# Semi-supervised and unsupervised kernel-based novelty detection with application to remote sensing images

THÈSE N$^{O}$ 6184 (2014)

PRÉSENTÉE LE 13 JUIN 2014

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 5
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Frank DE MORSIER

acceptée sur proposition du jury:

Prof. V. Gass, président du jury
Prof. J.-Ph. Thiran, Dr M. Borgeaud, directeurs de thèse
Prof. G. A. Camps -Valls, rapporteur
Dr D. Tuia, rapporteur
Prof. R. Vidal, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

*"Ce n'est point dans l'objet que réside le sens des choses, mais dans la démarche."*

— Antoine de Saint-Exupéry, *Citadelle*

# Acknowledgements

This Thesis would not have been possible without the help of many people that surrounded me all these years at EPFL and outside.

I would like to first say a great thank to my supervisor Prof. Jean-Philippe Thiran. He believed in me from my Master project and offered me this unique opportunity to continue for a Ph.D. He always managed (aside his busy agenda!) to keep time for questions, discussions and advices. He gave me many insights on the academic world, the freedom to pursue my individual research interests, and always considered me as a colleague rather than his student (at least I felt like that!). He also introduced me to the Malagasy colleagues, I had the great luck to meet in Madagascar (Sitraka, Mampion, le "petit-booster", both Prof. Solofo's and the rest of the LGET). Thanks Jean-Philippe !

If today I wish I had not finished, it is surely because of the great atmosphere that is present in the LTS5 group and with all the people across the corridor of the SPlabs. A very special thank to Anil for supporting me all these years in our office. You have been always there, helping, anticipating and.. whistling and singing ;). We have discovered your country, your family and I am every day a bit more ashamed that we won't be able to reach your level of hospitality.. Thanks to the "girls-band" Alia, Anna, Alessandra, Elda, Eleni, Laura, Christina, Tamara, Francesca, and the few surviving boys Thomas, Alessandro, Pascal, Martin, the "new boys" Gabriel, Murat, Didrik, Sasan, Tom and Marina, the "old-timers" Virginia, Jean-Louis, Matteo, Meri, Benoit, Ashkan, Lutz, Momo, Simon, Gilles, Luigi and many more ! Thanks Rosie for managing the administrative part of the large family ! It has been a great great time that I spent with all of you, from the christmas parties to the bbq at the lake, coffee time (with some "branches") and lunch at library, I know I have been really lucky to have you all around !

I would like also to thank my co-supervisor Dr. Maurice Borgeaud, who introduced me to the world of remote sensing and followed this Thesis even remotely from Italy. I would like also to thank my Thesis jury: Prof. Volker Gass who supported me from my Master thesis, Prof. Gustavo Camps-Valls who has been a constant inspiration along this Thesis, it was great to meet you "Gustao" ! Prof. Renè Vidal who gave me many insights on the final manuscript and last but not least, Dr. Devis Tuia. Devis has been my real coach during these years and I owe him so many things, that it is difficult to list them all here. He learned me a lot on research, writing papers, defending my ideas, but also helped me when I was in difficulties as we all go through during the PhD years. Thank you so much for your constant support, but also for your friendship ! I would like to thank the

**Acknowledgements**

Lausanne Remote Sensing Group (Devis, Jelena, Mich, Giona, Tim, Matthew, Diego..), who gathered the people on remote sensing issues, with a certain bias towards machine learning I have to admit.
I would like also to thank the people at IBM for their warm welcome. Maria, it has been a great time to work with you and with Benoit, David and Nathalie. I would like also to thank the Romanian boys and Hazard from the open space, and finally Rosetta my land lady that really enlightened this period in Zurich.
I would like to thank all the friends that were around and outside EPFL for the good times together that helped indirectly the completion of this Thesis. I thank my family, my mum and my brothers for being there supporting me, and my father from up there who inspired me to always go beyond, thanks !
Finally, Jeanne I thank you so much for your support and your love. You have been there when it was difficult, when I was lost in the desert, and you kept pushing me out of the water. I love you so much for what you are and all that you have done to me. You made me the most happy father with the birth of Alice ! My daughter, aside being the most beautiful and brilliant girl in the world, had the delicacy to wait that her father finished his Thesis to arrive among us !

*Lausanne, May 2014* Frank de Morsier

# Abstract

The main challenge of new information technologies is to retrieve intelligible information from the large volume of digital data gathered every day. Among the variety of existing data sources, the satellites continuously observing the surface of the Earth are key to the monitoring of our environment. The new generation of satellite sensors are tremendously increasing the possibilities of applications but also increasing the need for efficient processing methodologies in order to extract information relevant to the users' needs in an automatic or semi-automatic way.

This is where machine learning comes into play to transform complex data into simplified products such as maps of land-cover changes or classes by learning from data examples annotated by experts. These annotations, also called labels, may actually be difficult or costly to obtain since they are established on the basis of ground surveys. As an example, it is extremely difficult to access a region recently flooded or affected by wildfires. In these situations, the detection of changes has to be done with only annotations from unaffected regions. In a similar way, it is difficult to have information on all the land-cover classes present in an image while being interested in the detection of a single one of interest. These challenging situations are called novelty detection or one-class classification in machine learning. In these situations, the learning phase has to rely only on a very limited set of annotations, but can exploit the large set of unlabeled pixels available in the images. This setting, called semi-supervised learning, allows significantly improving the detection.

In this Thesis we address the development of methods for novelty detection and one-class classification with few or no labeled information. The proposed methodologies build upon the kernel methods, which take place within a principled but flexible framework for learning with data showing potentially non-linear feature relations. The thesis is divided into two parts, each one having a different assumption on the data structure and both addressing unsupervised (automatic) and semi-supervised (semi-automatic) learning settings.

The first part assumes the data to be formed by arbitrary-shaped and overlapping clusters and studies the use of kernel machines, such as Support Vector Machines or Gaussian Processes. An emphasis is put on the robustness to noise and outliers and on the automatic retrieval of parameters. Experiments on multi-temporal multispectral images for change detection are carried out using only information from *unchanged regions* or none at all.

The second part assumes high-dimensional data to lie on multiple low dimensional structures, called manifolds. We propose a method seeking a sparse and low-rank

representation of the data mapped in a non-linear feature space. This representation allows us to build a graph, which is cut into several groups using spectral clustering. For the semi-supervised case where few labels of one class of interest are available, we study several approaches incorporating the graph information. The class labels can either be propagated on the graph, constrain spectral clustering or used to train a one-class classifier regularized by the given graph. Experiments on the unsupervised and one-class classification of hyperspectral images demonstrate the effectiveness of the proposed approaches.

