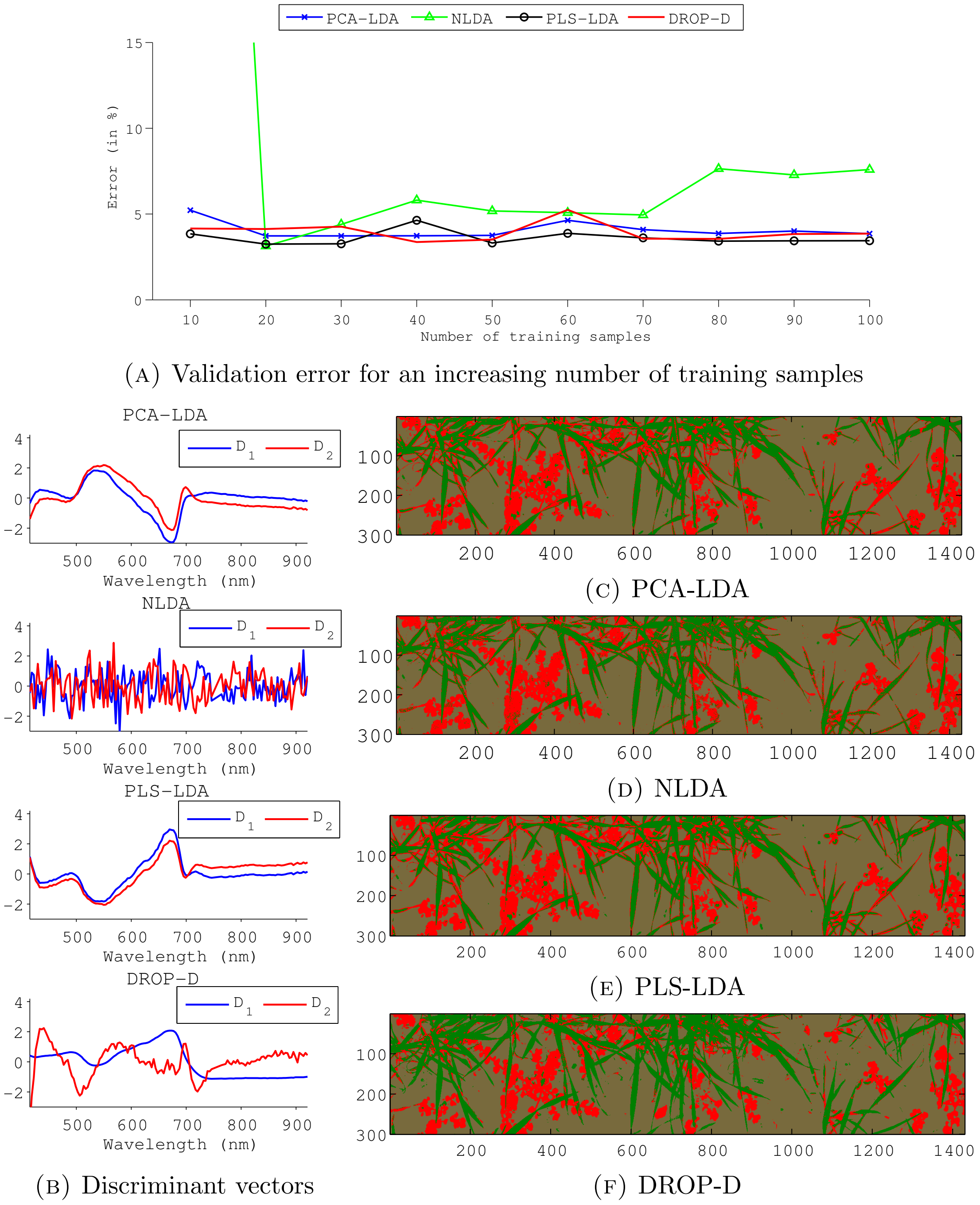
a Quadratic Discriminant Analysis (QDA) on the obtained scores (see Chapter 1 for details on QDA).

A first classification performance assessment is qualitative and is made by observing the shape of the obtained discriminant vectors. Indeed, as every dimension reduction method used is linear, the obtained discriminant vectors can be plotted as spectra and can be analyzed in the same way.

Then, for practical uses, it is interesting to assess the classification performances for different numbers of training samples. For this purpose, we randomly selected among the training set from 10 to the whole (100) spectra per class, by step of 10. Results are presented in Figure 4.13 and Figure 4.14 for data set RA1 and RB1 respectively. In these figures, we also provided the classification maps obtained for each method using the 100 spectra per class.

The obtained results in terms of classification peformance are very similar with these data sets. With our experience on other data sets using DROP-D (not represented here), the results highly depends on the data. DROP-D proves better than PLS-LDA in some cases only, but is generally better than PCA-LDA. In this sense, DROP-D offers an alternative way of reducing dimension in a supervised way.

All these methods appear to be relatively not sensitive to the lack of training samples. In particular, above 20 samples per class, the classification stabilizes to its optimal value, which is of great interest for practical uses.



(A) Validation error for an increasing number of training samples

(B) Discriminant vectors

(C) PCA-LDA

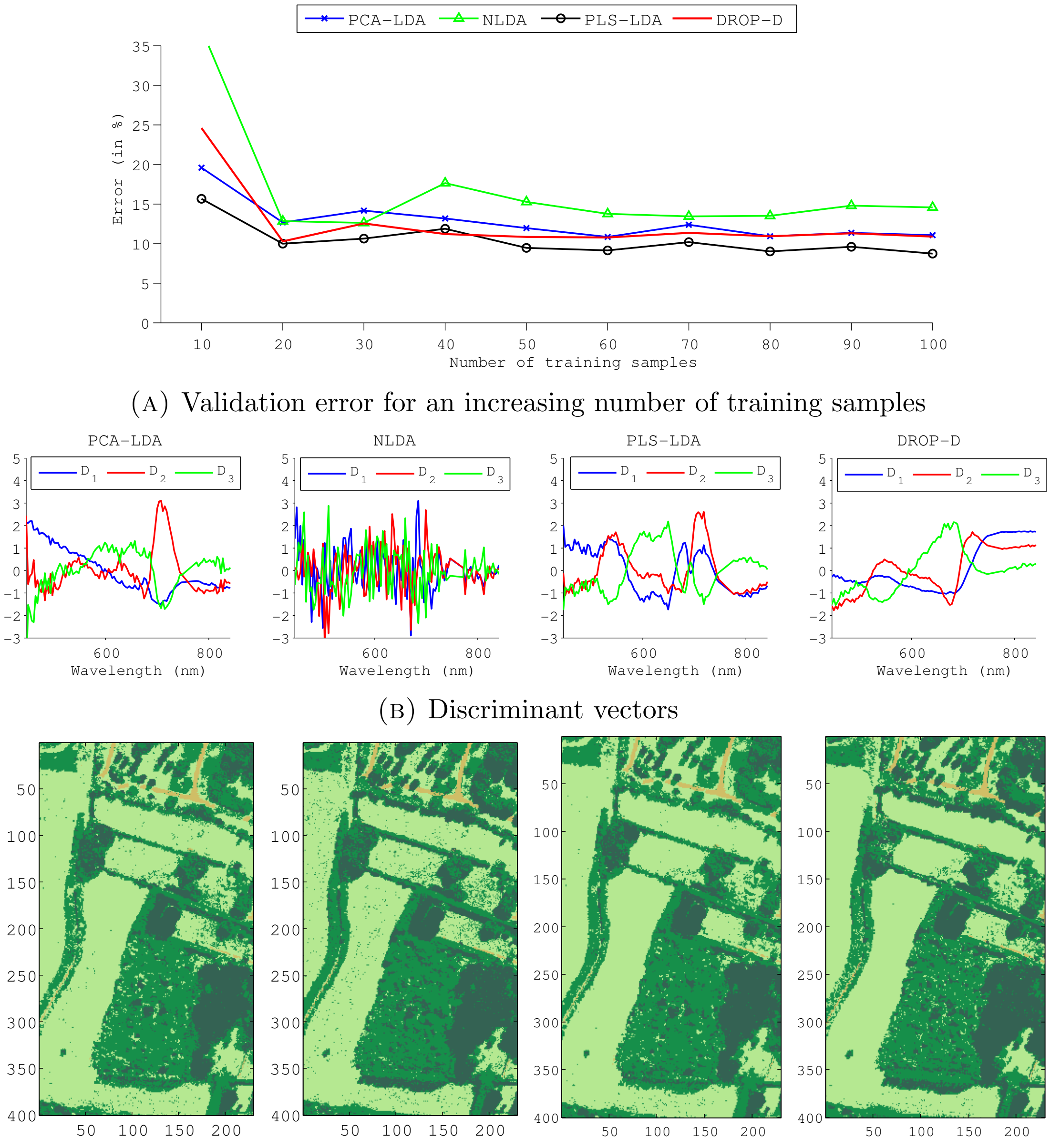
(D) NLDA

(E) PLS-LDA

(F) DROP-D

FIGURE 4.13: Dataset A1

For the discriminant vectors, only those of DROP-D and NLDA are orthogonal since they correspond to the eigenvectors of symmetric matrices. However, although it leads to high classification performances, the ‘shapes’ of NLDA discriminant vectors are not interpretable in practice, which probably explains the lack of interest from the chemometrics community. The DROP-D discriminant vectors that come from preserved between-class principal directions appear less noisy due to the averaging involved in the computation of these eigenvectors. However, the



(A) Validation error for an increasing number of training samples

(B) Discriminant vectors

(C) PCA-LDA (D) NLDA (E) PLS-LDA (F) DROP-D

FIGURE 4.14: Dataset RB1

discriminant vectors obtained by removing within-class axes are naturally noisier, but offer a different type of information from PLS-LDA or PCA-LDA (see Figure 4.13 for example).

In the following, we show that the relatively noisy aspect of classification maps can be dealt with using our spatial regularization approach.

4.3 Spatial regularization

In the previous chapter, we proposed a spatial regularization approach based on score images obtained in a supervised way. In a first first step, we show the interest of using a supervised dimension reduction by comparing the gradient images obtained with other methods. Then, for each data set, we show the influence of the tuning parameters of the spatial regularization on the classification outcome. We also illustrate the effect of this spatial regularization in both the image and the score space. Finally, using a benchmark data set, we compare our approach with other recently proposed spectral-spatial approaches.

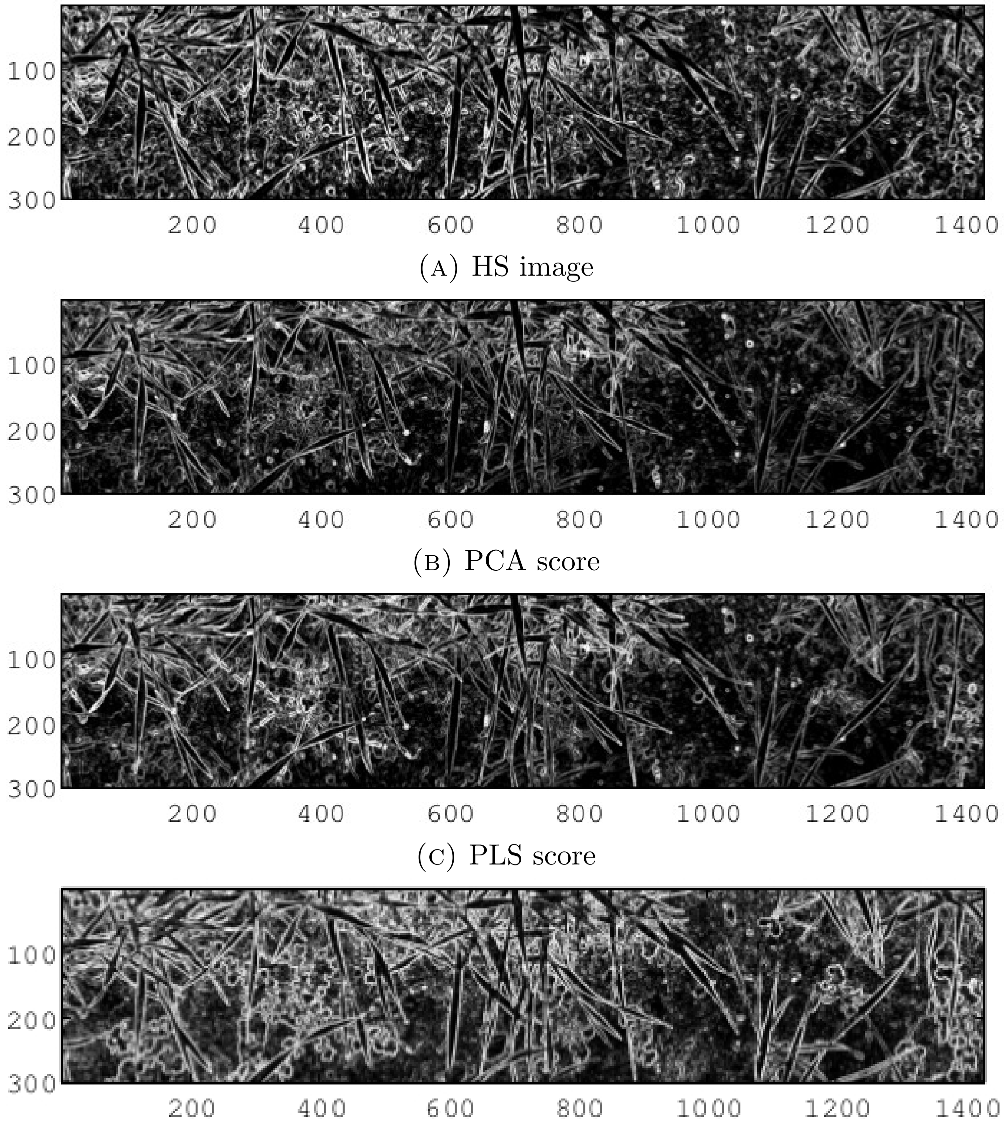
4.3.1 Validation of the approach

4.3.1.1 On 'what' to apply the regularization

In order to justify the choice of applying the diffusion process on a supervised score image, we represent in Figure 4.15 and Figure 4.16 the gradient images averaged over every channel for the data sets RA1 and RB1 respectively. These gradients were computed from the original HS image, from a PCA score image, from a PLS score image and from the DROP-D score image.

As expected, there is a high similarity between the HS image gradient and the PCA score image gradient because PCA summarizes most of the original HS image information in fewer components. Even though these two gradient images show high values at region borders, high values are also obtained inside regions because of the background variability, which is present in the original HS image and captured by the PCA.

On the contrary, the PLS and DROP-D score image gradients are different from the other two. By finding a trade-off between capturing spectral information and class information, the gradient images have low values within regions and high values mostly at the region borders. Therefore, because spatial regularization is guided by the image gradient intensity, it is more effectively used with a PLS or a DROP-D score image than with the original HS image or the PCA score image.



(A) HS image

(B) PCA score

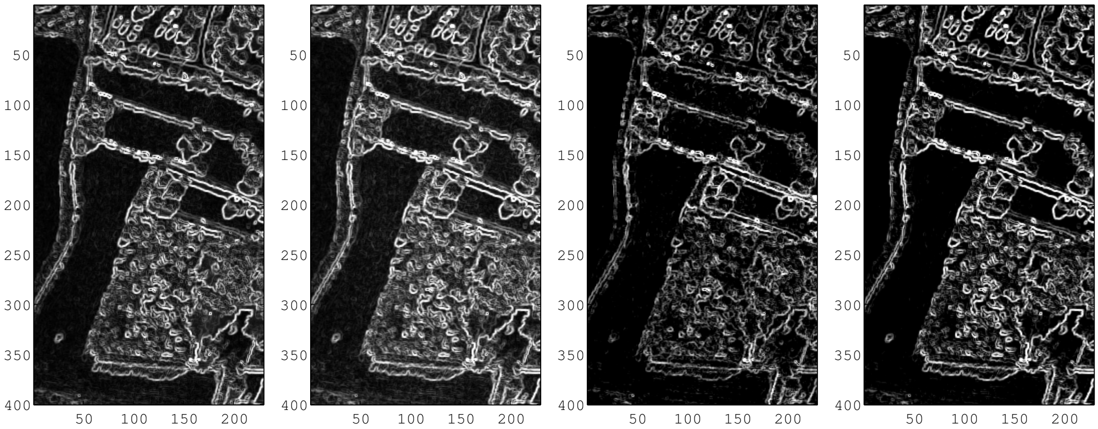
(C) PLS score

(D) DROP-D score

Figure 4.15: Gradient images computed from dataset RA1

4.3.1.2 'When' to apply regularization

In order to validate quantitatively the proposed approach, in Table 4.5 we provide the classification results obtained using the training set on the RA1 data set. In order to provide results also without dimension reduction, for this table we used a K-nearest neighbor (KNN) classifier with K=3. This table thus shows the results obtained without regularization (KNN, PCA-KNN, PLS-KNN and DROP-D-KNN), applying the regularization first (AR-KNN, AR-PCA-KNN, AR-PLS-KNN and AR-DROP-D-KNN), and applying the regularization after dimension reduction (PCA-AR-KNN, PLS-AR-KNN and DROP-D-AR-KNN).



(A) HS image (B) PCA score (C) PLS score (D) DROP-D score

Figure 4.16: Gradient images computed from dataset RB1

We first observe that the supervised approaches outperform the unsupervised ones. Then, for every method, applying a regularization usually improves the classification outcomes. For PCA, only a slight improvement is noted by applying the regularization before or after dimension reduction. However, PLS- and DROP-D-based results show that the regularization has to be performed after dimension reduction, which confirms the proposed approach detailed in the previous chapter.

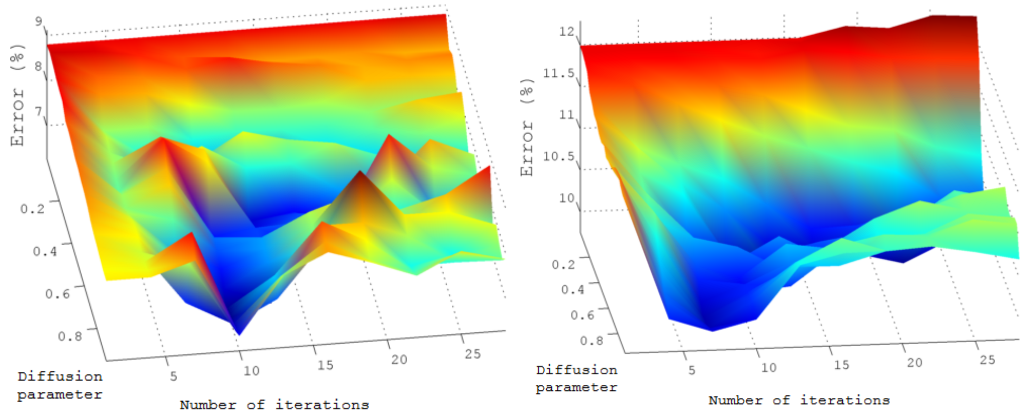
Table 4.5: Dataset RA1: Classification error (%) for different regularization procedures

Without regularization		Regularization first		Regularization second	
KNN	7.97	AR-KNN	7.62	n/a	n/a
PCA-KNN	7.35	AR-PCA-KNN	7.27	PCA-AR-KNN	7.26
PLS-KNN	6.84	AR-PLS-KNN	6.62	PLS-AR-KNN	**4.91**
DROP-D-KNN	7.14	AR-DROP-D-KNN	7.10	DROP-D-AR-KNN	**5.37**

4.3.2 Tuning robustness

In Figure 4.17, the evolution of the classification error is represented as a function of both the diffusion parameter (η) and the number of iterations using DROP-D-AR on data sets RA1 and RB1. The number of iterations ranged from 1 to 30, and η ranged from 0.01 to 1. Every DROP-D score was normalized to unit variance before regularization so that the diffusion parameter varied within the same range.