**Keywords:** Machine learning, Kernel methods, Change detection, Clustering, Manifold learning, Pattern recognition, Semi-supervised learning

# Résumé

Le principal défi des nouvelles technologies de l'information est de réussir à extraire de l'information intelligible à partir du déluge de données numériques collectées tous les jours. Parmi ces données, celles provenant des satellites d'observation de la Terre sont essentielles pour la surveillance de notre environnement. La nouvelle génération de satellites accroît les possibilités d'applications, mais aussi la nécessité de développer des méthodes capables d'extraire des informations pertinentes pour les utilisateurs de manière automatique ou semi-automatique. Dans ce but, l'apprentissage automatique permet de transformer des données complexes en produits simplifiés, tels que des cartes de changements multi-temporels ou d'occupation du sol. L'apprentissage s'effectue à partir d'exemples annotés par des experts. Ces annotations, également appelées étiquettes, peuvent être en réalité difficiles ou très coûteuses à obtenir, car établies sur la base de relevés de terrain. Il est par exemple difficile d'accéder à une zone ayant été inondée ou touchée par des incendies. Ainsi la détection de changement doit être effectuée uniquement avec des informations sur des régions intactes. Le problème est le même si l'on s'intéresse à une certaine classe d'occupation du sol et que l'on ne connaît pas toutes les classes contenues dans l'image. Ces situations particulièrement difficiles en terme d'apprentissage sont appelées "détection de nouveauté" ou "classification à une classe" dans le domaine de l'apprentissage automatique. Dans ces situations, la phase d'apprentissage doit se contenter du nombre très limité d'annotations mais peut utiliser le vaste ensemble de pixels non étiquetés toujours disponibles dans les images. Cet apprentissage, appelé semi-supervisé, permet d'améliorer grandement la détection.

Cette Thèse aborde le développement de méthodes, dites à noyaux, permettant l'apprentissage avec des données montrant potentiellement des caractéristiques non-linéaires pour ces situations de détection ou de classification n'ayant que très peu d'informations a priori. Cette Thèse est divisée en deux parties, chacune ayant une hypothèse différente sur la structure des données et traitant toutes les deux d'apprentissage non-supervisé et semi-supervisé.

La première partie fait l'hypothèse d'une structure en groupes qui se superposent. L'accent est mis sur la sélection automatique des paramètres requis par les différentes méthodes proposées. Ces méthodes permettent la détection automatique et semi-automatique de changements en utilisant seulement des exemples provenant de régions inchangées.

La deuxième partie s'intéresse aux données à haute dimensionnalité, telles que les images hyperspectrales, et fait l'hypothèse d'une composition de structures à faible dimensionna-

lité, appelées variétés. Les méthodes proposées cherchent une représentation parcimonieuse et dite à faible rang des données dans un espace projeté non-linéaire. Cette représentation permet de retrouver les multiples structures correspondant à différentes classes d'occupation du sol par exemple. Lorsque seules quelques étiquettes d'une classe d'intérêt sont disponibles, plusieures approches peuvent combiner la précédente représentation des données et les étiquettes. Elles peuvent être utilisées de différentes manières : propagées en respectant la structure des données, utilisées pour contraindre l'obtention des groupes ou pour l'apprentissage d'un classificateur qui tient compte de la structure découverte. Ces approches permettent la classification non-supervisée et semi-supervisée d'une classe d'intérêt dans des images hyperspectrales.

**Mots clefs :** Apprentissage non-supervisé, apprentissage semi-supervisé, méthodes à noyaux, détection de changements, analyse par regroupements, structures à faibles dimensions, reconnaissance de formes

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| Abbreviation | Description |
|---|---|
| AATSR | Advanced Along Track Scanning Radiometer |
| ADMM | Alternating Direction Method of Multipliers |
| ANN | Artificial Neural Network |
| AVHRR | Advanced Very High Resolution Radiometer |
| AVIRIS | Airborne Visible/Infrared Imaging Spectrometer |
| CCA | Canonical Correlation Analysis |
| CHRIS | Compact High Resolution Imaging Spectrometer |
| CMS | Cluster Merging System |
| CSC | Constrained Spectral Clustering |
| CS-SVM | Cost-Sensitive Support Vector Machine |
| CVA | Change Vector Analysis |
| CVM | Cluster Validity Measure |
| DB | Davies-Bouldin index |
| DIA | Difference Image Analysis |
| DMC | Direct Multi-date Classification |
| EM | Expectation Minimization |
| ESA | European Space Agency |
| HEO | High Earth Orbit |
| HGPC | Hierarchical Gaussian Process Clustering |
| GEO | Geostationary Earth Orbit |
| GFHF | Gaussian Functions and Harmonic Fields |
| GIS | Geographic Information Systems |
| GMM | Gaussian Mixture Model |
| GP | Gaussian Process |
| ICA | Independent Component Analysis |
| IR | InfraRed |
| LEO | Low Earth Orbit |
| LLGC | learning with Local and Global Consistency |
| LPU | Learning from Positive and Unlabeled examples |
| LRSC | Low-Rank Subspace Clustering |
| LRSSC | Low-Rank and Sparse Subspace Clustering |
| KCCA | Kernel Canonical Correlation Analysis |
| KKM | Kernel $k$-means clustering algorithm |
| KKT | Karush-Kuhn-Tucker conditions |
| KLRSC | Kernel Low-Rank Subspace Clustering |
| KLRSSC | Kernel Low-Rank and Sparse Subspace Clustering |
| KM | $k$-means clustering algorithm |
| KOMP | Kernel Orthogonal Matching Pursuit |

| | |
|---|---|
| $k$-NN | $k$ Nearest Neighbours |
| KPCA | Kernel Principal Component Analysis |
| KSSC | Kernel Sparse Subspace Clustering |
| LDA | Linear Discriminant Analysis |
| LRSC | Low-Rank Subspace Clustering |
| LRSSC | Low-Rank and Sparse Subspace Clustering |
| MAP | Maximum A posteriori Probability |
| MEO | Medium Earth Orbit |
| MERIS | MEdium Resolution Imaging Spectrometer |
| MLC | Maximum Likelihood Classifier |
| MMD | Maximum Mean Discrepancy |
| MR | Missrate |
| NASA | National Aeronautics and Space Administration |
| NCS-SVM | Nested Cost-Sensitive Support Vector Machine |
| NDVI | Normalized Difference Vegetation Index |
| NIR | Near-Infrared |
| NMI | Normalized Mutual Information |
| NOC-SVM | Nested One-Class Support Vector Machine |
| NSVC | Nested Support Vector Clustering |
| OA | Overall Accuracy |
| OC-SVM | One-Class Support Vector Machine |
| OMP | Orthogonal Matching Pursuit |
| OSP | Orthogonal Subspace Projection |
| PCA | Principal Component Analysis |
| PCC | Post Classification Comparison |
| PDF | Probability Density Function |
| RBF | Radial Basis Function |
| RGB | Red-Green-Blue |
| RKHS | Reproducing Kernel Hilbert Space |
| RMSE | Root Mean Square Error |
| RS | Remote Sensing |
| SAR | Synthetic Aperture Radar |
| SC | Spectral Clustering |
| SMCE | Sparse Manifold Clustering and Embedding |
| SNR | Signal-to-Noise Ratio |
| SOSP | Spectral Orthogonal Subspace Projection |
| SSC | Sparse Subspace Clustering |
| SSL | Semi-Supervised Learning |
| $S^2$OC-SVM | Semi-Supervised One-Class Support Vector Machine |
| $S^3$VDD | Semi-Supervised Support Vector Data Description |
| STSC | Self-Tuning Spectral Clustering |
| SV | Support Vector |
| SVC | Support Vector Clustering |
| SVDD | Support Vector Data Description |
| SVM | Support Vector Machine |
| SWIR | ShortWave InfraRed |
| TM | Thematic Mapper |
| VIS | Visible range (electromagnetic spectrum) |