For both data sets, optimal values were situated within a wide region, ranging from 0.4 to 0.6 for the diffusion parameter and from 8 to 25 for the number of iterations. This illustrates that the method is relatively robust to parameter variations and easy to tune, which is convenient when different types of images need to be classified.



(A) Dataset RA1

(B) Dataset RB1

FIGURE 4.17: Influence of the diffusion parameters regarding the classification error (%)

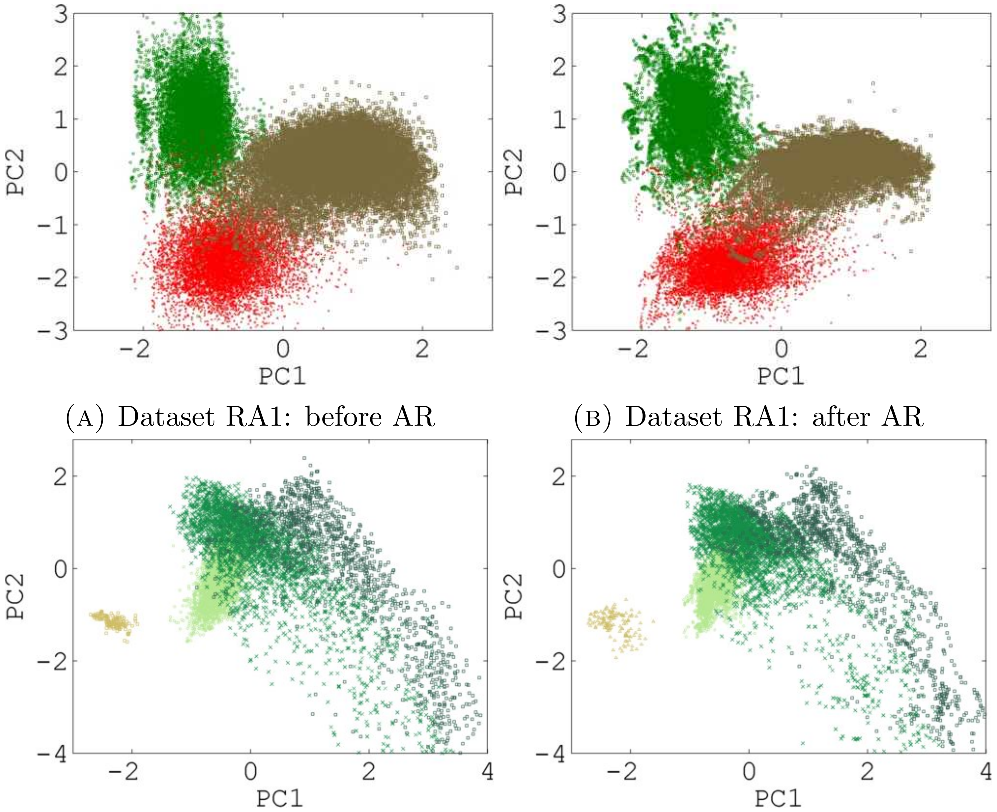
4.3.3 Effect on score versus spatial

Figure 4.18 presents the effects of AR in the spectral domain (scores) for both data sets. Similarly, Figure 4.19 and Fig. 4.20 present the effects of AR in the spatial domains for data sets RA1 and RB1 respectively.

The obtained regularized score images clearly indicate less variability within each class, which is confirmed by the images of differences (see Figure 4.19 and 4.20). On the other hand, class borders are well preserved, which emphasizes the importance of using an edge-preserving filter. Observing the scatter plots before and after regularization validates this observation, since every class was less spread out around its mean value, which leads to an increased class discriminability.

4.3.4 Classification results

In Figure 4.21 are represented the classification maps obtained before and after regularization for both data sets. In both cases, the classification noise is greatly decreased leading to more homogeneous classes. At the border, especially for data set RA1 (Figure 4.21) with the wheat leaves border, the classification error at the edges is also reduced.



(A) Dataset RA1: before AR

(B) Dataset RA1: after AR

(C) Dataset RB1: before AR

(D) Dataset RB1: after AR

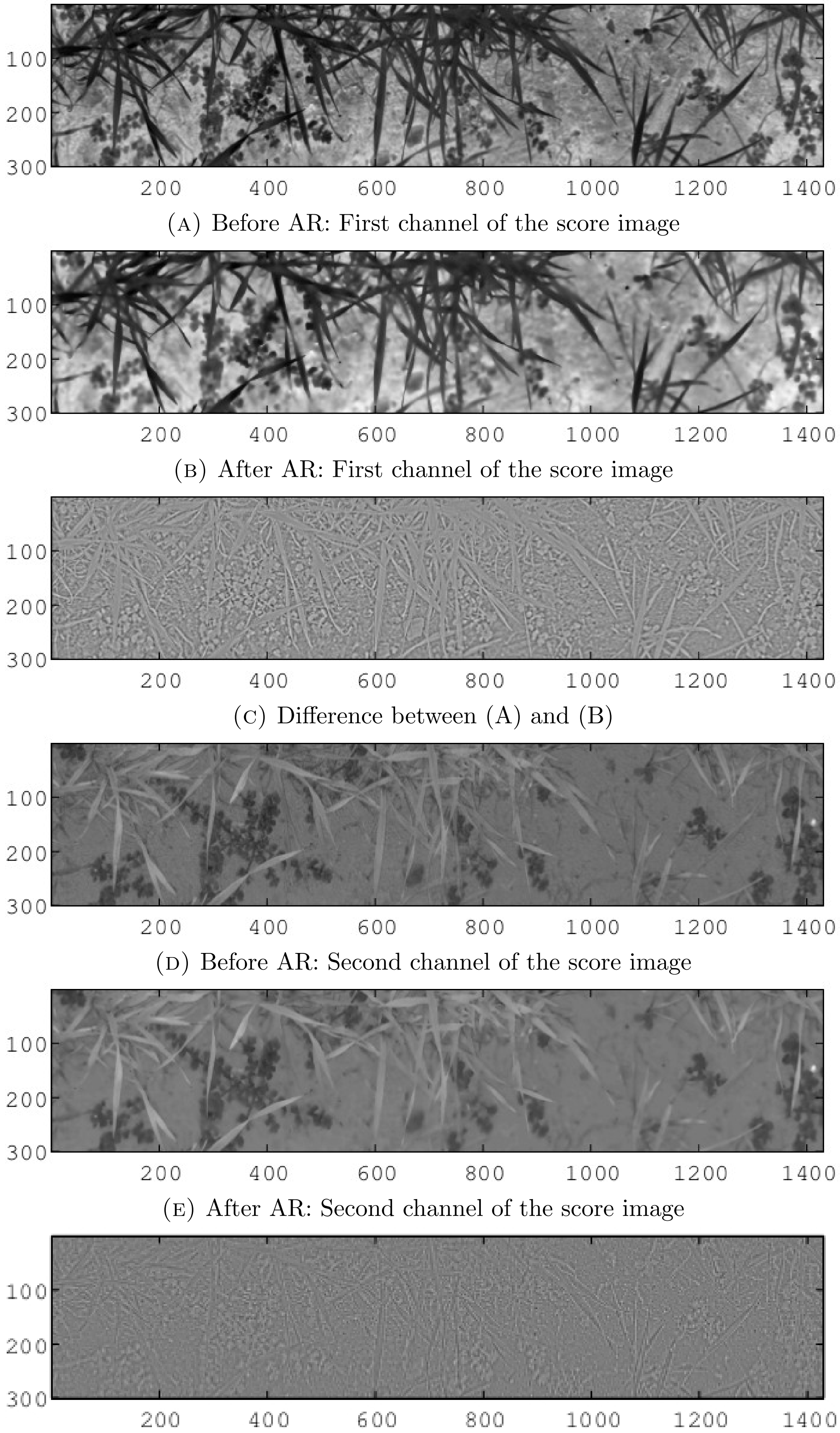
FIGURE 4.18: Effect of the regularization in the spectral domain. The two first score are represented before and after regularization

4.3.5 Comparison with other approaches

We finally compare our approach with some of the latest spectral spatial approaches developed by the remote sensing community. To do so, we used one of the remotely sensed hyperspectral images that is now considered as a benchmark for testing classification methods. Results on other benchmark images have been published [Hadoux et al., 2014] using PLS as a supervised dimension reduction.

The Salinas data set was acquired by the AVIRIS sensor over Salinas Valley, California at a spatial resolution of 3.7 meters per pixel. The image comprised 512×217 pixels and the ground truth contained sixteen classes. We discarded twenty bands affected by water absorption, in this case bands 108 to 112, 154 to 167 and 224. The ground truth map, class names and number of samples per class for each image are displayed in Figure 4.22.

For each method, every parameter was tuned using 10-fold cross-validation.



(A) Before AR: First channel of the score image

(B) After AR: First channel of the score image

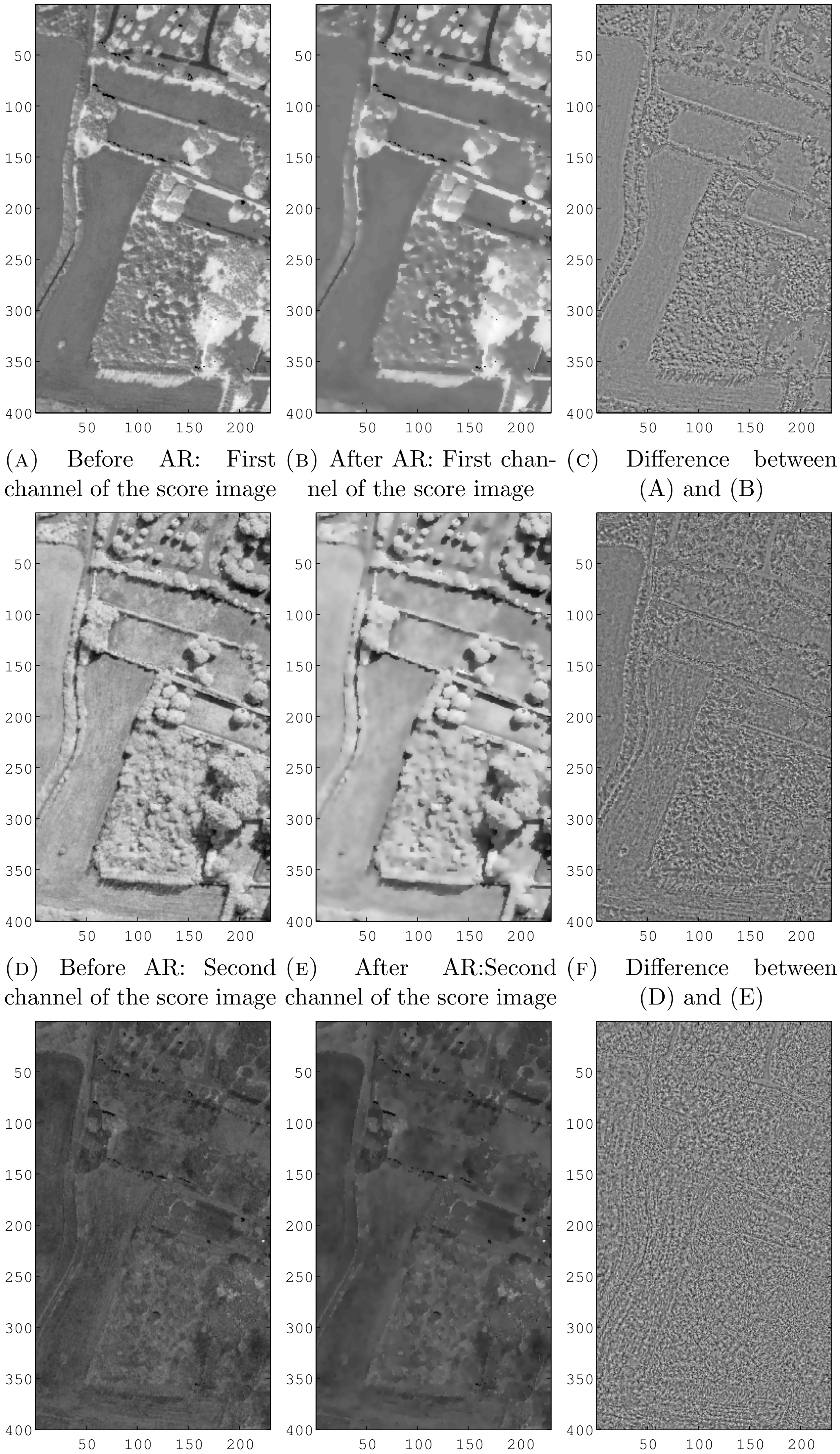
(C) Difference between (A) and (B)

(D) Before AR: Second channel of the score image

(E) After AR: Second channel of the score image

(F) Difference between (D) and (E)

Figure 4.19: Dataset RA1: Effect of the regularization in the spatial domain



(A) Before AR: First channel of the score image

(B) After AR: First channel of the score image

(C) Difference between (A) and (B)

(D) Before AR: Second channel of the score image

(E) After AR:Second channel of the score image

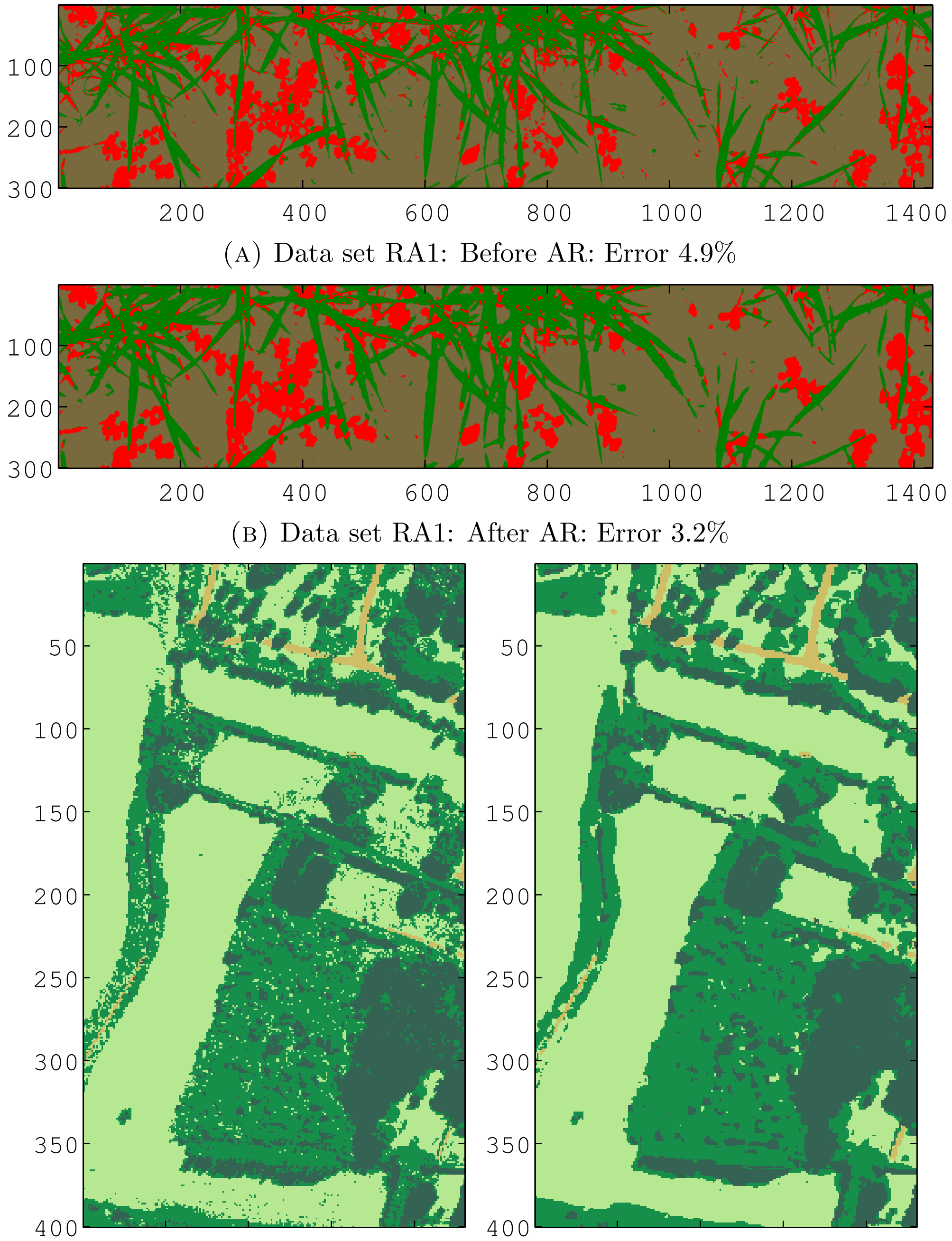
(F) Difference between (D) and (E)

(G) Before AR: Third channel of the score image

(H) After AR: Third channel of the score image

(I) Difference between (G) and (H)

Figure 4.20: Dataset RA1: Effect of the regularization in the spatial domain



(A) Data set RA1: Before AR: Error 4.9%

(B) Data set RA1: After AR: Error 3.2%

(C) Data set RB1: Before AR: Error 11%

(D) Data set RB1: After AR: Error 7%

FIGURE 4.21: Classification results before and after regularization on the two data sets

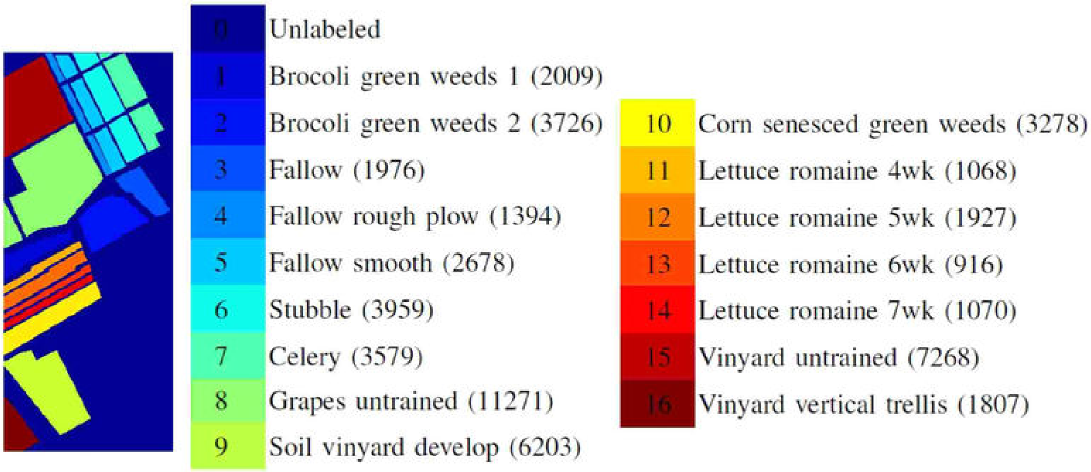

0 Unlabeled
1 Brocoli green weeds 1 (2009)
2 Brocoli green weeds 2 (3726)
3 Fallow (1976)
4 Fallow rough plow (1394)
5 Fallow smooth (2678)
6 Stubble (3959)
7 Celery (3579)
8 Grapes untrained (11271)
9 Soil vinyard develop (6203)
10 Corn senesced green weeds (3278)
11 Lettuce romaine 4wk (1068)
12 Lettuce romaine 5wk (1927)
13 Lettuce romaine 6wk (916)
14 Lettuce romaine 7wk (1070)
15 Vinyard untrained (7268)
16 Vinyard vertical trellis (1807)

FIGURE 4.22: Ground truth map with class labels of the Salinas data set.