# Introduction Part I

# 1 General introduction

## 1.1 Motivations

The massive amount of digital data gathered every day by various sensors on and above the surface of the Earth is far from being fully exploited. The constellation of Earth observation satellites, the networks of cameras in cities or the various measure stations are among the many continuous monitoring that occurs everyday on our planet. The diversity of the sensors allows to foresee applications and products with high impact but remains a poisoned gift in terms of required processing. The actual challenge is about extracting relevant information from a deluge of complex observations, whereas less than fifty years ago, it was about trying to acquire observations [Baraniuk, 2011].

Considering only the Earth observation satellites, more than a dozen terabytes of remote sensing measurements are collected every day. With the improvements in spectral, spatial and temporal resolutions of the sensors, the amount of data continues to grow. These sensors provide a unique chance to characterize the physical processes ongoing on the surface of the Earth over large areas and various time lapses in an accurate way. This new generation of satellites fosters applications like long-term monitoring of the ecosystems, precise weather forecasting, monitoring of agricultural fields and urban growth or early warning after natural catastrophes.

However, in order to meet the requirements of these applications an important processing chain is needed to transform the measured data into valuable products for the users. The processing is challenging: first there are the multiple perturbations occurring between the surface of the Earth and the sensor and secondly, the data is complex and massive. One of the most delicate part of the processing chain is the transformation of the data into a simplified product, such as thematic maps where every pixel of an image is attributed a thematic class. This stage, at the interface between machine learning, pattern recognition and image processing, is definitely the main bottleneck for having a complete automatic processing chain.

Automatic methods still require a learning phase where a set of annotated examples from experts are used to learn the desired transformation of the data. These annotations, also called labels, may be difficult or costly to obtain in practice, leaving the machine with only a few of these annotations. Fortunately, often a large set of unlabeled samples is available at no extra cost. In these situations, the combination of both labeled and unlabeled samples, usually referred to as *semi-supervised learning*, can greatly improve the prediction accuracies in classification or detection applications, since the unlabeled samples are providing valuable information on the data distribution and its structure.

The development of automatic methodologies would be beneficial for most of the monitoring applications. As for instance, the recent programs of the leading information technology companies, namely IBM's *smarter planet* [Zhu et al., 2009] or NASA-Cisco's *planetary skin* [Castilla-Rubio and Willis, 2009] whose common goal is the global monitoring of the environment and the early-warning and mitigation of disasters. These applications have a tremendous potential for life-safety and economic benefits and will rely on the success of automatic processing methods.

## 1.2 What is novelty detection about ?

In standard classification problems, the set of examples used for training a machine is assumed representative of the different categories desired. However, often in practice the labels available are not representative of all the data at hand. This requires the ability to identify data that was not represented by examples during the training stage. This type of situations are known as *novelty detection* or *one-class classification* in the machine learning community. More precisely, novelty detection is defined as the task of detecting new or unknown data based solely on some reference data of *normal* observations. A data sample that deviates significantly from the *normal* data distribution is assumed to be issued from another underlying process and is considered a *novelty* [Markou and Singh, 2003a,b].

In machine learning, the field of novelty detection is inter-related with the problem of one-class classification. These two different denominations for a similar problem are used to emphasize the desired task: novelty detection if we are more interested in the data that differs from the *normal* observations, or one-class classification if we are more interested in the *normal* observations. Furthermore, this type of problems is closely related to anomaly and outlier detection. The nuances between these appellations, driven by the assumptions made on the data and the supervised information available, are presented in Section 4.1. The ability to discover novelties is crucial for a large set of problems encountered in many applications which has raised the interest of researchers during last decade, within the related fields of novelty, outlier and anomaly detection as observed by the number of citations in these fields, illustrated in Figure 1.1.

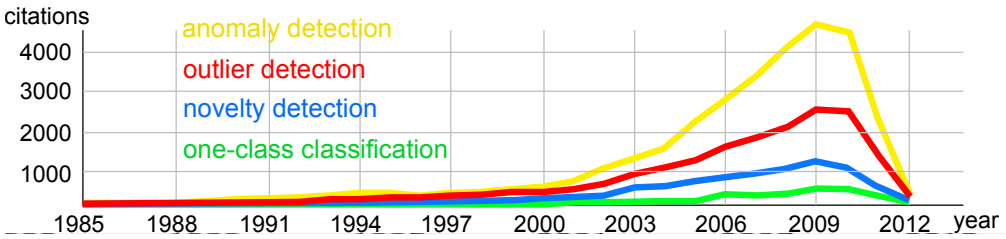
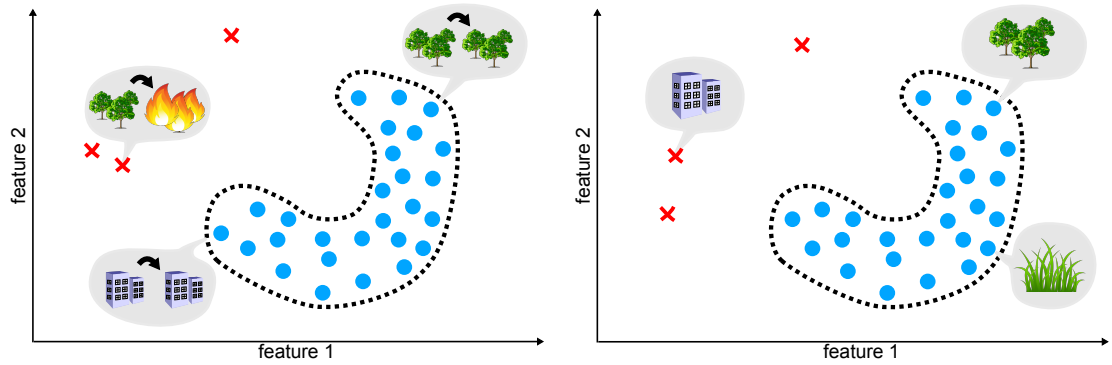



Figure 1.1: Number of citations over the years for novelty detection, anomaly detection, outlier detection and one-class classification. Source: http://academic.research.microsoft.com

We, as human beings, perform novelty detection on a daily basis. Our brain models events from the environment using a distributed neural network involving the hippocampus and many other different regions of the brain [Knight, 1996]. This network is able to detect unusual events which are potentially dangerous, by recognizing a bad prediction of the model. This would indicates that the event is different from what has already been experienced [Tiitinen et al., 1994]. The closely related topic of *concept learning*, i.e. how human beings build their rules for testing if a situation is novel, has been first investigated by psychologists [Hunt, 1962] and then transferred to the machine [Tenenbaum, 2000]. The understanding of how humans learn abstractions could help in designing models to detect a specific class of inputs. For a machine, discovering novelties is a delicate process, since it should detect something for which it has not been trained. The very first successful attempts were using auto-encoder Artificial Neural Networks (ANN) based on a model of the hippocampus, the part of the brain involved in learning and memorizing [Bishop, 1994, Japkowicz et al., 1995].