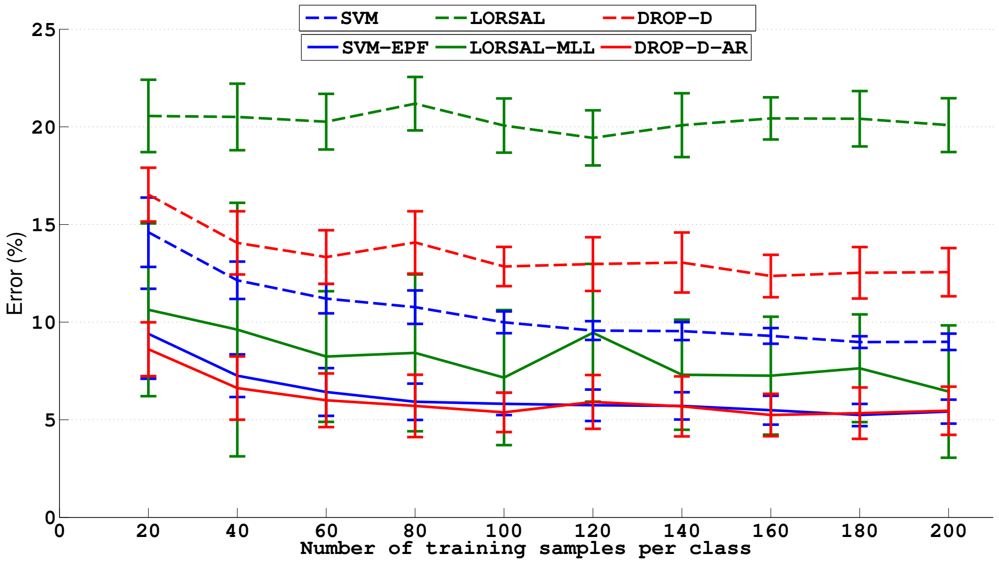


FIGURE 4.23: Comparison of classification methods for an increasing number of training samples using Salinas data set. Curves represent mean values obtained for a random selection repeated 30 times and error bars represent the standard deviations.

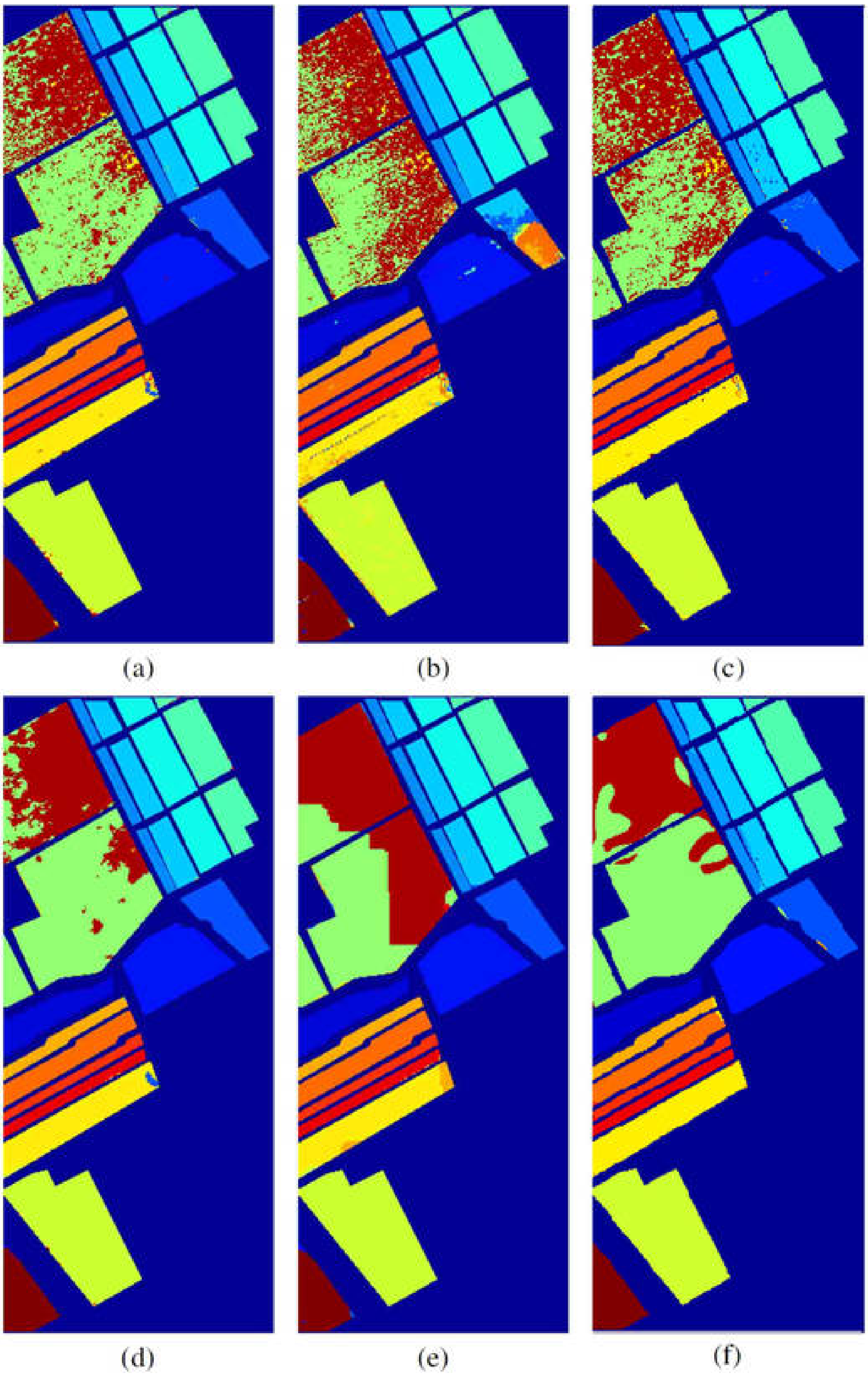


FIGURE 4.24: Classification maps of Salinas data set with 100 randomly selected training samples per class for (a) SVM, (b) LORSAL, (c) DROP-D, (d) SVM-EPF, (e) LORSAL-MLL, (f) DROP-D-AR.

We thus compared our approach with two of the leading state-of-the-art spectral-spatial methods. Every method presented here uses spatial information in addition to a purely spectral classification method. Therefore, in order to assess the synergy of the conjoint use of spectral and spatial information, we also present the results obtained without using spatial information.

We thus compare Support Vector Machine (SVM), SVM with Edge Preserving Filtering (SVM-EPF) [Kang et al., 2014], Logistic Regression via Splitting and Augmented Lagrangian (LORSAL) [Li, 2011], LORSAL Multilevel Logistic (LORSAL-MLL) [Li et al., 2012], and our methods DROP-D and DROP-D-AR.

We first compare these six methods in terms of classification error for an increasing number of training samples per class. Because classification results highly depend on the choice of the training set (especially for small sets), the selection was performed randomly and repeated 30 times for each number of training samples. Then, for each selection and each number of training samples, every model was trained and tuned using cross-validation. The classification error was evaluated on the samples that were not used for training. Figure 4.23 shows the results obtained with this data set and the six tested classification methods, where solid lines are used for methods using both spectral and spatial information, and dashed lines are used for methods using only spectral information. Our method proved to use effectively the spatial information: SVM gives better results than DROP-D, while DROP-D-AR is equivalent to SVM-EPF.

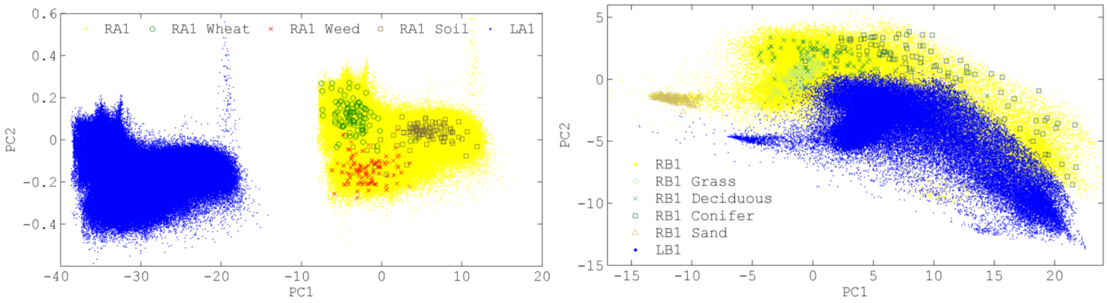
We also compare classification maps obtained by the six methods using 100 randomly selected samples per class in Figure 4.24. We observe that using spatial information improved the classification results at least by reducing the classification noise. Interestingly, both SVM-EPF and DROP-D-AR-based methods obtained well defined borders and few misclassified regions. LORSAL-MLL uses a graph cut technique, resulting in whole misclassified areas when the cut is not properly made, which explains the large variability observed in Figure 4.23. SVM-EPF increases the SVM classification rate by using neighboring information provided by a spatially regularized map (using PCA) in a majority of voting fashion. Therefore, if most pixels are misclassified within a region, after EPF, the region might be homogeneous but possibly with a wrong class label. With DROP-AR, however, since spatial information is used before class decision, some areas that are completely misclassified before regularization can be correctly classified afterwards (see the yellow region in the center of the Figure 4.24 for example) .

4.4 Reflectance correction

In this section we aim at proving that by using the log-radiance and by correctly estimating the translation in the feature space between images, radiance images can be used, avoiding the reflectance correction as detailed in the previous chapter. To do so, we first illustrate the translation occurring between the reflectance and radiance image in the score space for both data sets. We also show the translation occurring between the radiance images of our data set. Then, using the method described in the previous chapter, we prove that the translation can be estimated even when there is a missing class in the data set. We finally show the classification maps obtained with the registered scores and compared them with those obtained with reflectance images.

4.4.1 Reflectance correction effect on the reduced scores

In order to illustrate the translation occurring in the log-space between the reflectance spectra and the radiance spectra, we represented in Figure 4.25 the effect of reflectance correction on log-image. The models calibrated using the log-reflectance data RA1 and RB1 are applied to the log-radiance data LA1 and LB1. For both data sets, the differences clearly correspond to translations. There is therefore no loss of class separability when using the radiance image instead of a reflectance one. However, without a correct estimation of the translation, the classification model cannot be applied directly.



(A) Translation observed between RA1 and LA1

(B) Translation observed between RB1 and LB1

FIGURE 4.25: Effect of the reflectance correction on the score space after log-transformation. For both data sets, scores 1 and 2 from the DROP-D models are plotted. The model is calibrated with a log-reflectance image (yellow) and applied to a log-radiance image (blue). The training set is also represented.

4.4.2 Using log-radiance image for classification

The translation illustrated between reflectance and radiance image is very large. When dealing with radiance images, depending on weather conditions and hour of the day, the differences in radiance images can be a lot smaller.

For example in Figure 4.26 we show the differences between the two images of data set A that were acquired with one hour difference. Recall that A2 and A3 come from the same image and thus the same translation versus A1 is obtained. Note the class 'weed' (red) missing for data set 3 in the scatter plot. Because this translation is mostly along the soil-vegetation separation axis (PC1), classes tend to be more easily classified as wheat. As a result, the obtained classification map have green pixels even for the soil (which is normally easily discriminated).

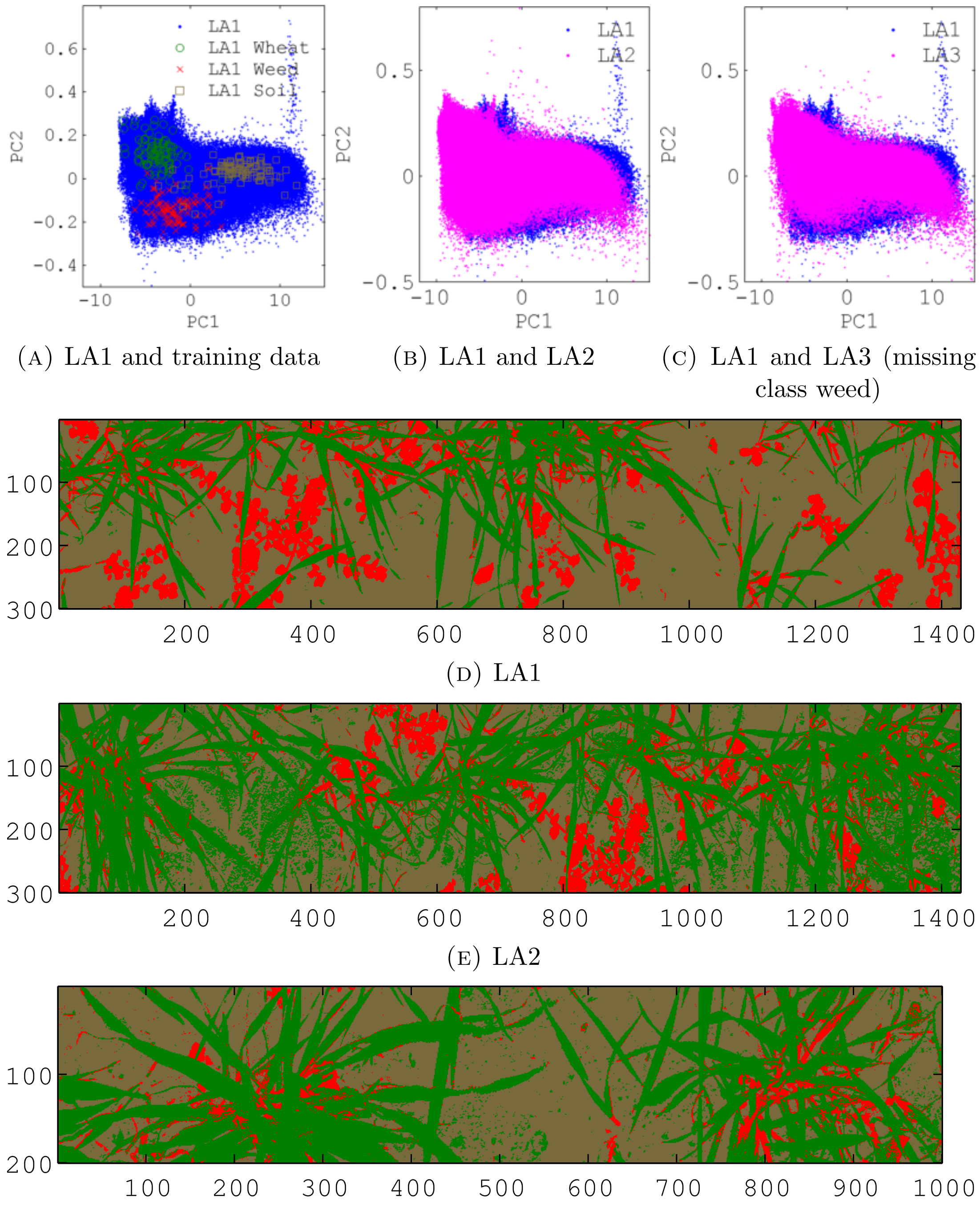
In Figure 4.27 we similarly show the translation occuring between the log-radiance image LB1 and LB2. With this data there are only two images but the optimal number of discriminant DROP-D axes was 3. Here, pixels in the classification maps are shifted more toward the 'deciduous' class.

4.4.3 Translation estimation

In this section, we illustrate the different steps of the approach proposed in the previous chapter in order to estimate the translation. Every step of the method is represented for both data sets in Figure 4.28 and Figure 4.29 for data sets A and B respectively.

For data set A, the estimation of each class is correctly performed by the EM algorithm. For LA3, the missing class was not problematic for EM which found a third class with a large variance and center next to the soil mean.

Then, the cross correlation between LA1 and the two other images lead to a unique maximum in each case. Because LA2 and LA3 come from the same image, the same lighting difference should be found with LA1. On these data, which were represented on a 200 × 200 pixel image, the difference between LA2 and LA3 translation was only of 4 pixels horizontal and 1 pixel vertical. This shows the robustness of the method even in presence of a missing class.



(A) LA1 and training data (B) LA1 and LA2 (C) LA1 and LA3 (missing class weed)

(D) LA1

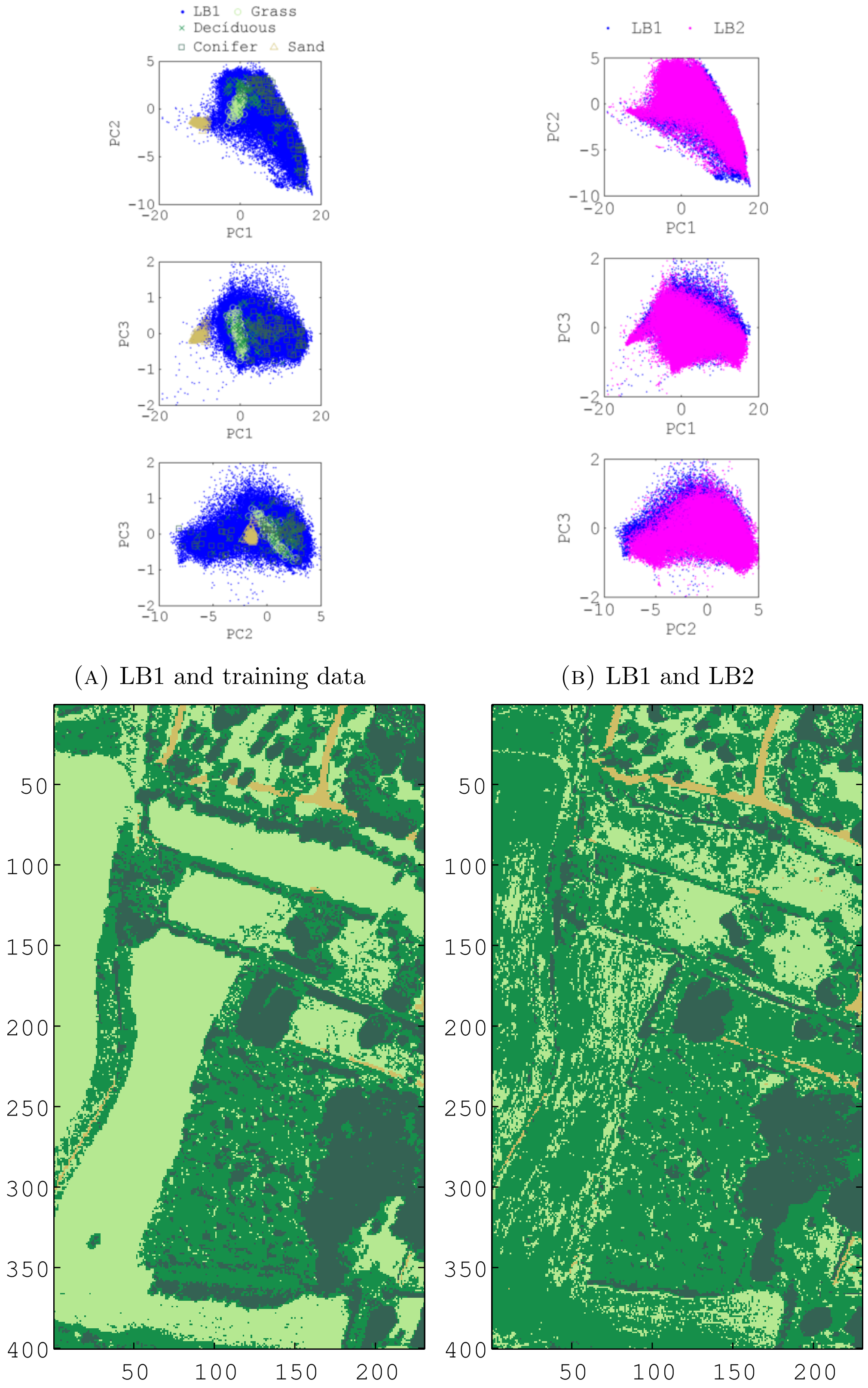
(E) LA2

(F) LA3 (missing class weed)

FIGURE 4.26: Scores plot and classification map calibrated with LA1

Observing the translation directly on the score plot shows satisfying results as the cloud points are now well overlapping. The resulting classification map shown in Figure 4.30 gives similar results as those obtained with the classical reflectance correction technique.

For data set B, the class estimation is less obvious because of the class 'sand' which is represented in only a very limited amount in the images. However, because the method is robust to missing classes, the further processing still proves to perform accurately to estimate the translation. The obtained cross-correlation presents a few local minima but the correct estimation of the translation can still be performed without any doubt by choosing the maximum. As with data set A, the registered score correctly overlaps the original ones.