The reason why there is a lack of information about the novelties to be detected is either their rare occurrence (e.g. rare engine default), their heterogeneity (e.g. different types of land-cover changes) or the cost and difficulty in collecting them (e.g. patterns failing after an expensive lithographic manufacturing process, bank frauds, or areas that are inaccessible after a natural disaster). In many classification problems, we only possess information from one or few classes of interest and new classes may appear. Aside novelty detection, the fields of *one-shot learning* [Miller et al., 2000, Fei-Fei et al., 2006], where knowledge from other classes is used to detect a novel class, and *zero-shot learning* [Palatucci et al., 2009], where new classes are discriminated solely based on training examples from different classes, are negating the "closed world" assumption. The latter assumes that all the data at hand during training is known in terms of possible classes and that unknown classes are not allowed as outputs, which is not reasonable and unrealistic in practice [Duin and Pekalska, 2005]. Instead the "open world" assumption should be preferred [Scheirer et al., 2013]. This assumption does not presume that all the knowledge is available at training and tolerate to have some outputs unknown. This is in favour of multi-class recognition (i.e. one-vs-all), where additional classes not present during training are detected as outliers. It is more flexible than the generally adopted




(a) (b)

Figure 1.2: Novelty detection and one-class classification examples in remote sensing. (a) Multi-temporal situation where the • unchanged pixels (*trees*) are used to train the model (dashed boundary) and detect the x changed pixels (*burnt trees*). (b) Classification situation where the • examples of a class of interest (*vegetation*) are used to train the model (dashed boundary) and detect the x pixels of different classes (*building, roads, etc..*) as novelties.

strategy which discriminates pairs of classes (i.e. one-vs-one) and it is more efficient since less computationally intensive and similar in terms of accuracy [Rifkin and Klautau, 2004].

In this Thesis, we address the problem of novelty detection by introducing methodologies specifically handling the restricted supervision level and the data structure, particularly in the context of remote sensing applications. As depicted in Figure 1.2, these situations can be encountered in multi-temporal or fixed time image analysis. Both cases are essentially about finding the best description of a class of interest to better detect data of a different nature, which requires flexible and robust methods for handling the complex statistical distribution of the data; hence the use of kernel methods in this Thesis. In particular in remote sensing, the assumption of a linear relationship among the data features does not hold. As an example, the data in Figure 1.2 shows a non-linear type of relation between its features producing a banana-shaped cluster.

On the one hand, we focus on the automatic change detection in multi-temporal remote sensing images, see Figure 1.2 (a). The rapid mapping of an area and the analysis of changes are decisive in early warning systems. In such small time laps, the experts cannot provide the machine with reliable information about the changes occurring at the surface of the Earth. Thus in the completely unsupervised case where no annotated examples are available, the machine will learn only from the structure of the data to recover the changed and unchanged areas. In the case where some unchanged areas can be reliably annotated, they can be used to detect unknown changes. Furthermore, this strategy can become advantageous for heterogeneous changes that would require many annotations for a complete characterization.

On the other hand, we consider the production of thematic land-cover maps under a low level of supervision. In the completely unsupervised case, the structure of the data is retrieved and the disconnected components are identified as classes. In the case of an expert interested in one specific land-cover class and cannot provide the system with all the different thematic classes present in the covered area, an accurate description of the class of interest is learned to enable its discrimination.

In all cases, we exploit the inherent structure of the data to cope with the low level of supervision.

## 1.3 Thesis objectives

In this Thesis, we aim at addressing the novelty detection problem in situations with no or partial supervision typically found in remote sensing. We plan to address this problem using methods that respect the non-linear nature of the data to provide more accurate products.

The objectives of the Thesis can be summarized as follows:

1. Provide methods for handling non-linear data characteristics and enabling novelty detection in situations of low level of supervision, such as semi-supervised and unsupervised settings.

2. Provide parameter selection strategies that support this low level of supervision.

3. Apply these methods to change detection and one-class classification problems in remote sensing.

The organization of the Thesis is as follows: An introduction to the motivations behind this Thesis and the related contributions are presented in this first Chapter. A concise introduction to the field of remote sensing for Earth observation is presented in Chapter 2, while the required background on machine learning and the state-of-the-art approaches to novelty detection are exposed in Chapter 3 and 4, respectively. The rest of the Thesis is organized along two main axes: the first about the type of assumption on the structure of the data and the second about the level of supervision. It has been natural to group the Chapters sharing the same assumptions on the nature of the data into two coherent parts (Part II and Part III). Inside each part, a first Chapter introduces some common background notions, a second Chapter presents a fully unsupervised approach to novelty detection, while the third Chapter deals with the semi-supervised setting.

## 1.4 Thesis contributions

Here we present the contributions made in this Thesis to the aforementioned objectives and their associated publications.

**Part II** studies the use of kernel machines for data formed by overlapping clusters. Emphasis is put on the robustness to noise and outliers and on the automatic retrieval of parameters. In **Chapter 6**, the unsupervised detection of changes is approached as a clustering problem where arbitrary-shaped and overlapping clusters need to be retrieved. We present the hierarchical Support Vector clustering and hierarchical Gaussian Processes clustering. We also propose a cluster validity measure for the selection of hyperparameters and a cluster merging system adequate to the problem of change detection. This Chapter relates to [de Morsier et al., 2012b, 2014b]. In **Chapter 7**, the semi-supervised detection of novelties is based on a discriminative approach between labeled and unlabeled samples. We propose to exploit entire solution path algorithms and establish a low-density criterion providing fast parameter selection since it avoids recurring to the usual cross-validation. This Chapter relates to [de Morsier et al., 2012a, 2013c].

**Part III** investigates the use of kernel methods with new paradigms for high-dimensional data lying on multiple low-dimensional manifolds. In **Chapter 9**, unsupervised classification is performed with the construction of a graph that represents the data manifold structure. The graph aims at finding a sparse and low-rank representation of the data mapped in a non-linear feature space. The graph is cut using spectral clustering. The formulation is flexible, handles multiple manifolds and has a natural out-of-sample extension. This Chapter relates to [de Morsier et al., 2014c,d]. In **Chapter 10**, we tackle the semi-supervised situation, i.e. when few labeled examples of one class are available. We investigate different approaches to novelty detection based on a graph representing the data structure based on either label propagation, constrained clustering or semi-supervised classification.

The contributions in this Thesis are concentrated around the problem of unsupervised and semi-supervised novelty and one-class detection under appropriate data assumptions. However, the methodologies developed range from clustering with noise, arbitrary-shaped clusters/manifolds to unbalanced classification problems. They could be extended to other related pattern recognition applications, such as face and object clustering as presented in the first experiments of Chapter 9, and provide new application opportunities.

**Contributions not presented in the Thesis** During the PhD, a certain number of other works have been done but are not presented in this Thesis. These works are most often the result of international collaborations or performed while doing an internship at IBM Zurich Research Lab. We simply list them hereafter and let the interested reader check the corresponding publication.

1. Dempster-Shafer consensus clustering for multi-source thematic classification [Ranoeliarivao et al., 2013].
2. Flood detection in SAR images with non-linear kernel clustering and topographic prior [de Morsier et al., 2013b].
3. Soil moisture estimation from airborne hyperspectral imagery over tilled agricultural fields based on Support Vector Regression (SVR) [Stamenkovic et al., 2013].
4. Semi-supervised change detection as a multi-view problem based on the semi-supervised kernel canonical correlation analysis (SSKCCA) [Volpi et al., 2013].
5. Robust registration of airborne videos exploiting motion estimation [de Morsier et al., 2013a].
6. Learning the geographic features for landscape attractiveness as a geographic one-class data problem [Produit et al., 2014].
7. Domain adaptation in remote sensing via cross-image synthesis based on dictionary learning [Matasci et al., 2014].
8. Hierarchical sparse representation for dictionary-based classification of hyperspectral images [Marcos Gonzalez et al., 2014].
9. Fast detection of novel problematic patterns based on dictionary learning and prediction of their lithographic difficulty [de Morsier et al., 2014a]
10. Two patents on fast novelty detection in very large VLSI layouts by means of compressive sampling and dictionary learning (developed while at IBM).
    - Fast novelty detection based on compressed sensing and dictionary learning (filed).
    - Fast detection of novel semiconductor patterns and prediction of their lithography difficulty based on online joint dictionary and regression learning (pending).

# 2 Introduction to remote sensing

We introduce in this Chapter the main notions on remote sensing (RS) and the optical imagery.

## 2.1 A brief history of remote sensing

For centuries the Earth has only been observed from the ground. Scientific progress have made possible the observations from a distance, i.e. sensing from a remote location: the space. In the 1840s, with the very beginning of photography and cameras, balloon pilots started to take pictures of the ground. During the First World War, cameras were mounted on airplanes and provided precious strategic images of the enemy's side. With the Second World War and the beginning of aerospace activities, the possibilities to observe the Earth were increased by jetplanes, rockets and finally satellites. Photo systems were quickly integrated into satellites for recognition and cartography by U.S intelligence agencies [Richards, 2013]. Nowadays, the observation satellites are not restricted to the defence sector anymore and several commercial satellites are now operational. However, the defence industry has an important technological advance over the commercial sector. The military satellites were capable of acquiring high-resolution optical images 30 years before the civil and commercial satellites as presented in Figure 2.1.

Earth observation by satellites quickly became an important tool for monitoring short and long-term phenomena occurring at the surface or in the atmosphere. It allows the systematic study of natural resources such as biomass content, moisture analysis, aerosols, land usage and natural disasters damage assessment. Many other disciplines related to spatial information are incorporating RS products, with the broad use of Geographic Information Systems (GIS). The GIS allow to unify all geo-spatial related information (land-cover, land-use, road networks, urban zones, demography, etc..) and helps for territorial management and decision making. More recently, RS has become a tool for geo-politics and humanitarian law surveillance, helping UN peacekeeping mission and

(a) (b)

Figure 2.1: High-resolution images acquired over the Eiffel tower, Paris, by (a) the U.S. military satellite KH-7 (20/03/1966) and (b) the Quickbird satellite (02/08/2004). Thirty years ago, military satellites could achieve a similar resolution than today's civilian satellites.

providing clear evidences of human rights violations. The Earth observation imagery starts to be recognized as evidence by international criminal courts [Marx and Goward, 2013]. As an example, the destruction of villages in Darfur (Sudan) could be observed as change in landcover classification from Landsat satellite imagery [Marx and Loboda, 2013].

The democratization of RS information is principally due to the technological advance in computational power of personal computers. Everyone can access the data from their personal computer on websites such as Google maps or softwares such as Google Earth and Earthwind. This widespread distribution of data is nowadays completed by the interaction with users. They are invited to contribute via collaborative platforms such as the open-street-maps database. This new source of spatial information forms the *neogeography*, also called *crowdsourcing*. These various sources of spatial information will provide us in the near future with what is called the Digital Earth, a powerful and rich source of information on our planet for scientific and non-scientific usage [Richards, 2013].

## 2.2 Remote sensing of the Earth by satellites/airborne platforms

The RS by satellite or airborne platforms aims at producing an image of the Earth revealing some of its surface characteristics. RS instruments can be separated into two distinct groups: the *passive* instruments which rely only on the solar radiation and the *active* instruments, which emit energy towards the surface and records the backscattered

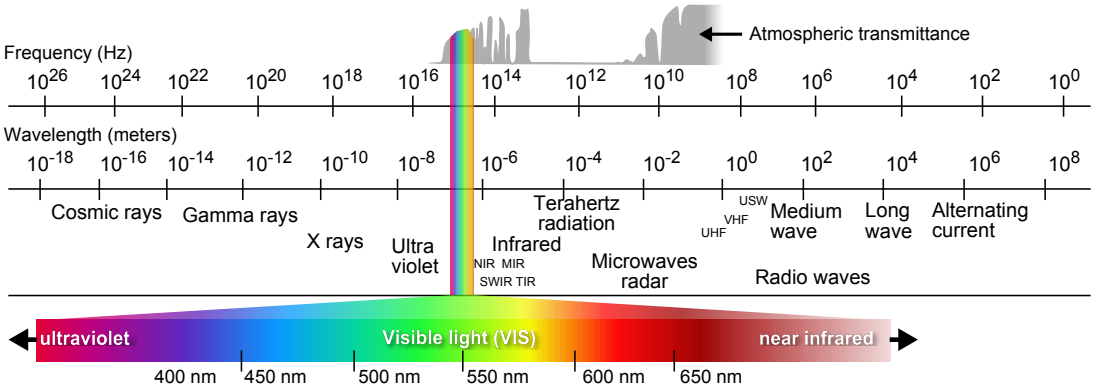



Figure 2.2: The electromagnetic spectrum is presented along a large range of frequency/wavelength. The part of the spectrum corresponding to the visible range is zoomed. The atmospheric transmittance is also reported on top.

energy. Examples of passive instruments are the multi-and hyperspectral sensors and examples of active instruments are radars and Synthetic Aperture Radars (SAR). These instruments are designed for observing a particular region of the electromagnetic spectrum, in the visible (VIS) and infrared range for passive optical instruments and in the microwave range for radars. See Figure 2.2 for a logarithm representation of the electromagnetic spectrum with its different parts. The VIS, where our human visual system is concentrated, is ranging between 400 and 700 nm as shows the zoom in Figure 2.2.