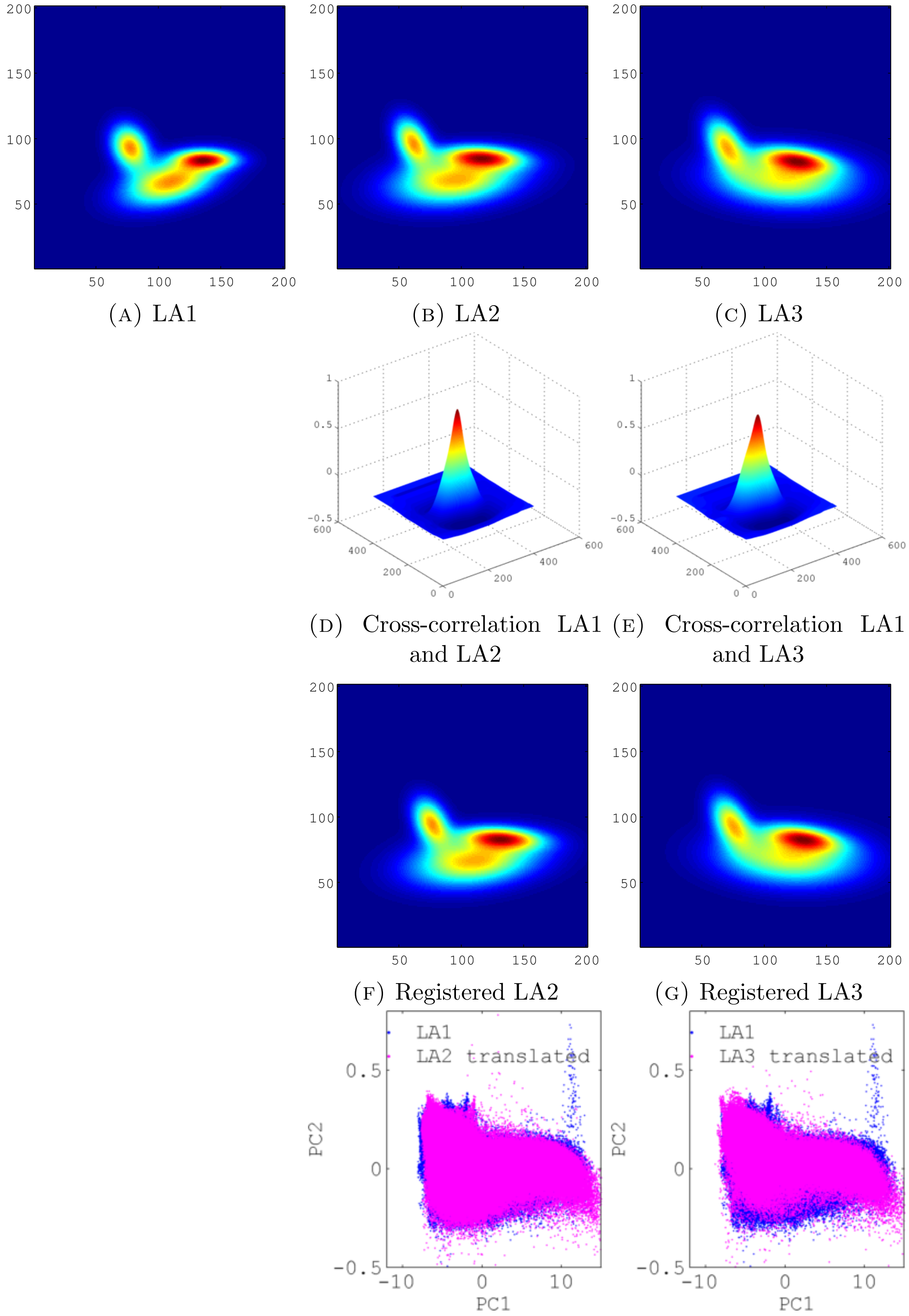
(A) LB1 and training data

(B) LB1 and LB2

(C) LB1

(D) LB2

Figure 4.27: Scores plot and classification map calibrated with LB1



(A) LA1

(B) LA2

(C) LA3

(D) Cross-correlation LA1 and LA2

(E) Cross-correlation LA1 and LA3

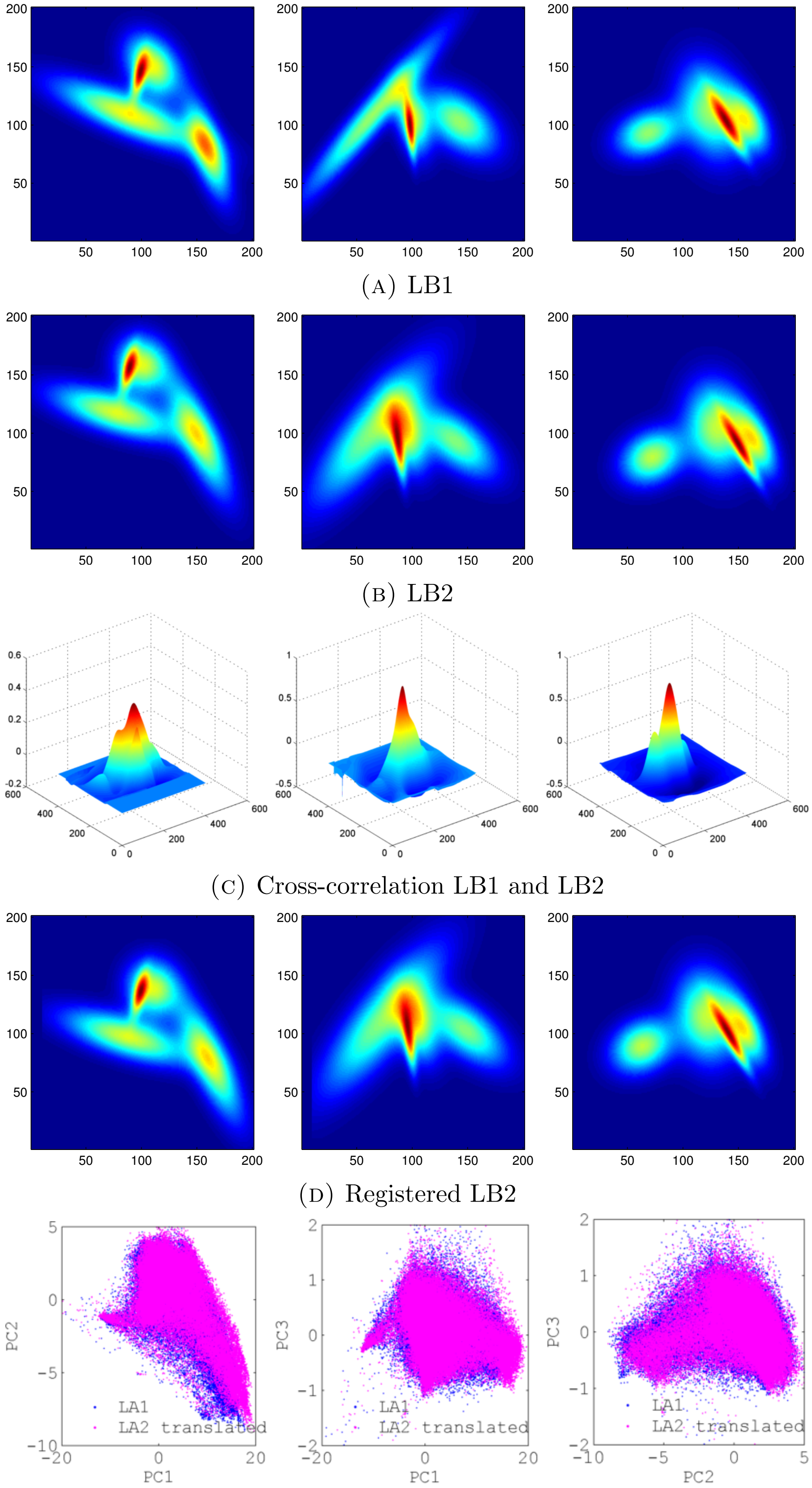
(F) Registered LA2

(G) Registered LA3

(H) LA1 and translated LA2 scores

(I) LA1 and translated LA3 scores

FIGURE 4.28: Registration process on LA data set. A,B,C: distribution density functions estimate by EM. F,G: distribution density functions after translation correction.



(A) LB1

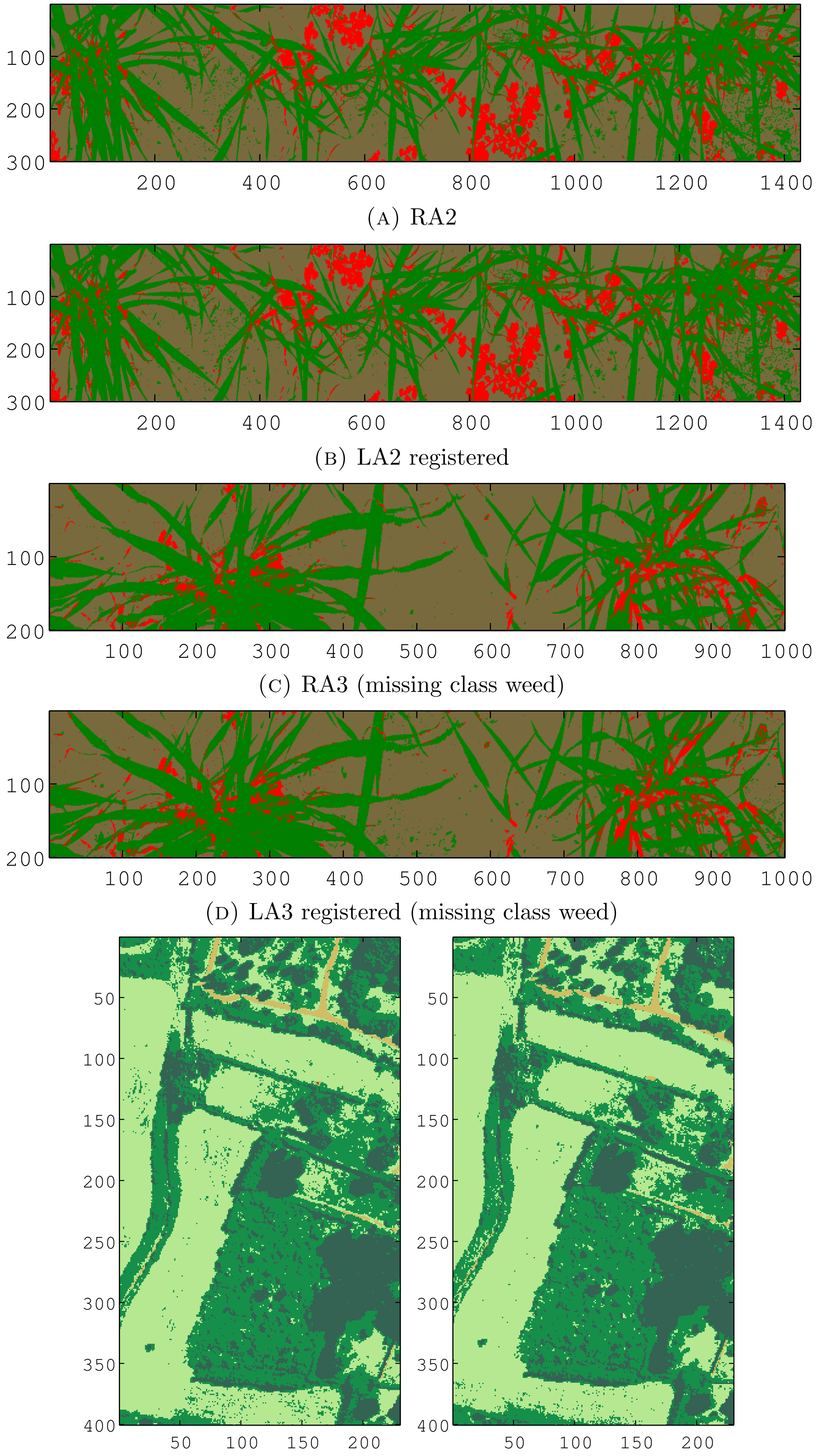
(B) LB2

(C) Cross-correlation LB1 and LB2

(D) Registered LB2

(E) LB1 and translated LB2 scores

FIGURE 4.29: Registration process on LB data set. From left to right axis (1,2), axes (1,3) and axes (2,3)



(A) RA2

(B) LA2 registered

(C) RA3 (missing class weed)

(D) LA3 registered (missing class weed)

(E) RB2

(F) LB2 registered

FIGURE 4.30: Comparison of the classification results obtained on the reflectance images and on the translated radiance images.

Chapter 5

Conclusions and future work

The objective of this thesis was to propose and validate new approaches to deal with some of the main issues in supervised classification of hyperspectral data. The focus was on three particular aspects: (1) spectral dimension reduction, (2) combination of spectral and spatial information and (3) compensation for variability in lighting conditions. In the next section we summarize the main contributions of the thesis. Then, we propose some research directions in order to continue and improve this work.

5.1 Conclusions

Hyperspectral image processing has been more and more used in many scientific and industrial fields over the last decades. Its growing interest comes from the possibility to obtain detailed spectral information for each pixel of the image. Using this spectral information, which is linked to the biochemical properties of the target, many useful characteristics can be retrieved regarding the imaged objects. HS images can therefore be used for environmental applications, earth monitoring, plant content mapping or even weed detection.

However, the counterpart of this very detailed spectral information is that the huge amount of data to process, in order to retrieve these characteristics, makes the usual processing techniques fail.

For example, in supervised classification, which is one of the main uses of HS imaging, the high dimensionality of the data leads to the failure of standard classifiers.

In addition, spectral data are highly correlated, which creates other conditioning problems for matrix computations. Fortunately, high collinearity also means high redundancy. Hence, methods that can summarize the spectral information have been investigated in order to deal with this type of data. For instance, in supervised classification, different methods have been proposed to reduce the dimensionality of the data. One classical approach is to model the class structure using statistical learning techniques such as Partial Least Squares. Although very effective, this method is prone to overfitting and therefore needs extra data to be collected to compensate for this issue. Another classical way of finding the 'best' linear subspace is to use the Fisher approach, that is to minimize the Wilk's Lambda by minimizing the within-class variability and maximizing the distance between classes. Fisher's paradigm is however not directly applicable to high dimensional and collinear data because of matrix inversion issues. Much research has thus been conducted to adapt Fisher's idea for high dimensional spaces. Among these methods, Nullspace LDA offers an interesting way to solve this issue but is dependent on the existence of this nullspace, which becomes empty as the number of observations increases. Another way is to perform a Principal Component Analysis prior to the Fisher LDA so that in the reduced space, the inversion problem is avoided. However, this method uses in its first step all the data information without considering class information in the dimension reduction and is therefore not optimal for classification purposes. In this thesis, we propose an alternative approach that uses orthogonal projection to clean the data before dimension reduction. The data cleaning is performed using the within-class principal directions. In that sense, it mimics the LDA, but instead of weighting the projection by the within-class inversion, it directly removes the information due to this within class-variations. We also show that without being very careful when removing the within-class information, the class separability can be lost because of non-orthogonality of the within- and between-class principal directions. Therefore, in the method we proposed, called DROP-D, a first step consists in preserving the most important between-class directions so that no cleaning can be performed on them. Once the data is cleaned, a classical Principal Component Analysis is performed in order to provide reduced data. This method provides similar results to PLS in terms of classification performances. However, contrary to PLS, by the nature of the method, overfitting can be prevented without using the cross-validation procedure. Indeed, by cleaning the data instead of learning a class structure, DROP-D classification results are affected if useful information is removed even during the training phase.

Another issue tackled in this thesis is the use of both spectral and spatial information to enhance classification performances. Contrary to pure spectrometric applications, the contextual spatial information provided by the hyperspectral image should not be ignored in the classification process. In this context, the hyperspectral community has recently been developing many different approaches in order to combine these complementary types of information. Among them, edge preserving filtering techniques have received much attention due to their ability to reduce spatial noise within homogeneous objects while preserving their borders. The HS image quality is thus improved for visual analysis and for classification performance owing to the reduced noise. However, the existing approaches either use this spatial filtering directly on the hyperspectral image, and thus need to redefine high-dimensional gradients, or use the regularization on reduced variables obtained from unsupervised dimension reduction methods. In both cases, the variables used for regularization contain both the within-class natural variations as well as the information due to class differences. Hence, spatial filtering is not optimal because the natural variability creates edges in the reduced image that are, by definition, preserved by the EPF. To compensate for this issue, we proposed in this thesis to use a supervised approach for dimension reduction before applying the spatial EPF. We show that, by having edges that are mostly due to the objects borders, within-object smoothing is increased. Therefore, it results in increased classification performances when compared with the other approaches.

Finally, the last issue tackled in this thesis concerns the reflectance correction necessary to model the transfer between images. In order to process data that do not vary with atmospheric and lighting conditions, hyperspectral images have first to be calibrated into reflectance images. This operation requires the measure of lighting conditions in each image and is thus constraining for some applications. For instance, at Irstea, a calibrated reference surface is positioned at each acquisition in the field of view of the camera. Then, because the reflectivity of this surface is known, the lighting condition on the scene can be estimated and the image corrected. Other techniques exist, but also require the measurement or the estimation of the light received by the objects at the moment of image acquisition. In this thesis, we show that in the framework of supervised classification this effect can be corrected automatically without prior information if the surfaces are Lambertian. In particular, we show that the difference in lighting can be expressed, after logarithm pre-treatment, as an additive effect which is constant for each pixel within an image. This effect remains additive after using linear dimension reduction method such as DROP-D and can thus be estimated in the low-dimensional space. We thus propose a method to estimate this translation that is based on class density estimation in the reduced space. Owing to the use of a supervised dimension reduction, the classes form clusters in this reduced space which can be retrieved by automatic methods such as Expectation Maximization. Once the class distribution is modeled, we propose to use a classical image registration technique (cross-correlation) in order to estimate this translation, which is robust to missing classes. On the two data sets presented in this thesis, the obtained results offered by this method are comparable to those obtained with a classical reflectance correction technique.

These three approaches can also be combined in a general hyperspectral processing framework: DROP-D can be first used as a spectral supervised dimension reduction method. Then, on the obtained scores, a spatial regularization technique that preserves spatial borders can be effectively applied. Finally, without prior reflectance correction, other HS images can be classified as well using the score translation method.

5.2 Future work

In this thesis, we have tackled three main issues of hyperspectral image classification. For each proposed approach, different research directions can be taken and improvements can be made.

In the actual implementation of DROP-D, the eigenvectors associated to the largest eigenvalues are preserved. However, it would be interesting to chose some combinations of between and within axis to be kept or removed, not necessarily starting from the main ones. Although obvious when there are few classes, it would require optimization techniques to be implemented as the number of possible combinations increases.

Another approach would be to relax the orthogonality constraint on the projection. It is expected that non-orthogonal projection would be able to reduce better the within class variability without affecting the between class distances.

Another possibility with DROP-D would be to use the cleaned data with some more complex classifier such as SVM in order to reduce the amount of support vectors to use.

Finally, from a theoretical point of view, it would be interesting to compare the cleaning made by DROP-D with the chemometric methods developed for regression analysis such as Orthogonal Signal Correction for example.

We implemented the spectral-spatial approach using the original version of Anisotropic Diffusion in conjunction with the supervised dimension reduction. Hence, as the image processing community has worked extensively in the field of spatial regularization, investigating more sophisticated approaches would surely be beneficial. Also, observing the residual maps (differences before and after regularization), some obtained patterns in terms of textural analysis seem of potential interest to keep enhancing the use of spatial information in the classification process. These textural features could be added to the process as input of a classifier in the same way as spectral data.