### 2.2.1 Optical remote sensing

In this Thesis we focus on the analysis of optical images which are acquired by passive instruments. Radar and SAR images are of a very different nature in terms of interpretation for the user and require specific (pre-)processing techniques out of the scope of this Thesis.

**Effect of the atmosphere** The energy captured by instrument, mounted on a platform like a satellite, corresponds to sunlight energy backscattered by the surface of the Earth. However, multiple disturbances, attenuations and additional source of energy are impacting on the received energy. The spectrum of the solar radiations arriving at the top of the atmosphere (TOA) is similar to a blackbody at temperature 5700 K, as presented in Figure 2.3. After passing through the atmosphere, the *solar irradiance* at the surface of the Earth is 1000 $W/m^2$ in average, which is 30% less than at the TOA. The backscattered radiations will again lose energy on their way back to the sensor through the atmosphere. The content of the atmosphere, a mixture of nitrogen, oxygen, argon, water vapour, ozone and aerosols varying with altitude, actually interacts with the electromagnetic radiation causing *scattering* and *absorption*.

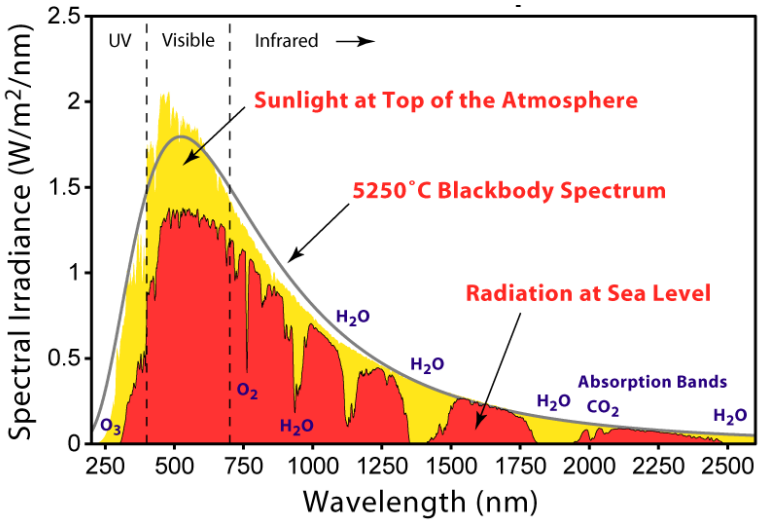



Figure 2.3: The solar radiation at the TOA and at surface of the Earth after passing through the atmosphere. The blackbody spectrum is overlayed and the absorption bands between the atmospheric windows are well observable.

The absorption of the atmosphere acts as a barrier and protects us from radiations such as ultraviolet (UV) which are harmful for the human body. However, it drastically restricts the capacity of any RS system willing to observe the Earth through the atmosphere. Hopefully, there are regions of the electromagnetic spectrum where the absorption by the atmosphere is almost non-existent. These regions are called the *atmospheric windows*. Between these atmospheric transmission windows are the *absorption bands*, corresponding most often to specific atmospheric gases. In Figure 2.3 is presented the solar irradiance which clearly shows the absorption bands. As expected from our everyday experience, the VIS is located inside the principal atmospheric window, allowing us to see the stars from the ground even if they are faint lights.

The scattering of the atmosphere has for principal effect to scatter back to the sensor solar radiation which did not reach the ground and produce a hazy effect on the images. Moreover, the scattering is spreading the energy from a zone to another and results in inaccurate brightness at the sensor.

The optical instruments are sensing in the VIS and usually extend to near infrared (VNIR or NIR), short-wave infrared (SWIR), middle infrared (MIR) or even thermal infrared (TIR) corresponding to the principal range of the sunlight energy. Towards the UV range the atmospheric transmittance decreases which prevent from sensing this part of the electromagnetic spectrum. The signals acquired in the spectral bands of the blue region ($\sim 400nm$) are hazy and only used for specific tasks related to the atmosphere and not the ground surface.

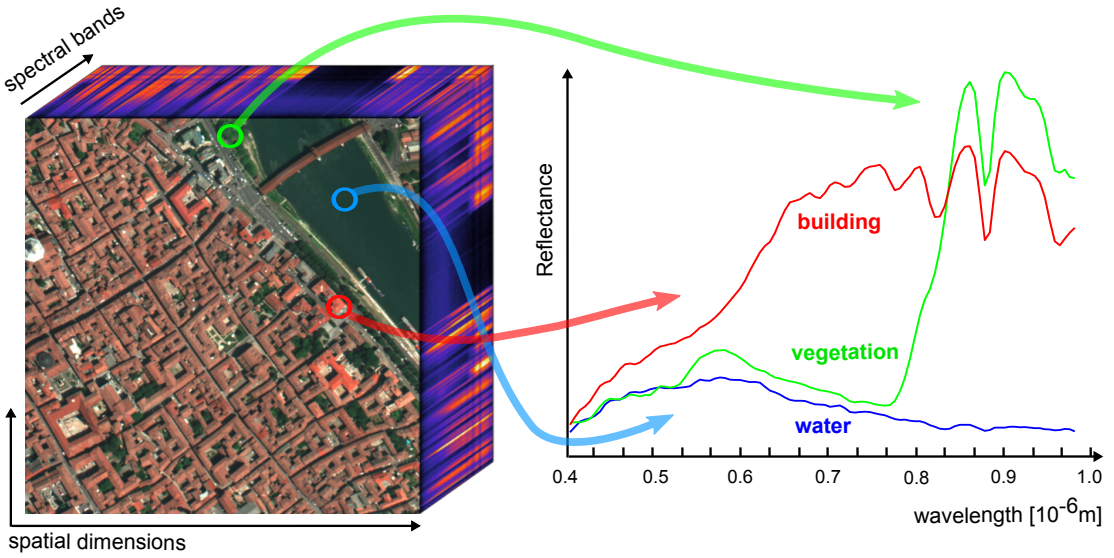



Figure 2.4: The spectral response of the surface of the Earth allows to characterize its material composition. Three different spectral signatures are presented corresponding to vegetation, water and building roofs.