Finally, with the translation estimation, several improvements can be made in the approach. For instance, as a first step, alternatives to the very costly cross-correlation to find a translation would be beneficial (using Fourier Transform for example). Then, using the spatial information (i.e., shapes in the classification map as a feedback) in addition to the correlation information should help to find the translation when classes do not cluster well in the reduced space. Finally, solutions to deal with non-Lambertian objects would be highly beneficial since even reflectance correction cannot accurately deal with this particular case. We do believe that this approach using a logarithm transformation as a pre-processing could also be applicable to deal with non-Lambertian cases. In this case each class would translate independently from the other, but with the same covariance matrix. Thus, using non-rigid registration techniques by matching the covariance could possibly be performed on some cases.

Appendix A

Contributions

A.1 Journal

Xavier Hadoux, Nathalie Gorretta, Jean-Michel Roger, Ryad Bendoula, and Gilles Rabatel. *Comparison of the efficacy of spectral pre-treatments for wheat and weed discrimination in outdoor conditions.* Computer and Electronics in Agriculture, 2014.

Xavier Hadoux, Sylvain Jay, Gilles Rabatel and Nathalie Gorretta. *A spectral-spatial approach for hyperspectral image classification using spatial regularization on supervised score image.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014.

Sylvain Jay, Gilles Rabatel, **Xavier Hadoux**, Daniel Moura and Nathalie Gorretta. *In-field crop row phenotyping from 3D modeling performed using structure from motion.* Computer and Electronics in Agriculture, 2014.

Sylvain Jay, **Xavier Hadoux**, Nathalie Gorretta, Fabienne Maupas and Gilles Rabatel. *Leaf nitrogen content mapping in sugar beet using close-range hyperspectral imaging under field conditions.* Field crop research [submitted 2014].

Xavier Hadoux, Jean-Michel Roger, Gilles Rabatel and Rutledge Douglas. *DROP-D : Dimension reduction by orthogonal projection for discrimination.* [To submit in Chemometrics and Intelligent Laboratory Systems (2014)].

Xavier Hadoux, Nathalie Gorretta, Marc Sarossy and Dinesh Kant Kumar. *A spectral-region selection for multi-variate regression and discrimination.* [To submit in Chemometrics and Intelligent Laboratory Systems (2014)].

A.2 Conference

Marc Sarossy, **Xavier Hadoux**, and Jess Tang. *Using statistical decomposition to improve the identification of the photopic negative response of the erg.*In European Association for Vision and Eye Research EVER congress, Nice, France, October 2014.

Adrian Bingham, **Xavier Hadoux**, and Dinesh Kant Kumar. *Implementation of a safety system using ir and ultrasonic devices for mobility scooter obstacle collision avoidance.* In 5th ISSNIP-IEEE Biosignals and Biorobotics Conference, Salvador, Brazil, June 2014.

Nathalie Gorretta, Sylvain Jay, and **Xavier Hadoux**. *Classification spectrale-spatiale d'images hyperspectrales par régularisation anisotropique.* In 3ème colloque scientifique de la Société Française de Télédétection Hyperspectrale, Porquerolles, France, May 2014.

Sylvain Jay, **Xavier Hadoux**, Nathalie Gorretta, and Gilles Rabatel. *Potentiel de l'imagerie hyperspectrale de proxy-détection pour l'estimation et la cartographie de la teneur en azote dans les feuilles de betteraves sucrières.* In 3ème colloque scientifique de la Société Française de Télédétection Hyperspectrale, Porquerolles, France, May 2014.

Sylvain Jay, **Xavier Hadoux**, Nathalie Gorretta, and Gilles Rabatel. *Potential of hyperspectral imagery for nitrogen content retrieval in sugar beet leaves.* In International Conference on Agricultural Engineering (AGENG), Zurich, Proceedings, July 2014.

Xavier Hadoux. *Using spectral/spatial information in hyperspectral image processing : application on weed/crop discrimination.* In Séminaire UMR ITAP, Montpellier, France, March 2014.

Nathalie Gorretta and **Xavier Hadoux**. *Prise en compte de la dimension spatiale des données dans le traitement des images hyper-spectrales: état des lieux.* In

14èmes Rencontres HélioSPIR "Spectrométrie proche infrarouge et hétérogénéité", Agropolis International, Montpellier, France, November 2013.

Xavier Hadoux, Nathalie Gorretta, and Jean-Michel Roger. *Réduction de dimension pour la classification spectrale.* In Conférence Francophone de Chimiométrie, Brest, France, Septembre 2013. [**Best junior presentation award**]

Xavier Hadoux, Nathalie Gorretta, Jean-Michel Roger, Olivier Strauss, and Gilles Rabatel. *Spectral spatial pre-processing using multi-resolution 3d wavelets for hyperspectral image classification.* In IEEE 5th workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing, Gainesville, Florida, June 2013.

Xavier Hadoux, Nathalie Gorretta, and Gilles Rabatel. *A comparative study of spectral pre-processing for supervised classification : application on weeds-wheat discrimination using hyperspectral imagery.* In 16th International Conference on Near Infrared Spectroscopy, La Grande Motte, France, June 2013.

Nathalie Gorretta, **Xavier Hadoux**, T. Brunel, Christophe Guizard, and A. Paulhe- Massol. *Wheat kernel vitreousness classification by hyperspectral imaging andspectral-spatial data analysis.* In 16th International Conference on Near Infrared Spectroscopy, La Grande Motte, France, June 2013.

Xavier Hadoux, Nathalie Gorretta, and Gilles Rabatel. *Weeds-wheat discrimination using hyperspectral imagery.* In International Conference of Agricultural Engineering, Valencia, Spain, July 2012.

Appendix B

Résumé Étendu

Contents

Introduction

Une caméra hyperspectrale (HS) peut enregistrer des images avec une information spectrale très détaillée pour chaque pixel. L'information spectrale fournie par ces capteurs est liée aux propriétés biochimiques de l'échantillon mesuré, et a donc été largement utilisée pour la mesure non-destructive dans les domaines scientifiques et industriels ces dernières décennies. À Irstea, et en particulier dans l'unité de recherche ITAP (Information, Technologies, Analyse environnementale, Procédés agricoles), cette information spectrale spatialisée permet d'augmenter les possibilités de caractérisation déjà offertes par les spectromètres et les caméras couleurs classiques pour les applications environnementales et agricoles. En effet, si la dimension spectrale fournit une source d'information sur l'état de la récolte (physiologique ou pathologique, par exemple), l'information spatiale permet de récupérer des informations de structure (phase de développement, présence d'adventices, etc.). La mise en œuvre de la technologie HS et le traitement des données obtenues sont cependant complexes et nécessitent des procédures adaptées.

Dans le cadre de la classification, les différences biochimiques entre pixels spectraux peuvent être exploitées pour créer un modèle de classification qui permet d'affecter chaque pixel de l'image HS à une catégorie unique. En classification supervisée, les échantillons d'apprentissage de catégories connues sont nécessaires afin de définir la règle d'affectation.

Au début de ce travail de thèse, le contexte spécifique de la discrimination des adventices dans les cultures de blé a été étudié. En particulier, la comparaison de différents pré-traitements spectraux en regard des méthodes de classification a été proposée dans le cadre de la proxi-détection au champ. Toutefois, il a été constaté que certains des problèmes rencontrés pourraient être abordés de manière plus générique.

Par conséquent, par la suite, des questions plus générales concernant la classification supervisée des images hyperspectrales ont été étudiées, et constituent l'essentiel du présent document. Trois contributions principales sont notamment développées. La première est une nouvelle méthode supervisée de réduction de dimension, développée spécifiquement pour faire face à la dimension élevée et à la colinéarité des données spectrales. La seconde est une approche spectro-spatiale qui consiste à utiliser une méthode de régularisation spatiale en combinaison avec une réduction supervisée de la dimension, dans le but d'optimiser son effet sur les performances de classification. La troisième et dernière est une méthode automatique qui permet à des images HS en radiance d'être classifiées même avec des conditions différentes d'éclairage, et évite donc la correction en réflectance.

Notons que d'autres méthodes ont également été étudiées, que nous avons choisi de ne pas inclure dans le présent document. En particulier, une collaboration avec le Pr. Dinesh Kant Kumar et le Dr. Marc Sarossy du Royal Melbourne Institute of Technology, en Australie, a été accomplie afin de développer une approche multi-résolution pour l'analyse spectrale. Toutes ces contributions sont disponibles en annexe A.

Ce manuscrit de thèse est organisé en cinq chapitres. Le premier chapitre présente le contexte de la classification d'images hyperspectrales. La description ainsi que les plus importantes définitions concernant les images hyperspectrales y sont données. Le contexte spécifique de la classification supervisée y est ensuite détaillé. À la fin de ce chapitre, les principaux enjeux concernant l'application des méthodes de classification aux images hyperspectrales sont résumés, à savoir la dimension élevée et la colinéarité des données spectrales, la façon d'introduire l'information spatiale dans le processus de classification et les principales étapes de correction pour obtenir une image HS en réflectance. Le deuxième chapitre dresse un état de l'art des méthodes principales qui répondent aux problèmes mentionnés précédemment. Il permet enfin de statuer sur les inconvénients des approches existantes. Dans le troisième chapitre, nous proposons trois contributions originales pour répondre à ces questions. Le quatrième chapitre est consacré à la validation des approches proposées par la présentation de résultats en utilisant des exemples réels. Le cinquième chapitre conclut cette thèse en synthétisant les points les plus importants des approches développées et propose quelques perspectives ainsi que de futures directions de recherche pour continuer ce travail.

B.1 Classification en imagerie hyperspectrale

L'imagerie hyperspectrale, également connue sous le nom d'imagerie chimique ou encore de spectro-imagerie, est une technologie d'imagerie relativement récente qui permet à la fois l'acquisition de l'information spectrale et spatiale des objets ciblés. Les images hyperspectrales (HS) sont des images multivariées qui peuvent être représentées comme des cubes de données, avec deux dimensions

spatiales (x, y) et une dimension spectrale (λ). Chaque pixel de l'image contient une mesure spectrale échantillonnée, qui peut être interprétée pour identifier les matériaux présents dans la scène. Cette représentation est généralement perçue de deux manières équivalentes :
- Du point de vue de la spectrométrie, le contenu de l'image HS est considéré comme de l'information spectrale spatialisée : les spectromètres deviennent spatialement résolus.
- Du point de vue du traitement de l'image, le contenu de l'image de HS est considéré comme de l'information spatiale spectralisée : les pixels d'une image deviennent spectralement résolus.
Dans les deux cas, chaque position spatiale dans l'image HS est associée à un spectre qui contient l'information chimique de l'objet imagé.

L'objectif de la **classification** est d'identifier la nature des objets en termes de classes, sur la base de certaines caractéristiques [Bishop, 2007, Fukunaga, 1990]. En classification supervisée, toutes les classes sont supposées être connues et mutuellement exclusives. Quelques observations pour chaque classe sont également supposés être disponibles pour étalonner un modèle. Ces observations, qui forment ce qu'on appelle *les échantillons d'apprentissages*, sont attribuées manuellement, et nécessitent l'établissement préalable d'une vérité terrain.
Avec une image HS, les caractéristiques peuvent prendre différentes formes, par exemple spectres bruts, variables spectrales réduites, formes des objets, textures.

Définissons un espace des caractéristiques $\mathcal{X} \in \mathbb{R}^P$ et un jeu fini des classes possibles $\mathcal{Y} = \{\mathcal{Y}_1, \cdots, \mathcal{Y}_C\}$, où $\mathcal{Y}_c$ représente l'une des C classes. Les N observations de l'ensemble d'apprentissage sont regroupées dans une matrice $\mathbf{X} = \{\mathbf{x}_i \in \mathcal{X}\}$ et les classes associées dans $\mathbf{Y} = \{\mathbf{y}_i \in \mathcal{Y}\}$, où $i = 1, \cdots, N$. Avec cette notation, $\mathbf{x}_i$ correspond au i^{eme} vecteur et $\mathbf{y}_i$ à sa classe. Les classes sont généralement notées en codage disjonctif, par exemple $\mathbf{y} = [0\ 0\ 1\ 0\ 0]^\intercal$ code la classe 3 parmi 5.
La classification consiste à assigner chaque vecteur à l'une des C classes d'intérêt en utilisant une fonction $g : \mathcal{X} \mapsto \mathcal{Y}$.

B.2 Problématiques et état de l'art en classification des données HS

Le type et la quantité d'informations fournies par les capteurs HS doivent être considérés lors de la mise en place d'une procédure de classification. En effet, bien que le grand nombre de bandes spectrales fournies par la caméra HS signifie également plus d'information potentiellement discriminatoire, cela pose également des problèmes

B.2.1 Problèmes avec la dimension spectrale

Il existe plusieurs problèmes liés à l'utilisation des données spectrales à des fins de classification, qui sont dues au fait que nous essayons de modéliser une structure de faible dimension contenue dans un espace de grande dimension et en utilisant seulement quelques observations [Donoho, 2000, Jimenez et al., 1998, Tormod and Bjorn-Helge, 2001]. **La concentration de la mesure** stipule que les régions d'un espace de grande dimension sont presque vides parce que les données ont tendance à se concentrer dans une couche mince à la frontière des régions. Chaque voisinage des observations dans l'espace des caractéristiques est donc susceptible d'être vide. Par conséquent, les estimations de densités statistiques doivent être réalisées en utilisant une large bande passante et donc en perdant les détails spectraux. **Les estimations statistiques** nécessitent un nombre croissant d'échantillons d'apprentissage lorsque la dimensionnalité des données augmente [Hughes, 1968]. **La colinéarité** entre les variables est un problème bien connu avec des données spectrales. Ce problème, lié au conditionnement des matrices, est dû à la très forte inter-corrélation des variables spectrales mesurées.

Toutefois, du fait que les espaces de grande dimension sont presque vides, une structure de dimension inférieure contenant la même quantité d'information est susceptible d'exister. Pour compenser ces problèmes, une approche classique consiste donc à effectuer une *réduction de dimension* avant la classification [Geladi, 2003, Wold et al., 2001]. Pour la réduction de dimension spectrale, lorsque l'objectif est la classification, les méthodes non supervisées conduisent à des scores sous-optimaux car ne prenant pas en compte l'information de classe lors du processus de réduction [Barker and Rayens, 2003].

Parmi les approches supervisées, les méthodes type moindres carrés partiels (PLS) [Bylesjö et al., 2006, Fearn, 2000, Trygg and Wold, 2002, Wold et al., 2001] tendent à modéliser la structure de classe des données en maximisant la capture de covariance entre les variables et les classes lors de la construction des scores. Ces approches ont donc naturellement tendance à "sur-apprendre" et leurs paramètres doivent être réglés en utilisant des procédures de validation croisées [Esbensen and Geladi, 2010]. Sans prendre de précaution particulières, ces validations croisées peuvent conduire à des résultats trop optimistes et peuvent même trouver une structure de classe quand il n'y en a pas.

D'autre part, des méthodes type analyse discriminante de Fisher (LDA) [Guo et al., 2006, Witten and Tibshirani, 2011, Ye et al., 2005] tendent à résoudre le problème d'inversion de la matrice de covariance en utilisant des astuces mathématiques comme la pseudo-inverse ou l'inversion de la matrice de variance totale à la place de la matrice de variance intra-classe. Une autre méthode consiste à utiliser une analyse en composantes principales (PCA) [Grahn and Geladi, 2007] comme une première étape en vue d'obtenir moins de variables sur lesquelles une LDA peut ensuite être effectuée. Cette approche est cependant sous-optimale, car la première étape sélectionne des composantes qui ne sont pas liées à des différences entre classes. Enfin, nullspace LDA [Chen et al., 2000] est une méthode mathématiquement très prometteuse car répondant parfaitement au paradigme de la LDA. Cependant, cette méthode nécessite que le noyau de la matrice de covariance intra-classe existe, ce qui n'est le cas que lorsque le nombre de variables est plus grand que le nombre d'observations. Cela signifie qu'en cas de nouvelles observations acquises pour améliorer un modèle, la méthode ne peut plus être utilisée, limitant son champ d'applications pour les données HS.

Dans cette thèse, nous proposons une démarche dans laquelle, contrairement à la nullspace LDA, la suppression de la variabilité intra-classe est contrôlée, ce qui permet également de préserver explicitement les axes discriminants les plus importants.

B.2.2 Utilisation de l'information spatiale

Les classifieurs décris précédemment ne traitent les données HS que comme des listes de mesures spectrales sans tenir compte des relations spatiales entre pixels adjacents, écartant ainsi des informations importantes. En effet, les résultats de classification pourraient être améliorés en utilisant l'information contextuelle fournie par le spatial en plus de l'information spectrale [Dalla Mura et al., 2011, Gorretta et al., 2012, Tarabalka et al., 2010a]. Selon l'échelle d'acquisition, différentes sources de variabilité spectrale sont présentes au sein des objets et pourraient être gérées en utilisant l'information spatiale [Bioucas-Dias et al., 2013]. À cette fin, depuis la méthode originale "extraction et classification des objets homogènes" (ECHO) développée par Kettig and Landgrebe [1976], un grand nombre de recherches ont été menées pour trouver des classifieurs spectro-spatiaux efficaces [Fauvel et al., 2013].

Ces méthodes se répartissent en trois catégories [Valero, 2011] :
(1) Si les objets à classer ont de fortes caractéristiques discriminatoires spatiales, ces caractéristiques sont extraites et utilisées comme variables pour un classifieur. Par exemple, la segmentation d'image [Tilton, 2010], la Morphologie Mathématique [Aptoula and Lefèvre, 2007, Soille, 2003, Tilton, 2010], les filtres de régularisation à préservation de contours [Lennon et al., 2002, Wang et al., 2010].
(2) Si les objets à classer ont de fortes caractéristiques discriminatoires spectrales et spatiales, les deux sont extraites puis utilisées simultanément dans un classificateur par des techniques de noyaux [Camps-Valls et al., 2006, Fauvel, 2007], des champs de Markov [Rellier, 2002, Tarabalka et al., 2010b], ou à l'aide d'une analyse croisée [Gorretta, 2009].
(3) Si les objets à classer ont de fortes caractéristiques spectrales discriminatoires, l'information spectrale est d'abord traitée et l'information spatiale des pixels voisins est ensuite utilisée pour améliorer les résultats de la classification par segmentation ou régularisation des cartes de classifications [Kang et al., 2014, Li, 2011, Tarabalka, 2007].

Concernant les approches spectro-spatiales, les méthodes qui utilisent des filtres à préservation de contours (EPF) semblent être particulièrement bien adaptées à la classification des images HS. En effet, être en mesure de réduire la variabilité au sein des classes en utilisant un EPF semble très intéressant pour compléter la réduction de variabilité spectrale déjà obtenue par la méthode de réduction de dimension spectrale supervisée. Cependant, parmi les approches proposées dans la littérature, cette régularisation spatiale est effectuée uniquement soit sur les images HS brutes soit sur des images de scores obtenues de manière non supervisée. Dans les deux cas, la variabilité naturelle au sein de chaque classe peut conduire à des images très texturées. Ainsi, en utilisant le filtrage EPF, des bords sont aussi préservés à l'intérieur des classes qui doivent être homogénéisées. Par conséquent, nous proposons dans cette thèse d'utiliser l'EPF de façon légèrement différente : il est appliqué sur une image de scores, obtenue à partir d'une méthode de réduction de dimensions supervisée. En effet, en utilisant une méthode supervisée, la variabilité au sein des classes est réduite et la distance entre classes est augmentée, ce qui aide la régularisation spatiale à trouver des bords seulement aux frontières des classes.

B.2.3 Obtention d'images en réflectance

Dans le scénario idéal, chaque objet à classer peut être représenté par sa signature spectrale. Cependant, de nombreuses sources de variabilité incontrôlables tels que l'angle de la source de lumière incidente, l'angle d'acquisition, les conditions atmosphériques et un certain nombre d'autres variables affectent sensiblement la mesure spectrale [Barrett, 2013]. La correction en reflectance est donc indispensable pour chaque analyse d'image HS acquise en extérieur. Les méthodes disponibles sont généralement classés en trois catégories : modèles de transfert radiatifs, méthodes basées sur l'image et méthodes basées sur la scène [Shaw and Burke, 2003]. Des revues complètes concernant ces méthodes sont disponibles dans [Gao et al., 2009, Griffin and Burke, 2003].

Le modèle général [Gao and Goetz, 1990, Hamm et al., 2012] à partir duquel toutes les méthodes de corrections sont basées est donné par :

$$L_{obs}(\lambda) = \Big(E_{\downarrow}(\lambda)T_{\downarrow}(\lambda)\cos\theta + L_{\downarrow}(\lambda)\Big)T_{\uparrow}(\lambda)\pi^{-1}\rho(\lambda) + L_{\uparrow}(\lambda), \tag{B.1}$$

où, $\rho(\lambda)$ est la réflectance de surface, $L_{obs}(\lambda)$ la radiance observée par le capteur, $L_{\uparrow}(\lambda)$ la radiance montante (trajet cible → capteur) causé par la diffusion de l'atmosphère, $L_{\downarrow}(\lambda)$ l'irradiance descendante (illumination diffuse), $E_{\downarrow}(\lambda)$ la radiance exo-atmosphérique, θ l'angle du soleil par rapport à la surface, $T_{\downarrow}(\lambda)$ la transmission atmosphérique soleil → cible et $T_{\uparrow}(\lambda)$ la transmission cible → capteur.

On peut noter qu'une relation linéaire existe entre la radiance observée et la réflectance de surface :

$$L_{obs}(\lambda) = a(\lambda)\rho(\lambda) + b(\lambda). \tag{B.2}$$

L'objectif des méthodes de correction atmosphérique est donc de donner une estimation précise de $a(\lambda)$ et $b(\lambda)$. **Les modèle de transferts radiatifs** simulent le spectre de rayonnement solaire, calculent les effets de scènes, la position du soleil et mesurent ou estiment le taux de particules absorbantes et diffusantes de l'atmosphère [Kruse, 2000]. **Les méthodes de corrections basées sur la scène** utilisent des sources d'informations supplémentaires afin d'estimer empiriquement les termes additif et multiplicatif [Moran et al., 2001, Smith and Milton, 1999, Vain et al., 2009]. **Les méthodes de corrections basées sur l'image** utilisent uniquement les informations qui peuvent être récupérées à partir de l'image pour effectuer la correction atmosphérique.

B.3 Approches proposées

B.3.1 Introduction

Lorsque les données HS sont utilisées à des fins de classification, les différences entre réponses spectrales sont utilisées pour attribuer un label à chaque pixel de l'image HS. Si la classification est supervisée, des échantillons d'apprentissage avec labels connus sont nécessaires afin d'étalonner le modèle de classification. Cependant, des questions spécifiques sont soulevées quand un modèle de classification fiable doit être créé avec ces données complexes. Dans cette thèse, nous présentons trois approches pour faire face à certaines de ces questions principales, à savoir, la réduction de la dimension spectrale, la combinaison des informations spectrale et spatiale et la correction en réflectance.

B.3.2 Réduction de dimension

La classification supervisée consiste, en utilisant une matrice de données $\mathbf{X}$ et une matrice de classe $\mathbf{Y}$ d'échantillons d'apprentissages, à trouver un modèle capable de prédire la classe de toute nouvelle observation $\mathbf{x}$ en utilisant ses P descripteurs. Avec les données spectrales, la classification se fait généralement en deux étapes:
(1) projection de l'observation dans un sous-espace de dimension plus faible;
(2) affectation de l'observation à une classe.

L'efficacité de la deuxième étape est fortement influencée par la première. Par conséquent, nous recherchons un sous-espace dans lequel les centres des classes sont bien séparés et la répartition des classes autour de leurs centres est faible. D'un point de vue mathématique, cela correspond à trouver des facteurs $\mathbf{D}$ $\big(P \times Q\big)$ tels que la projection de $\mathbf{X}$ sur $\mathbf{D}$ minimise le lambda de Wilk's:

$$\Lambda_{Wilks} = \frac{\text{trace}(\mathbf{W})}{\text{trace}(\mathbf{W}+\mathbf{B})} \tag{B.3}$$

qui correspond au ratio de la variabilité intra-classe sur variabilité totale (somme de la variabilité inter- et intra-classes). Dans les cas “bien conditionnés”, une solution est donnée par l'analyse factorielle de Fisher (LDA) :

$$\mathbf{D} = \arg\max_{\mathbf{D}} \Big(\text{trace}\big(\mathbf{D}^{\intercal}\mathbf{W}^{-1}\mathbf{B}\mathbf{D}\big)\Big) = \mathbf{E}_Q\big(\mathbf{W}^{-1}\mathbf{B}\big) \tag{B.4}$$

où pour toute une matrice carré diagonalisable $\mathbf{A}$, la notation $\mathbf{E}_Q\big(\mathbf{A}\big)$ correspond aux Q vecteurs propres associés à ses Q plus grandes valeurs propres. Cependant, avec des données mal conditionnées, l'inversion de $\mathbf{W}$ devient problématique. Par conséquent, La LDA est incapable de traiter des données spectrales directement, et plusieurs solutions ont été proposées dans la littérature pour résoudre ce problème.

Néanmoins, la construction d'un modèle de classification correspond à trouver un sous-espace de l'espace des variables qui “copie” la structure de classe observée dans l'espace des individus. La LDA le fait en contractant le sous-espace porté par la variance intra-classe et en se focalisant sur celui porté par la variance inter-classes.

La méthode proposée dans cette thèse offre une autre façon de réaliser cette copie. L'idée est d'utiliser les variance inter- et intra-classes pour décomposer l'espace des variables en différents sous-espaces, de sorte que l'un d'eux porte une grande partie de la variance inter-classes et une petite partie de l'intra-classe. Cependant, la séparation des sources de variance n'est pas évidente en raison de la colinéarité potentielle entre les sous-espaces $\mathcal{F}_{\mathbf{B}}$ et $\mathcal{F}_{\mathbf{W}}$. Ainsi, selon la configuration des classes, la suppression de la variance intra-classe n'améliore pas nécessairement la séparabilité. Dans cette thèse, nous proposons une méthode, appelée DROP-D, qui permet une suppression contrôlée de la variabilité intra-classe, c'est à dire, en préservant ses axes colinéaires à $\mathcal{F}_{\mathbf{B}}$.

DROP-D: Réduction de dimension par projection orthogonale pour la discrimination est décomposée en trois étapes.
La première étape consiste à supprimer de $\mathbf{X}$ les b directions principales de la variabilité inter-classes, tel que :

$$\mathbf{X}_b^{\perp} = P_{\mathbf{B},b}^{\perp}\big(\mathbf{X}\big). \tag{B.5}$$

Dans la seconde étape, la variabilité intra-classe est calculée sur $\big(\mathbf{X}_b^{\perp}, \mathbf{Y}\big)$. Ensuite, les w directions principales liée à cette variabilité intra-classe ($\mathbf{W}^*$) sont éliminées suivant l'équation :

$$\mathbf{X}_{clean} = P_{\mathbf{W}^*\left(\mathbf{X}_B^{\perp}, \mathbf{Y}\right), w}^{\perp}\big(\mathbf{X}\big). \tag{B.6}$$

La troisième étape consite à extraire les Q directions principales de $\mathbf{X}_{clean}$ qui sont données par :

$$\mathbf{D} = \mathbf{E}_Q\Big(\mathbf{T}\big(\mathbf{X}_{clean}\big)\Big). \tag{B.7}$$

En résumé, DROP-D défini trois sous-espaces de $\mathbb{R}^P$, $\mathcal{F}_B$, $\mathcal{F}_{W^*}$ et $\mathcal{F}_D$ tel que :
- $\mathcal{F}_B$ est lié aux b directions principales de la variabilité inter-classes
- $\mathcal{F}_{W^*}$ contient les w directions principales de la variabilité intra-classe (orthogonal à $\mathcal{F}_B$)
- $\mathcal{F}_D$ contient les Q directions qui incluent les b directions principales de la variabilité inter-classes et les $Q - b$ direction principales qui sont orthogonales à la variabilité intra-classe.

Ce faisant, DROP-D élimine les directions principales de la variabilité intra-classe, tout en préservant les directions les plus importantes de la variabilité inter-classes. Une simple projection orthogonalement à $\mathbf{W}$ risquerait de supprimer aussi des axes importants de $\mathbf{B}$, car $\mathcal{F}_B$ et $\mathcal{F}_W$ peuvent avoir des parties colinéaires. En ce sens, l'étape 1 de DROP-D garantie de préserver au moins les b axes les plus importants de $\mathcal{F}_B$. En outre, les axes de $\mathcal{F}_B$ qui n'étaient pas inclues à l'étape 1, mais qui sont orthogonaux à $\mathcal{F}_W$, sont également préservés.

B.3.3 Approche spectro-spatiale

Dans l'état de l'art, nous avons vu plusieurs approches spectro-spatiales visant à améliorer les performances de classification. Parmi elles, les EPF ont prouvé leur efficacité au travers de différentes études. Dans cette thèse, nous proposons également une approche spectro-spatiale qui utilise la régularisation spatiale EPF afin d'améliorer les résultats de la classification purement spectrale. L'hypothèse retenue lors de l'utilisation de la régularisation spatiale EPF pour améliorer les résultats de classification est que les bords sont supposés être présents seulement aux frontières des classes et non à l'intérieur des classes. Cependant, dans les images réelles, des bords sont également trouvés ailleurs qu'aux frontières des classes en raison du bruit de fond provoqué par la texture, les non homogénéités de couleur, d'éclairage, etc. Par conséquent, l'utilisation d'EPF directement sur une image HS conserve les bords dus au bruit de fond et ne parvient pas à en réduire la variabilité. L'EPF appliqué à une image de scores obtenus par une méthode de réduction de dimension non-supervisée échoue de manière similaire car les caractéristiques extraites comprennent également le bruit de fond.

Pour compenser ce problème, nous proposons donc une approche dans laquelle la régularisation spatiale est appliquée à une image de scores obtenue par une méthode de réduction de dimension supervisée (comme DROP-D). L'idée de base est que, puisque l'image des scores décrit déjà les classes à discriminer en minimisant la variabilité due au bruit de fond, les bordures correspondent principalement aux frontières de classes et le processus de régularisation spatial est plus efficace.

Construction de l'image des scores

Une image hyperspectrale $\mathcal{H}$ de dimension $I \times J \times P$, c'est à dire, I lignes, J colonnes et P longueur d'ondes, peut être dépliée dans une matrice $\mathbf{H}$ de taille $M \times P$ où $M = I \cdot J$. La notation $\mathcal{H}_i$ correspond au i^{eme} canal de l'image HS. L'image de scores $\mathcal{S}$ de taille $I \times J \times Q$ est obtenue de manière similaire en repliant la matrice de scores $\mathbf{S}$ de taille $M \times Q$ donnée par :

$$\mathbf{S} = \mathbf{HD}. \tag{B.8}$$

Chaque canal $\mathcal{S}_i$ de l'image de scores correspond donc au i^{eme} score.

Notons que comme les scores de DROP-D sont obtenues à partir d'une PCA, les facteurs ainsi que les scores sont orthogonaux. Les différents canaux de l'image de scores sont donc supposés non corrélés.

Régularisation anisotropique

Nous mettons en œuvre notre approche en utilisant la méthode de diffusion anisotropique de Perona and Malik [1990] pour augmenter l'homogénéité au sein des régions tout en gardant intact les frontières entre régions adjacentes. Cette méthode a été développée pour débruiter des images en niveau de gris en lissant l'image sans en enlever les bordures principales.

La méthode de Perona and Malik [1990] est un processus itératif dans lequel, à chaque itération, la quantité de lissage est pondérée par l'intensité locale du gradient.

$$\frac{\partial \mathcal{I}(x,y,t)}{\partial t} = div \left[g\big(\parallel \nabla \mathcal{I}(x,y,t) \parallel \big) \nabla \mathcal{I}(x,y,t) \right] \tag{B.9}$$

où la fonction g doit être décroissante par rapport à la norme du gradient $\alpha = \parallel \nabla \mathcal{I} \parallel$. Dans [Perona and Malik, 1990], les auteurs ont utilisés une fonction Gaussienne déterminée seulement par un paramètre, correspondant à une largeur de noyau de lissage η. Cette fonction est donnée par :

$$g(\alpha) = \exp\big(- \big(\alpha/\eta\big)^2 \big). \tag{B.10}$$

Ce processus de diffusion est donc anisotropique et permet donc de conserver les bordures principales.

Régularisation des images de scores

En fonction des scores obtenus, deux schémas de régularisation peuvent être envisagés. **Avec des scores non-orthogonaux**, une régularisation multidimensionnelle est préférable afin d'éviter les valeurs aberrantes. **Avec des scores orthogonaux**, on peut trouver des régions homogènes sur un score alors qu'il y a une transition de classe sur un autre. Dans ce cas, chaque canal de l'image des scores peut être traité de façon indépendante, ce qui conduit à une méthode très simple et parallèle.

Le processus de diffusion est donc dans notre cas appliqué individuellement sur chaque canal $\mathcal{S}_i$ de l'image des scores $i \in [\![1, \cdots, Q]\!]$. Le processus est initialisé avec $\mathcal{S}_{i,0} = \mathcal{S}_i$. Ensuite, à l'itération $k+1$, la diffusion est appliquée numériquement au canal $\mathcal{S}_{i,k}$ en suivant l'équation :

$$\mathcal{S}_{i,k+1} = \mathcal{S}_{i,k} + \epsilon \cdot div \left[g\big(\parallel \nabla \mathcal{S}_{i,k} \parallel \big) \cdot \nabla \mathcal{S}_{i,k} \right] \tag{B.11}$$

où ϵ règle le taux de change à chaque itération du processus de diffusion.

B.3.4 Correction en réflectance

Dans cette section, nous proposons un approche automatique de correction de l'éclairement qui, dans le cas de la classification, évite l'utilisation de mesures de références. L'hypothèse principale est que les matériaux à discriminer sont lambertiens.

Hypothèse lambertienne
La quantité mesurée par une caméra HS est, après correction radiométrique, une radiance spectrale $L(\lambda)$, c'est à dire, une irradiance mesurée dans une direction spécifique (en $W.sr^{-1}.m^{-2}.nm^{-1}$). Avec des matériaux lambertiens, pour un pixel i, j, la radiance mesurée est :

$$L_{i,j}(\lambda) = r_{i,j}(\lambda)E(\lambda) \tag{B.12}$$

où $r_{i,j}(\lambda)$ est la réflectance en radiance et $E(\lambda)$ l'irradiance descendante (en $W.m^{-2}.nm^{-1}$) supposée identique pour chaque pixel.

Hypothèse du modèle de discrimination
Considérons une matrice $\mathbf{X}$ de taille $(N \times P)$ qui correspond aux N spectres de P longueurs d'ondes extraits de l'image HS. Définissons une matrice $\mathbf{Y}$ de dimension $(N \times C)$ qui code le degré d'appartenance de chaque spectre de $\mathbf{X}$. Considérons également une méthode de classification qui calcule des scores (comme DROP-D) et qui les utilise ensuite pour la discrimination. Rappelons que les projections d'un espace de grandes dimensions vers un espace de dimensions plus faibles favorisent la Gaussianité des distributions, dans la suite, nous utilisons donc un classifieur de Bayes sur les scores réduits. On a donc un modèle de discrimination comme suit :
(1) Réduction de dimension linéaire $\mathbf{D}$ qui décompose la matrice de spectres en une matrice de scores $\mathbf{S}$:

$$\mathbf{S} = \mathbf{XD} \tag{B.13}$$

de dimensions $N \times Q$, avec $Q \ll P$.

(2) Pour chaque classe $c \in [1, \cdots, C]$ une estimation du vecteur moyen $\hat{\boldsymbol{\mu}}_c$ et de la matrice de covariance $\widehat{\boldsymbol{\Sigma}}_c$ est calculée en utilisant les observations de l'ensemble d'apprentissage.

(3) La décision de classification pour une nouvelle observation est faite en utilisant le classifieur de Bayes définie par :

$$\hat{c} = \arg\max_{c} \frac{1}{(2\pi)^{\frac{Q}{2}} |\widehat{\boldsymbol{\Sigma}}_c|^{\frac{1}{2}}} \exp^{\left(\frac{1}{2}\left(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_c\right)^{\mathsf{T}} \widehat{\boldsymbol{\Sigma}}_c^{-1} \left(\boldsymbol{s}-\hat{\boldsymbol{\mu}}_c\right)\right)} \tag{B.14}$$

Définition du problème

Nous avons vu que réflectance et radiance sont liés par un terme multiplicatif, constant pour chaque pixel de l'image, $E(\lambda)$. La clé de la méthode est de transformer l'équation B.12 en utilisant le logarithme :

$$log(L_{i,j}(\lambda)) = log\big(r_{i,j}(\lambda)E(\lambda)\big) = log\big(r_{i,j}(\lambda)\big) + log\big(E(\lambda)\big). \tag{B.15}$$

On peut montrer que la seule différence qui se produit dans le modèle en utilisant la log-réflectance par rapport à un modèle utilisant la log-radiance est l'estimation du vecteur moyen par classe, c'est à dire que la matrice de réduction de dimension ainsi que les matrices de covariances sont inchangées. D'un point de vue classification, la correction en réflectance peut être évitée si l'on connait la translation $\widehat{\boldsymbol{\mu}^{(E)}}$, en changeant la règle de décision par $\hat{\boldsymbol{\mu}}_c + \widehat{\boldsymbol{\mu}^{(E)}}$ à la place de $\hat{\boldsymbol{\mu}}_c$.

Estimation de la translation

Nous proposons également dans cette thèse une méthode pour estimer la translation automatiquement. Cette méthode, basée sur une technique utilisée pour le recalage d'image est robuste aux classes manquantes ainsi qu'au nombre variable d'observations par classe. Avant recalage, une image de dimension Q doit être crée à partir des scores de dimension Q. La création de l'image est basée sur l'estimation de distribution de classes dans l'espace de dimension Q pour chaque image HS. Dans notre cas où la distribution des classes est supposée gaussienne, un outil puissant pour l'estimation est l'algorithme Espérance-Maximisation (EM) [Moon, 1996]. Finalement, la corrélation croisée est utilisée comme mesure d'appariement pour estimer la translation des scores de chacune des images.

B.4 Résultats

L'objectif de cette section est de montrer la pertinence des approches développées à partir d'images hyperspectrales acquises dans des conditions extérieures réelles. Dans l'ensemble de données [1], des images HS de proxi-détection ont été acquises avec une caméra Hyspex V-NIR 1600 (Norks Elektro Optikk, Norvège). Les images ont été acquises sur le terrain à l'aide d'un rail de translation monté sur un tracteur, à 1 mètre au-dessus du sol (résolution spatiale de $0,2$ mm/pixel). En utilisant une surface de référence calibrée dans chaque image, les images de radiance (LXX) ont été transformées en images de réflectance (RXX). Deux acquisitions ont été réalisées à une heure d'intervalle à des endroits différents dans le même champ. **Ensemble de données A1** a été utilisé pour étalonner le modèle, dans lesquels trois classes doivent être discriminées : *blé*, *adventices* et *sol*. Pour l'étalonnage du modèle, 100 spectres par classe (300 au total) ont été extraits au hasard à partir de la carte de vérité terrain disponible (cela qui correspond à environ $0,6\%$ des données disponibles). Par souci de cohérence avec les résultats de la dernière section, dans laquelle une transformation logarithmique doit être effectuée, nous avons utilisé les spectres transformés en logarithme dans toute la thèse. **Les données A2** correspondent au même champ que les données A1 mais acquises une heure plus tard. Cet ensemble de données est utilisé dans la dernière section pour illustrer le problème de la transformation en réflectance.