**Spectral response of the ground surface** The different materials on the ground such as water, soil, vegetation, bricks or asphalt interact with the electromagnetic radiations in different ways. For example, the water will absorb most of the visible and infrared light, whereas the snow will reflect almost everything. The spectral reflectance signatures allow to characterize accurately different materials. In Figure 2.4, the spectral signatures of three different materials corresponding to vegetation, water and a building roof are represented. The three spectra show different absorption regions. The vegetation presents a very specific signature with a strong chlorophyll absorption in the visible (blue, red) and a peak of reflectance in the near-infrared region ($> 900$ nm). We can also observe the absorption band due to $H_2O$ in both the vegetation and building signatures.

### 2.2.2 Sensors characteristics

Considering the area of optical passive instruments, we describe here concisely the principal characteristics of the sensor-platform system from the data acquisition to the satellite orbit. Among the existing sensors, two main characteristics allow to categorize them: their spatial and spectral resolutions.

**Spatial resolution** The pixel size of a sensor is defined by its instantaneous field of view (IFOV). The IFOV is the finest angular resolution covered by the optical sensor at a fixed time instant. From the known platform altitude, the IFOV is directly related to the ground sampling distance (GSD), which is defined as the smallest resolvable element

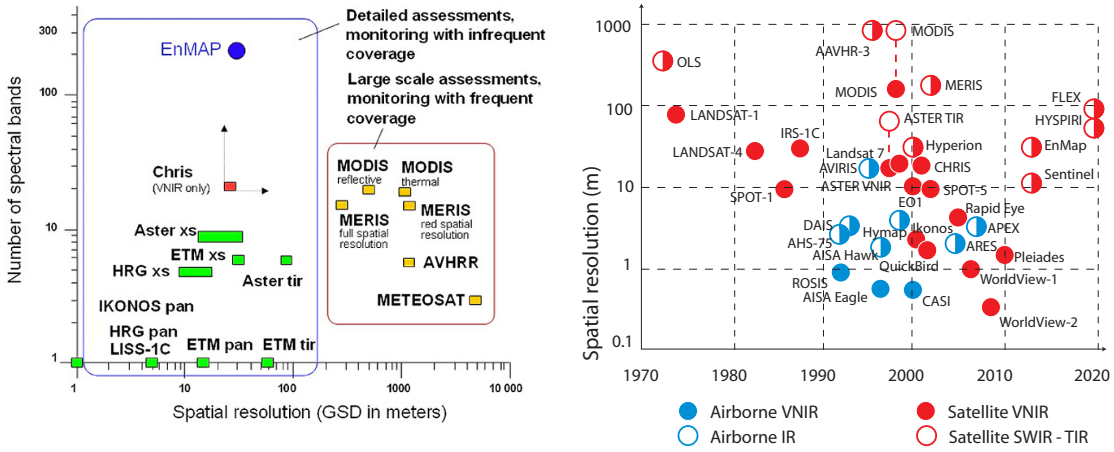
in the image [Ian Dowman and Sandau, 2012]. Most of the sensors are based on the "push-broom" techniques which scan across track as the satellite moves forward. This defines the field of view (FOV) of the sensor which corresponds on the ground to its swath width. The spatial resolution of sensors varies depending on the targeted application, low resolution ($\sim km$) for global mapping to very high-resolution ($< 1m$) for precise ground characterization, typically in urban environment. The highest spatial resolutions are at sub-meter level with WorldView-2 from DigitalGlobe. See Figure 2.5 for a detailed summary of the sensors and their characteristics over the years. The trade-off between spatial and spectral resolution can be observed.

**Spectral resolution** The sensor is designed to acquire the signal within a certain range of the spectrum and with a certain spectral resolution. The *panchromatic* images are the lowest spectral resolution images possible. A single spectral band is acquired over a broad region of the spectrum ($\sim$ 400nm, typically in the VIS). However, the panchromatic image has a high Signal-to-noise ratio (SNR) producing fine spatial resolution images. The *multispectral* images acquire data in a small number $(3 - 15)$ of large spectral bands ($\sim$ 100nm). This provides a way to start discriminating some materials. Finally, *hyperspectral* images also called *imaging spectroscopy* are made of hundreds of narrow spectral bands ($<$ 10nm). The characterization of the surface becomes extremely accurate with the distinction of fine differences in the ground composition. The hyperspectral images, or *hyperspectral cubes* (see Figure 2.4), with hundreds of spectral bands carrying information need to be simplified in order to be usable by a user.

**Radiometric resolution** The radiometric resolution is defined by the sensor system and the digital quantization of the data. The data acquired by the sensor should have the best possible SNR to describe as closely as possible the surface of the ground observed. Among the many components which affect the signal received at the sensor, we can separate the SNR into the following components [Ian Dowman and Sandau, 2012]

$$SNR = \frac{n_s}{\sqrt{\sigma_s^2 + \sigma_{fp}^2 + \sigma_{rms}^2}} \tag{2.1}$$

where $n_s$ is the number of electrons at the sensor generated by the signal, $\sigma_s^2$ is the variance of the signal electrons $n_s$, $\sigma_{fp}^2$ is the variance of the local responsivity difference (fixed pattern noise due to optical system and detector elements differences) and $\sigma_{rms}^2$ is the variance of the time-dependent noise of all other components involved. The typical range of SNR values, determining the finest difference in incoming energy detected by the sensor, is related to the sensitivity of the CCD of the sensor. This acquired signal within a certain range of distinct values is then coded into digital numbers (DN) of $n$ bits. The latter may respect the original electronic sensitivity or compress the data by quantizing it in the range $[0, 2^n - 1]$. Typically, between 8 and 16 bits are used to encode




(a) (b)

Figure 2.5: (a) Comparison of the main air/space-borne multi/hyperspectral systems in terms of spectral and spatial resolution. Credits: http://www.enmap.de. (b) Evolution of spatial resolution through the years. Credits: [Camps-Valls et al., 2011]

each spectral channel.

**Temporal resolution** Another important characteristic is the temporal resolution or *revisit time*, which quantifies the frequency at which a same region is covered by the satellite. This directly depends on the orbit of the satellite, its capability of observing off nadir and if it is part of a constellation of multiple satellites which allows to reduce the revisit time. Typical revisit times range from about 16 days for Landsat, to less than a day for WorldView 3.

**Satellites and airborne sensors** The principal orbits on which satellites are found are the low Earth orbits (LEO) between $0-2000$ km, medium Earth orbits (MEO) above 2000 km and under the geostationary Earth orbits (GEO) which are at 35'786 km. The latter orbit is in synchrony with the Earth rotation which allows to cover a fixed point at the surface. Above are the high Earth orbit (HEO). The privileged orbits for RS are the LEO which allows high spatial resolution, SNR and short revisit time, while GEO satellites allow a continuous and large coverage but achieve low spatial resolution.