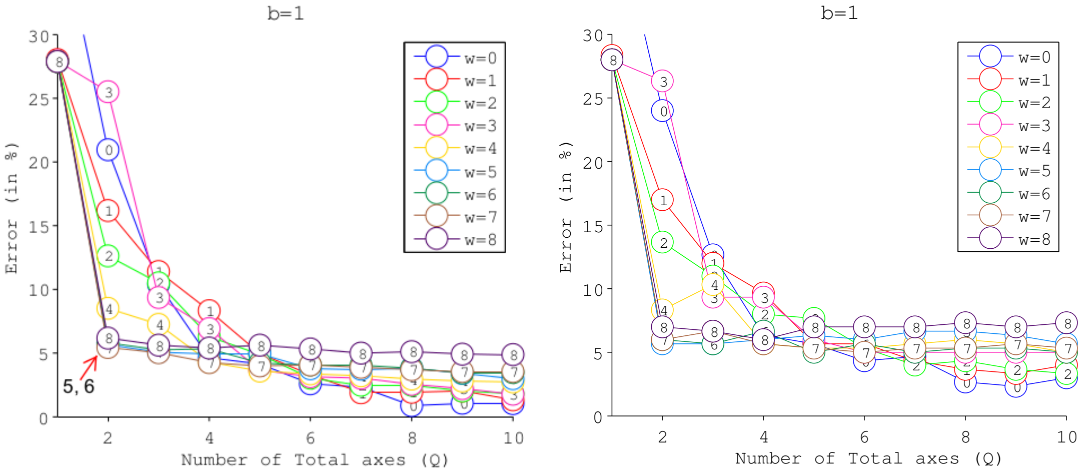
L'erreur de classification correspond au ratio de pixels mal classifiés (en pourcentage). Les résultats de validation croisée sont utilisés pour le réglage des paramètres des différents modèles. Dans ce cas, nous avons utilisé une procédure "10-fold" sur l'ensemble d'apprentissage [Esbensen and Geladi, 2010]. Pour validation, les résultats sont présentés sur le jeu de données globales moins les données d'apprentissage.

B.4.1 Réduction de dimension

La figure B.1, montre l'erreur de classification obtenue avec l'ensemble d'apprentissage et celle obtenue avec une 10-fold validation croisée. Ces graphiques présentent

[1] Dans ce résumé, un seul jeu de données est présenté (jeu de données A). La notation utilisée pour décrire les jeux de données sont : 1^{er} caractère: lettre R pour réflectance et L pour un luminance; 2^{eme} caractère : lettre A ou B pour l'ensemble de données; 3^{eme} caractère : numéro du sous-ensemble

l'erreur de classification en fonction du nombre d'axes discriminants retenus pour différents nombres d'axes intra-classes retirés (nombre inscrit dans le cercle). Avec $w = 0$ (correspond à une PCA) les erreurs de d'étalonnage et de validation croisée décroissent lentement sans minimum évident. Pour $w = 1$ à $w = 4$, l'erreur de classification varie fortement. Ensuite, à partir de $w = 5$ (optimal) jusqu'à $w = 7$, la même erreur de classification est obtenue. En plus, un minimum évident apparaît pour Q ($Q = 2$). Ensuite, si plus d'axes intra-classes sont retirés, l'erreur commence à augmenter (même sur l'ensemble d'étalonnage), permettant donc de régler les paramètres sans validation croisée. En terme de résultats de classification, DROP-D offre les mêmes possibilités que les approches présentées précédemment, tout en évitant le risque de sur-apprentissage.



(A) Calibration error

(B) Cross-validation error

Figure B.1: Jeu de données RA1: Erreur d'étalonnage (A) et de validation croisée (B) pour différent paramètres pour w et Q et avec $b = 1$.

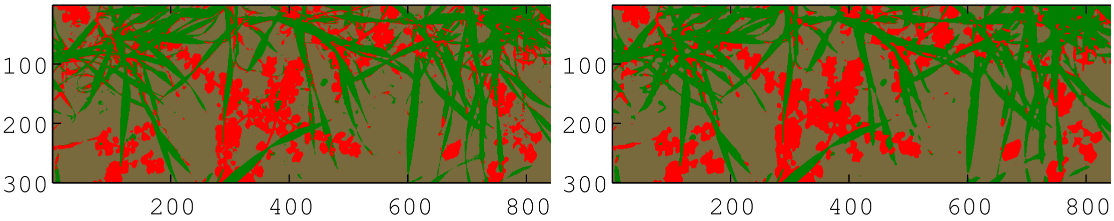
B.4.2 Régularisation spatiale

Afin de valider quantitativement notre approche, le Tableau B.1 donne des résultats de classification pour différent schémas de régularisation en utilisant les K plus proche voisins (KNN) comme méthode de classification avec K=3. On observe déjà que l'utilisation de réduction de dimension supervisée donne de meilleurs résultats que les méthodes non-supervisées. Ensuite, pour chaque méthode, la régularisation spatiale permet d'améliorer les résultats de classification. Pour l'ACP, une amélioration légère est obtenue lorsque la régularisation est effectuée après réduction de dimension. En revanche, pour la PLS ou DROP-D, on voit

Table B.1: RA1: erreur de classification (%) pour différentes procédures de régularisation

Without regularization		Regularization first		Regularization second	
KNN	7.97	AR-KNN	7.62	n/a	n/a
PCA-KNN	7.35	AR-PCA-KNN	7.27	PCA-AR-KNN	7.26
PLS-KNN	6.84	AR-PLS-KNN	6.62	PLS-AR-KNN	**4.91**
DROP-D-KNN	7.14	AR-DROP-D-KNN	7.10	DROP-D-AR-KNN	**5.37**

clairement que la régularisation doit être effectuée après réduction de dimension comme proposé dans cette thèse. La Figure B.2 représente les cartes de classification obtenues avant et après régularisation spatiale. On observe que le bruit de classification est largement diminué, montrant des classes plus homogènes. Aux bordures des classes, notamment avec les feuilles de blé (vert), l'erreur de classification est aussi nettement diminuée.



(a) Données RA1 avant AR : Erreur 4.9% (b) Données RA1 après AR : Erreur 3.2%

Figure B.2: Résultats de classification avant et après régularisation.

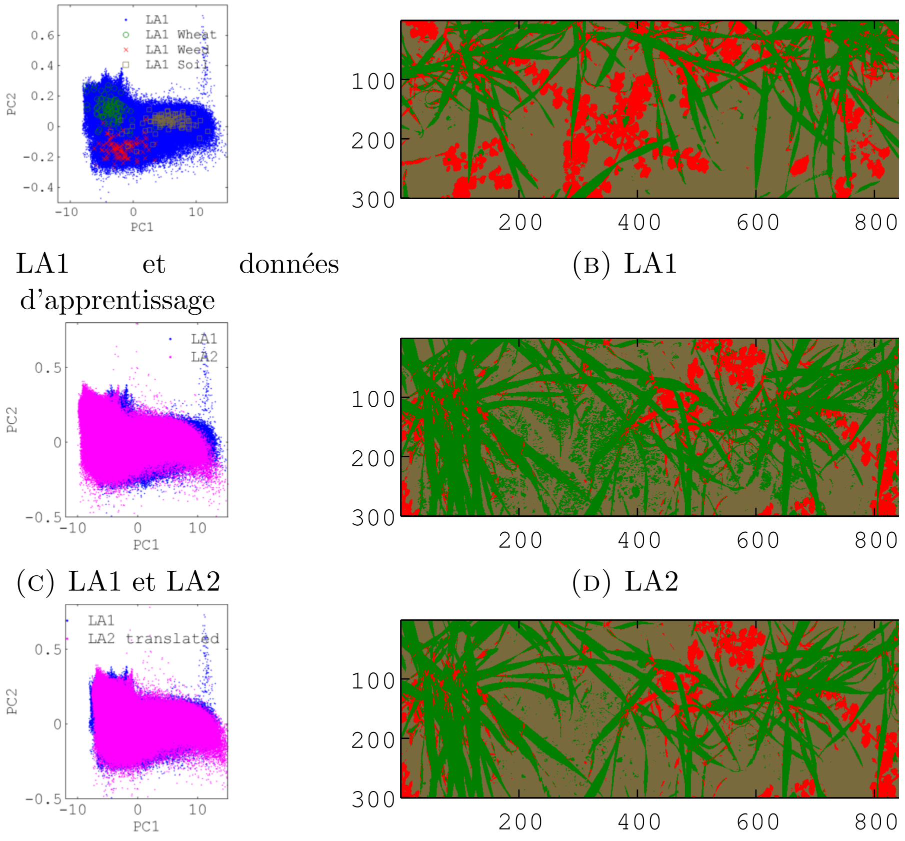
En comparant par rapport à d'autres approches spectro-spatiales récentes, notamment SVM-EPF [Kang et al., 2014], LORSAL Multilevel Logistic (LORSAL-MLL) [Li et al., 2012] notre approche offre des résultats au moins aussi bons sur les images classiquement utilisées pour comparer ce type de méthodes.

B.4.3 Correction en réflectance

Dans cette section nous souhaitons montrer qu'en utilisant la log-radiance et en estimant correctement la translation, la correction en réflectance pouvait être évitée.

Par exemple, la Figure B.3 montre les différences de scores et les résultats de classification obtenus sur notre jeu de données avec des acquisitions à une heure d'intervalle. Comme cette translation est principalement selon l'axe (PC1) (figure c), la carte de classification obtenue est plus verte (figure d), même pour classifier du sol qui est normalement aisément discriminable de la végétation. Après estimation de la translation et recalage, les scores deviennent parfaitement alignés (figure e) et la carte obtenue est de nouveau correctement classifiée (figure f).

Des résultats similaires sont obtenus en l'absence de la classe adventice (voir les résultats complets de la thèse).



(A) LA1 et données d'apprentissage

(B) LA1

(C) LA1 et LA2

(D) LA2

(E) scores LA1 et scores LA2 après translation

(F) LA2 après recalage

Figure B.3: Scores et cartes de classification obtenues avec un modèle étalonné sur LA1

B.5 Conclusion

Dans cette thèse, nous avons abordé trois principales questions portant sur la classification d'image hyperspectrale. Pour chaque approche proposée, différentes directions de recherche ainsi que des améliorations sont possibles.

Dans la mise en œuvre de DROP-D, les vecteurs propres associés aux plus grandes valeurs propres sont conservés. Toutefois, il serait intéressant de choisir d'autres

combinaisons d'axes inter- et intra-classes à conserver, non nécessairement à partir des axes principaux. Ce choix est évident quand il y a très peu de classes. Il faudrait cependant mettre en œuvre des techniques d'optimisation dès que le nombre de classes augmente.
Une autre approche serait de relaxer la contrainte d'orthogonalité de la projection. On espère qu'une projection non-orthogonale serait capable de réduire davantage la variabilité intra-classe sans affecter la distance entre classes.
Une autre possibilité serait d'utiliser des données nettoyées par DROP-D avec d'autres méthodes de classification plus performantes comme les machines à vecteurs de support (SVM) afin de réduire le nombre de vecteurs de support à utiliser.
Enfin, d'un point de vue plus théorique, il serait intéressant de comparer le nettoyage réalisé par DROP-D avec d'autres méthodes chimiométriques développées pour la régression, comme l'Orthogonal Signal Correction par exemple.

Nous avons mis en œuvre l'approche spectro-spatiale en utilisant la version originale de diffusion anisotropique. Cependant, comme la communauté de traitement d'image a beaucoup travaillé dans le domaine de la régularisation spatiale, l'utilisation d'approches plus sophistiquées serait sûrement bénéfique.
De plus, en observant les cartes des résidus (différences avant/après régularisation), certains motifs obtenus semblent d'un intérêt potentiel, en termes d'analyse de texture, pour continuer à augmenter les performances de classification. Ces caractéristiques texturales pouvant être ajoutées comme entrée d'un classifieur de la même manière que les données spectrales.

Finalement, pour l'estimation de la translation, plusieurs améliorations peuvent être apportées. Dans un premier temps, une alternative à la corrélation croisée (très coûteuse) afin d'estimer la translation serait bénéfique (en utilisant la transformée de Fourier par exemple). Puis, l'utilisation de l'information spatiale (ex: les formes obtenues dans la carte de classification comme feedback), en plus de l'information de corrélation, devrait aider à estimer la translation lorsque les classes ne sont pas évidentes à regrouper dans l'espace des scores.
Enfin, une solution pour gérer les objets non-lambertiens serait très bénéfique, car même les corrections en réflectance classiques ne peuvent correctement traiter ce cas particulier. Nous pensons que cette approche utilisant une transformation logarithmique comme un pré-traitement pourrait également être applicable pour traiter les cas non-lambertiens. Dans ce cas, chaque classe se translaterait indépendamment l'une de l'autre, mais avec la même matrice de covariance. Ainsi,

en utilisant des techniques de recalage non-rigides, en faisant correspondre les matrices de covariances, la correction pourrait être envisageable dans certains cas.

Bibliography

Adler-Golden, S., et al., 1994: FLAASH, A MODTRAN4 Atmo- spheric Correction Package for Hyperspectral Data Retrievals and Simulations. *Proc. 7th Ann. JPL Airborne Earth Science Workshop, JPL Publication 97-21, Pasadena, Calif.*, 9–14.

Aptoula, E. and S. Lefèvre, 2007: A comparative study on multivariate mathematical morphology. *Pattern Recognition*, **40**, 2914–2929.

Balabin, R. M. and S. V. Smirnov, 2011: Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Analytica chimica acta*, **692**, 63–72.

Barker, M. and W. Rayens, 2003: Partial least squares for discrimination. *Journal of Chemometrics*, **17 (3)**, 166–173.

Barrett, E. C., 2013: *Introduction to environmental remote sensing.* Routledge.

Baudat, G. and F. Anouar, 2000: Generalized discriminant analysis using a kernel approach. *Neural computation*, **12**, 2385–2404.

Bellman, R. and R. E. Kalaba, 1965: *Dynamic programming and modern control theory.* Academic Press New York.

Bertrand, D., P. Courcoux, J.-C. Autran, R. Meritan, and P. Robert, 1990: Stepwise canonical discriminant analysis of continuous digitalized signals: Application to chromatograms of wheat proteins. *Journal of Chemometrics*, **4**, 413–427.

Bioucas-Dias, J. M., A. Plaza, G. Camps-Valls, P. Scheunders, N. M. Nasrabadi, and J. Chanussot, 2013: Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine.*

Bishop, C. M., 2007: *Pattern Recognition and Machine Learning.*

Bishop, C. M., J. Lasserre, et al., 2007: Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, **8**, 3–23.

Boulet, J.-C. and J.-M. Roger, 2012: Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, **117**, 61–69.

Brown, P., T. Fearn, and M. Haque, 1999: Discrimination with many variables. *Journal of the American Statistical Association*, **94 (448)**, 1320–1329.

Brown, R. and S. Noble, 2005: Site-specific weed management: sensing requirements-what do we need to see? *Weed Science*, **53 (2)**, 252–258.

Bylesjö, M., M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg, 2006: OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, **20**, 341–351.

Camps-Valls, G., L. Gomez-Chova, J. Muñoz Marí, J. Vila-Francés, and J. Calpe-Maravilla, 2006: Composite Kernels for Hyperspectral Image Classification. *Geoscience and Remote Sensing Letters, IEEE*, **3 (1)**, 93–97.

Chang, C.-I., 2007: *Hyperspectral Data Exploitation: Theory and Applications.* 440 pp.

Chen, L.-F., H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, 2000: A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, **33 (10)**, 1713–1726.

Cheng, Y., Y. Zhuang, and J. Yang, 1992: Optimal Fisher discriminant analysis using the rank decomposition. *Pattern recognition*, **25 (1)**, 101–111.

Chevallier, S., D. Bertrand, A. Kohler, and P. Courcoux, 2006: Application of PLS-DA in multivariate image analysis. *Journal of Chemometrics*, **20**, 221–229.

Clerbaux, C., et al., 2009: Monitoring of atmospheric composition using the thermal infrared iasi/metop sounder. *Atmospheric Chemistry and Physics*, **9 (16)**, 6041–6054.

Cooley, T., et al., 2002: FLAASH, a MODTRAN4-based atmospheric correction algorithm, its application and validation. *IEEE International Geoscience and Remote Sensing Symposium*, **3**.

Cover, T. and P. Hart, 1967: Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13 (1)**, 21–27.

Dahm, D. and K. Dahm, 2001: The physics of near-infrared scattering. *Near-Infrared Technology in the Agricultural and Food Industries*, Vol. 2, chap. 1.

Dalla Mura, M., A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, 2011: Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis. *Geoscience and Remote Sensing Letters, IEEE*, **8 (3)**, 542–546.

De Maesschalck, R., D. Jouan-Rimbaud, and D. L. Massart, 2000: The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, **50 (1)**, 1–18.

Diaconis, P. and D. Freedman, 1984: Asymptotics of graphical projection pursuit. *The annals of statistics*, 793–815.

Donoho, D., 2000: High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.

Duarte-Carvajalino, J. M., M. Vélez-Reyes, and P. Castillo, 2006: Scale-space in hyperspectral image analysis. *Proc. SPIE 6233, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII*, Vol. 6233, 623 312–623 315.

Duin, R. P., M. Loog, and T. K. Ho, 2006: Recent submissions in linear dimensionality reduction and face recognition. *Pattern Recognition Letters*, **27 (7)**, 707–708.

Esbensen, K. H. and P. Geladi, 2010: Principles of Proper Validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, **24 (3-4)**, 168–187.

Evans, A. N. and X. U. Liu, 2006: A morphological gradient approach to color edge detection. *Image Processing, IEEE Transactions on*, **15 (6)**, 1454–1463.

Fan, J., Y. Fan, and Y. Wu, 2011: No Title. *High-dimensional Data Analysis. (Cai, T.T. and Shen, X., eds.)*, World Scientific, New Jersey, 3–37.

Fauvel, M., 2007: Spectral and spatial methods for the classification of urban remote sensing data. Ph.D. thesis.

Fauvel, M., Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, 2013: Advances in Spectral-Spatial Classification of Hyperspectral Images. *Proceedings of the IEEE*, **101 (3)**, 652–675.

Fearn, T., 2000: On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, **50 (1)**, 47–52.

Fearn, T., 2008: Principal component discriminant analysis. *Statistical applications in genetics and molecular biology*, **7 (2)**.

Fearn, T., 2011: Principal component analysis and classification. *NIR News*, **22 (3)**, 22.

Fisher, R. A., 1936: The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7 (2)**, 179–188.

Foley, D. H. and J. W. Sammon, 1975: An Optimal Set of Discriminant Vectors. *IEEE Transactions on Computers*, **C-24 (3)**, 281–289.

Fukunaga, K., 1990: *Introduction to statistical pattern recognition 2nd edition*. Academic p ed., 591 pp.

Fukunaga, K. and R. R. Hayes, 1989: Effects of sample size in classifier design. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **11 (8)**, 873–885.

Gao, B. and A. Goetz, 1990: Column atmospheric water vapor and vegetation liquid water retrievals from airborne imaging spectrometer data. *Journal of Geophysical Research*, **95 (D4)**, 3549–3564.

Gao, B.-C., K. B. Heidebrecht, and A. F. Goetz, 1993: Derivation of scaled surface reflectances from AVIRIS data. *Remote Sensing of Environment*, **44 (2-3)**, 165–178.

Gao, B.-C., M. J. Montes, C. O. Davis, and A. F. Goetz, 2009: Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote Sensing of Environment*, **113**, S17–S24.

Gat, N., 2000: Imaging spectroscopy using tunable filters: a review. *AeroSense 2000, International Society for Optics and Photonics*, H. H. Szu, M. Vetterli, W. J. Campbell, and J. R. Buss, Eds., Vol. 4056, 50–64.

Geladi, P., 2003: Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy*, **58 (5)**, 767–782.

Geladi, P., J. Burger, and T. Lestander, 2004: Hyperspectral imaging: calibration problems and solutions. *Chemometrics and Intelligent Laboratory Systems*, **72 (2)**, 209–217.

Ghamisi, P., J. Benediktsson, and J. Sveinsson, 2014: Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, **52 (9)**, 5771–5782.

Goetz, A. F., G. Vane, J. E. Solomon, and B. N. Rock, 1985: Imaging spectrometry for earth remote sensing. *Science*, **228 (4704)**, 1147–1153.

Gonzalez, R. C., R. E. Woods, and S. L. Eddins, 2009: *Digital image processing using MATLAB*, Vol. 2. Gatesmark Publishing Knoxville.

Gorretta, N., 2009: Proposition d ’ une approche de segmentation d ’ images hyperspectrales. Ph.D. thesis, Université de Montpellier 2 Sciences et Techniques du Languedoc, Montpellier.

Gorretta, N., G. Rabatel, C. Fiorio, C. Lelong, and J.-m. Roger, 2012: An iterative hyperspectral image segmentation method using a cross analysis of spectral and spatial information. *Chemometrics and Intelligent Laboratory Systems*, **117**, 213–223.

Gorretta, N., J. Roger, M. Aubert, V. Bellon-Maurel, F. Campan, and P. Roumet, 2006: Determining vitreousness of durum wheat kernels using near infrared hyperspectral imaging. *Journal of near infrared spectroscopy*, **14 (4)**, 231.

Gowen, A., C. O’Donnell, P. Cullen, G. Downey, and J. Frias, 2007: Hyperspectral imaging–an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, **18 (12)**, 590–598.

Grahn, H. and P. Geladi, 2007: *Techniques and applications of hyperspectral image analysis*. John Wiley & Sons.

Green, A., M. Berman, P. Switzer, and M. Craig, 1988: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, **26 (1)**, 65–74.

Griffin, M. and H. Burke, 2003: Compensation of hyperspectral data for atmospheric effects. *Lincoln Laboratory Journal*, **14 (1)**, 29–54.

Guo, Y.-F., L. Wu, H. Lu, Z. Feng, and X. Xue, 2006: Null Foley–Sammon transform. *Pattern Recognition*, **39 (11)**, 2248–2251.

Hadoux, X., S. Jay, G. Rabatel, and N. Gorretta, 2014: A spectral-spatial approach for hyperspectral image classification using spatial regularization on supervised score image. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 1–9.

Hagen, N., R. T. Kester, L. Gao, and T. S. Tkaczyk, 2012: Snapshot advantage: a review of the light collection improvement for parallel high-dimensional measurement systems. *Optical engineering (Redondo Beach, Calif.)*, **51 (11)**.

Hagen, N. and M. W. Kudenov, 2013: Review of snapshot spectral imaging technologies. *Optical Engineering*, **52 (9)**, 090 901.

Hall, P. and K.-C. Li, 1993: On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, 867–889.

Hamamoto, Y., T. Kanaoka, and S. Tomita, 1993: On a theoretical comparison between the orthonormal discriminant vector method and discriminant analysis. *Pattern Recognition*, **26 (12)**, 1863–1867.

Hamm, N., P. Atkinson, and E. Milton, 2012: A per-pixel, non-stationary mixed model for empirical line atmospheric correction in remote sensing. *Remote Sensing of Environment*, **124**, 666–678.

Harsanyi, J. C. and C.-I. Chang, 1994: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, **32 (4)**, 779–785.

Hastie, T., A. Buja, and R. Tibshirani, 1995: Penalized discriminant analysis. *The Annals of Statistics*, **23 (1)**, 73–102.

Hastie, T. and R. Tibshirani, 1996: Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 155–176.

Hoffbeck, J. and D. Landgrebe, 1994: Effect of radiance-to-reflectance transformation and atmosphere removal on maximum likelihood classification accuracy of high-dimensional remote sensing data. *IEEE International Geoscience and Remote Sensing Symposium.*

Hong, Z.-Q. and J.-Y. Yang, 1991: Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *Pattern Recognition*, **24 (4)**, 317–324.

Huang, R., Q. Liu, H. Lu, and S. Ma, 2002: Solving the small sample size problem of LDA. *IEEE 16th International Conference on Pattern Recognition, 2002.*, IEEE Comput. Soc, Vol. 3, 29–32.

Hughes, G. F., 1968: On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, **14 (1)**, 55–63.

Indahl, U. G., N. S. Sahni, B. Kirkhus, and T. Næ s, 1999: Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise. *Chemometrics and Intelligent Laboratory Systems*, **49 (1)**, 19–31.

Jimenez, L. O. and D. Landgrebe, 1996: High dimensional feature reduction via projection pursuit. *ECE Technical Reports*, 103.

Jimenez, L. O., A. Member, D. A. Landgrebe, and L. Fellow, 1998: Supervised Classification in High-Dimensional Space : Geometrical , Statistical , and Asymptotical Properties of Multivariate Data. *IEEE Transactions on Systems, Man and Cybernetics*, **28 (1)**, 39–54.

Jolliffe, I., 2005: *Principal Component Analysis*. John Wiley & Sons, Ltd.

Kang, X., S. Li, and J. Benediktsson, 2014: Spectral–spatial hyperspectral image classification with edge-preserving filtering. *IEEE Transactions on Geoscience and Remote Sensing*, **52 (5)**, 2666–2677.

Kemsley, E., 1996: Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, **33 (1)**, 47–61.

Kendall, M., 1961: *A Course in the Geometry of n-dimensions*. New york: Hafner Publishing Co., 35 pp.

Kettig, R. L. and D. Landgrebe, 1976: Classification of multispectral image data by extraction and classification of homogeneous objects. *Geoscience Electronics, IEEE Transactions on*, **14 (1)**, 19–26.

Kruse, F. A., 2000: Introduction to hyperspectral data analysis. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, HI, USA.*

Krzanowski, W. and P. Jonathan, 1995: Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied statistics*, **44 (1)**, 101–115.

Landgrebe, D. a., 1980: The development of a spectral-spatial classifier for earth observational data. *Pattern Recognition*, **12 (3)**, 165–175.

Lee, C. and D. A. Landgrebe, 1993: Feature extraction based on decision boundaries. *IEEE Transactions on Geoscience and Remote Sensing*, **15 (4)**, 388–400.

Lennon, M., G. Mercier, and L. Hubert-Moy, 2002: Nonlinear filtering of hyperspectral images with anisotropic diffusion. *IEEE International Geoscience and Remote Sensing Symposium.*

Li, J., 2011: Discriminative Image Segmentation : Applications to Hyperspectral Data. Ph.D. thesis.

Li, J., J. M. Bioucas-Dias, and A. Plaza, 2012: Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, **50 (3)**, 809–823.

Li, J., P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, 2013a: Generalized Composite Kernel Framework for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, **51 (9)**, 4816–4829.

Li, Q., X. He, Y. Wang, H. Liu, D. Xu, and F. Guo, 2013b: Review of spectral imaging technology in biomedical engineering: achievements and challenges. *Journal of biomedical optics*, **18 (10)**, 100 901.

Lu, G. and B. Fei, 2014: Medical hyperspectral imaging: a review. *Journal of biomedical optics*, **19 (1)**, 1–23.

Mercier, G. and M. Lennon, 2003: Support vector machines for hyperspectral image classification with spectral-based kernels. *IEEE International Geoscience and Remote Sensing Symposium*, Vol. 00, 288–290.

Moon, T., 1996: The expectation-maximization algorithm. *Signal processing magazine, IEEE.*

Moran, M., R. Bryant, T. Clarke, and J. Qi, 2001: Deployment and calibration of reference reflectance tarps for use with airborne imaging sensors. *Photogrammetric Engineering and Remote Sensing*, **67 (3)**, 273–286.

Naes, T., T. Isaksson, T. Fearn, and T. Davies, 2002: *A user friendly guide to multivariate calibration and classification.* NIR publications Chichester.

NASA, 2010: *Introduction to The Electromagnetic Spectrum.* National Aeronautics and Space Administration, Science Mission Directorate.

Nicolaï, B. M., K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron, and J. Lammertyn, 2007: Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, **46 (2)**, 99–118.

Nocairi, H., E. Mostafa Qannari, E. Vigneau, and D. Bertrand, 2005: Discrimination on latent components with respect to patterns. Application to multicollinear data. *Computational Statistics & Data Analysis*, **48 (1)**, 139–147.

Okada, T. and S. Tomita, 1985: An optimal orthonormal system for discriminant analysis. *Pattern Recognition*, **18 (2)**, 139–144.

O'Neill, N. T., F. Zagolski, M. Bergeron, a. Royer, J. R. Miller, and J. Freemantle, 2014: Atmospheric Correction Validation of casi Images Acquired over the Boreas Southern Study Area. *Canadian Journal of Remote Sensing*, **23 (2)**, 143–162.

Osborne, B. G., T. Fearn, A. R. Miller, and S. Douglas, 1984: Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, **35 (1)**, 99–105.

Perona, P. and J. Malik, 1990: Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12 (7)**, 629–639.

Pesaresi, M. and J. Benediktsson, 2001: A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, **39 (2)**, 309–320.

Rao, C. R., 1948: The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10 (2)**, 159–203.

Rellier, G., 2002: Analyse de textures dans l'espace hyperspectral par des méthodes probabilistes. Ph.D. thesis.

Richter, R., A. Müller, and U. Heiden, 2002: Aspects of operational atmospheric correction of hyperspectral imagery. 145–157 pp.

Roger, J., B. Palagos, S. Guillaume, and V. Bellon-Maurel, 2005: Discriminating from highly multivariate data by Focal Eigen Function discriminant analysis; application to NIR spectra. *Chemometrics and Intelligent Laboratory Systems*, **79 (1-2)**, 31–41.

Sammon, J. W., 1970: An optimal discriminant plane. *Computers, IEEE Transactions on*, **(September)**, 826–829.

Schölkopf, B., A. Smola, and K. Müller, 1998: Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, **1319**, 1299–1319.

Schölkopf, B. and A. J. Smola, 2002: *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Schwartz, C. R., M. T. Eismann, J. N. Cederquist, and R. O. Johnson, 1996: Thermal multispectral detection of military vehicles in vegetated and desert backgrounds. *Aerospace/Defense Sensing and Controls*, International Society for Optics and Photonics, 286–297.

Schweitzer, D., M. Hammer, J. Kraft, E. Thamm, E. Königsdörffer, and J. Strobel, 1999: In vivo measurement of the oxygen saturation of retinal vessels in healthy volunteers. *IEEE transactions on bio-medical engineering*, **46 (12)**, 1454–65.

Shaw, G. A. and H.-h. K. Burke, 2003: Spectral Imaging for Remote Sensing. *Lincoln Laboratory Journal*, **14 (1)**, 3–28.

Smet, S., G. Sicot, and M. Lennon, 2010: Evaluation des capacités de le télédétection hyperspectrale et développement de méthodes innovantes de traitement d'images pour des applications défense en zone littorale (hyplitt). Tech. rep., , contrat de recherche DGA 2010 34 0014.

Smith, G. M. and E. J. Milton, 1999: The use of the empirical line method to calibrate remotely sensed data to reflectance. 2653–2662 pp.

Soille, P., 2003: *Morphological Image Analysis: Principles and Applications*. 2d ed., Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Tarabalka, Y., 2007: Classification of Hyperspectral Data Using Spectral-Spatial Approaches. Ph.D. thesis, University of Iceland; University of Grenoble.

Tarabalka, Y., J. Chanussot, and J. Benediktsson, 2010a: Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, **43 (7)**, 2367–2379.

Tarabalka, Y., M. Fauvel, J. Chanussot, and J. A. Benediktsson, 2010b: SVM-and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, **7 (4)**, 736–740.

Tian, Q., Y. Fainman, and S. H. Lee, 1988: Comparison of statistical pattern-recognition algorithms for hybrid processing. II. Eigenvector-based algorithm. *Journal of the Optical Society of America A*, **5 (10)**, 1670–1682.

Tilton, J., 1998: Image segmentation by region growing and spectral clustering with natural convergence criterion. *IEEE International Geoscience and Remote Sensing Symposium, Seattle, WA*.

Tilton, J. C., 2010: Split-remerge method for eliminating processing window artifacts in recursive hierarchical segmentation. Google Patents.

Tom M., M., 2005: Chapter 1. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression.

Tomasi, C. and R. Manduchi, 1998: Bilateral Filtering for Gray and Color Images. *Proceedings of the Sixth International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, 839–846, ICCV '98.

Tormod, N. and M. Bjorn-Helge, 2001: Understanding the collinearity problem in regression and discriminant. *Journal of Chemometrics*, **426 (1998)**, 413–426.

Trygg, J. and S. Wold, 2002: Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16 (3)**, 119–128.

Vain, A., S. Kaasalainen, U. Pyysalo, A. Krooks, and P. Litkey, 2009: Use of Naturally Available Reference Targets to Calibrate Airborne Laser Scanning Intensity Data. *Sensors*, **9**, 2780–2796.

Valero, S., 2011: Arbre de partition binaire: Un nouvel outil pour la représentation hiérarchique et l'analyse des images hyperspectrales. Ph.D. thesis.

Van Der Maaten, L. J. P., E. O. Postma, and H. J. Van Den Herik, 2009: Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, **10**, 1–41.

Vapnik, V., 1998: Statistical Learning Theory Wiley-Interscience. *New York*.

Vigneau, N., 2010: Potentiel de l ' imagerie hyperspectrale de proximité comme outil de phénotypage : application à la concentration en azote du blé. Ph.D. thesis.

Vigneau, N., M. Ecarnot, G. Rabatel, and P. Roumet, 2011: Potential of field hyperspectral imaging as a non destructive method to assess leaf nitrogen content in Wheat. *Field Crops Research*, **122 (1)**, 25–31.

Wang, Y., R. Niu, and X. Yu, 2010: Anisotropic diffusion for hyperspectral imagery enhancement. *IEEE Sensors Journal*, **10 (3)**, 469–477.

Weickert, J., 1997: A review of nonlinear diffusion filtering. *Scale-space theory in computer vision*, M. Haar Romeny, Bart and Florack, Luc and Koenderink, Jan and Viergever, Ed., Springer Berlin Heidelberg, lecture no ed.

Weickert, J., 1998: *Anisotropic diffusion in image processing.*

Whitaker, R. and G. Gerig, 1994: Vector-valued diffusion. *Geometry-driven diffusion in computer vision.*

Wilks, S. S., 1962: Mathematical Statistics.

Williams, P. and K. Norris, 2001: *Near-Infrared Technology in the Agricultural and Food Industries (2nd Ed.).* Amer Assn of Cereal Chemists, St. Paul, Minnesota.

Witten, D. M. and R. Tibshirani, 2011: Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **73 (5)**, 753–772.

Wold, H., 1966: *Estimation of Principal Components and Related Models by Iterative Least squares*, 391–420. Academic Press, New York.

Wold, S., H. Antti, F. Lindgren, and J. Öhman, 1998: Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44 (1-2)**, 175–185.

Wold, S., K. Esbensen, and P. Geladi, 1987: Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2 (1-3)**, 37–52.

Wold, S., M. Sjöström, and L. Eriksson, 2001: PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**, 109–130.

Xiaobo, Z., Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, 2010: Variables selection methods in near-infrared spectroscopy. *Analytica chimica acta*, **667 (1-2)**, 14–32.

Yang, J. and J.-y. Yang, 2003: Why can LDA be performed in PCA transformed space? *Pattern Recognition*, **36 (2)**, 563–566.

Ye, J., S. Member, and Q. Li, 2005: A Two-Stage Linear Discriminant Analysis. **27 (6)**.

Yuhas, R., A. F. Goetz, and J. Boardman, 1992: Discrimination among semi-arid landscape endmembers using the spectral angle mapper(sam) algorithm. *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop.*, Vol. 1.

Zhang, Z., G. Dai, C. Xu, and M. Jordan, 2010: Regularized discriminant analysis, ridge regression and beyond. *The Journal of Machine Learning Research*, **11**, 2199–2228.

Zitová, B. and J. Flusser, 2003: Image registration methods: a survey. *Image and Vision Computing*, **21 (11)**, 977–1000.

Résumé Cette thèse présente trois approches pour gérer différentes problématiques de la classification supervisée des images hyperspectrales (HS) : réduction de la dimension spectrale, combinaison de l'information spectrale et spatiale et indépendance vis-à-vis de l'éclairement. La grande dimension et forte colinéarité des données spectrales nécessitent un traitement adapté avant classification. Pour pallier à ce problème, nous proposons une approche originale de réduction de dimension supervisée utilisant les projections orthogonales. La projection est réalisée afin que les scores obtenus minimisent la variabilité intra-classe tout en préservant les distances entre classes. De plus, la méthode étant basée sur de la suppression d'information, le sur-apprentissage peut être empêché sans nécessiter une validation croisée. Ensuite, afin de combiner l'information spectrale et spatiale, nous développons une approche de régularisation spatiale sur les canaux d'images de scores obtenus de manière supervisée. Ces scores, mettant en évidence les différences entre les classes, permettent dans le domaine spatial, d'obtenir des bordures correspondant aux variations entre classes et non au bruit de fond. Par conséquent, une régularisation spatiale qui préserve les contours, appliquée aux canaux de l'image des scores, réduit la variabilité intra-classe restant et facilite la classification. Enfin, nous présentons une démarche permettant, dans le contexte de classification supervisée, de s'affranchir de la correction en réflectance préalable des images HS. En faisant l'hypothèse que les classes ont des réflectances lambertiennes, nous montrons que, après une transformation logarithmique, la différence d'éclairement correspond à une translation dans l'espace spectral ainsi que dans l'espace des scores obtenu à partir d'une réduction de dimension supervisée linéaire. Grâce à l'utilisation de la méthode de réduction de dimension supervisée, les classes forment des clusters dans l'espace réduit. Nous proposons donc une méthode d'estimation de cette translation dans l'espace des scores, robuste aux variations du nombre d'individus par classe ainsi qu'aux aux classes manquantes. Ces trois approches ont été évaluées et validées sur deux jeux de données HS réels, i.e., classification d'adventices dans les champs de blé à partir d'image HS en proxi-détection et classification d'une zone rurale à partir de données HS en télédétection.

Mots clés: Hyperspectral; Classification; Analyse multivariée; Réduction de dimension; Méthode spectrale-spatiale; Correction en reflectance

Abstract This thesis presents three approaches to deal with different issues concerning supervised classification in hyperspectral (HS) images: spectral dimension reduction, spectral spatial combination and light source independence. The high dimensionality and collinearity of spectral variables necessitate specific processing methods to be used before classification. To tackle this issue, we propose an original supervised spectral dimension reduction method that uses orthogonal projections. The projection is performed so that the obtained scores minimize the within-class variability and preserve between-class distances. In addition, since the method is based on removing information, overfitting is prevented without the need for cross-validation. Then, in order to combine the spectral and spatial information, we propose using a spatial regularization on score image channels obtained with a supervised dimension reduction method. These channels, that are built to highlight class differences, allow edges to be obtained, in the spatial domain, that correspond to the actual class borders and not to the background variability. Therefore, applying an edge-preserving spatial regularization to the channels of this score image reduces the remaining within-class variability and thus leads to an easier classification. Finally, we propose an approach that allows, in the context of supervised classification, the prerequisite reflectance correction of the HS images to be unblocked. Under the assumption that classes have Lambertian reflectance, we show that, after log-transformation, the difference in lighting corresponds to a translation in the spectral space as well as in a score space obtained through linear supervised dimension reduction. Owing to the use of a supervised dimension reduction, classes form clusters in the low-dimensional score space. Using these clusters, we propose a method to estimate the translation that is robust against an unbalanced number of samples and missing classes. These three approaches have been evaluated and validated on two real HS datasets, i.e., classification of weeds in a wheat crop using close-range HS images, and classification of a rural area using remotely-sensed HS images.

Keywords: Hyperspectral; Classification; Multivariate analysis; Dimension reduction; Spectral-spatial method; Reflectance correction