Satellites for Earth observation are numerous and ensured a continuous monitoring of the Earth for more than fourty years. The Committee on Earth Observation Satellites (CEOS) reports that 286 Earth observation platforms are in orbit and roughly half of them are operational. Among others, the series of LANDSAT satellites from 1 to 8 (only 7 and 8 operational today) from the NASA is the longest Earth observation program, which started in 1972 until today. The 6 SPOT satellites and more recently the Pléiades constellation (two satellites following each other on the same orbit) from the french CNES

have been among the earliest high-resolution imagers. The ESA Sentinel program is made of constellations of pairs of satellites and had its first satellite launched the $3^{rd}$ of April 2014. Three Sentinel programs are established, each one with its specific mission and characteristics (radar, optical, etc..). The Disaster Monitoring Constellation (DMC) is made of several satellites (4 operational) mainly dedicated to disaster relief under the International Charter for Space and Major Disaster. The GOES (Geostationary Operational Environmental Satellite) are geostationary satellites dedicated to weather monitoring. In the meantime, commercial companies started to launch their own satellites. BlackBridge, ex-RapidEye is a commercial constellation of 5 identical satellites particularly efficient in vegetation health monitoring. DigitalGlobe is probably the largest commercial company owning Earth observation satellites. It owns IKONOS, Quickbird, GeoEye-1 and WorldView-1,2 & 3, the latter planned for 2014. They are among the finest spatial resolution ($\sim$ 1m) available today among the satellite multispectral sensors.

## 2.3 Remote sensing classification and change detection

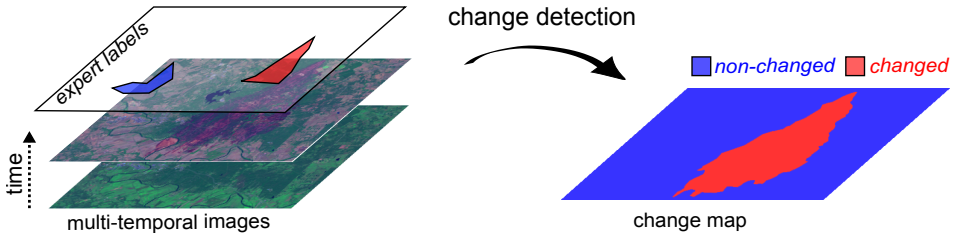
Many of the remote sensing applications are about monitoring the Earth through time and delivering products in the form of thematic maps. The repeated measures performed by the satellites allow to update ground information and build time series for monitoring. This is one of the great advantage of satellite for Earth observation, whereas airborne flight campaign are fastidious and costly with limited capability in terms of coverage. The major use cases of remote sensing images are for land-cover or land-use maps, urban expansion monitoring, deforestation, monitoring of biophysical parameters and detection of changes or damages. All these applications involve a certain number of processing steps after the image acquisition to produce the final product. They can be summarized as


Atmospheric correction aims at reducing the wavelength-dependent attenuation. The geo-registration transforms the image geometry to match a geo-referenced grid. To be able to obtain a pixel or sub-pixel correspondence, a precise co-registration of the images based on several tie points is required. The learning step is transforming the data into maps of the variables of interest (land-cover, biophysical parameters, etc...) and is the central point of attention in this Thesis.

Methodologies for learning from data are evolving very quickly and follow the increasing processing power of computers. This brings researchers to consider more and more automatic processing and to exploit pattern recognition and machine learning techniques. Moreover, the need for fast product delivery, denoted as *rapid mapping*, encourages the development of efficient automatic or semi-automatic methods.

This Thesis focuses on two learning tasks in remote sensing, which are under the novelty detection paradigm and part of two majors problematics. In change detection: the detection of unknown changes when considering multi-temporal imagery. In classification: the detection of a class of interest using hyperspectral imagery. The next Sections detail these two processing tasks.

### 2.3.1 Change detection in remote sensing




Change detection in remote sensing [Singh, 1989, Radke et al., 2005] is defined as the detection of class transitions between two or more co-registered images acquired at different time instants. It has a wide range of applications from land use change analysis, study on shifting cultivation, monitoring of pollution or urban expansion, to the assessment of the extent of deforestation, floods or wildfires burnt areas. We are going toward a near real-time monitoring of the Earth, where satellites are part of the surveillance system and provide alerts. This is crucial for example in early warning systems of natural hazards. In order to be operational, constellations of satellites are used to shorten the revisit time.

**Pre-processing** In order to obtain accurate change detection maps, certain pre-processing steps are needed.

First of all, two images taken at different time will present various differences, with some due to atmospheric conditions, sunlight angle (time) and sensor calibration, and others due to seasonal changes and land-cover changes. The latter are the change of interest in most applications, while the others can be qualified as uninteresting changes. These changes should be cancelled out to enhance the detection of the changes of interest. Several approaches are aiming at removing the impact of those changes. Atmospheric corrections methods would be ideal, however they are complex to preset and can be replaced by appropriate rescaling of the data in change detection problems [Song et al., 2001]. The centering of the images, which removes the mean on each image separately, will reduce the homogeneous perturbations part of the atmospheric disturbances. Furthermore, the spectral bands may be rescaled to unit standard deviation. This assumes normally distributed samples and will match the radiometric range of the two images. Alternatively, the spectral bands may be rescaled inside a fixed interval, e.g. $[0, 1]$. This preserves the data distribution but matches the ranges of the different spectral bands. The most used methods are the histogram equalization and histogram matching. The histogram

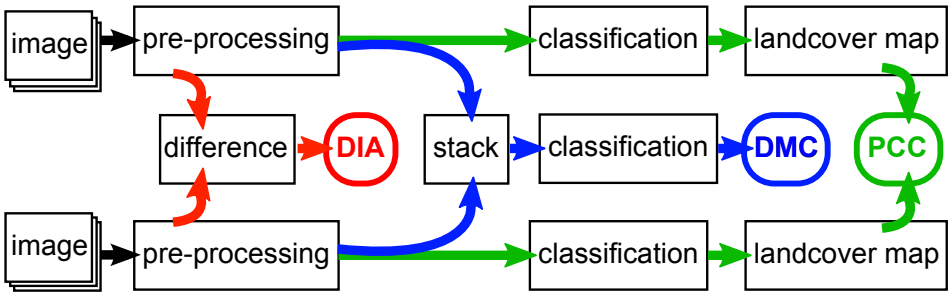



Figure 2.6: Flowchart of the three different approaches to CD: DIA, DMC and PCC

equalization may distort too much the data distributions and induce artefacts. The histogram matching will reduce efficiently the shift present between the images if matched on unchanged regions or if the changes are assumed small. Finally, the pixels may be rescaled to unit norm. This will remove their magnitude and only leave their spectral angle as a measure of difference. Some methods aim at directly detecting the changes of interest while reducing the impact of the others by a particular rescaling of the data [Theiler, 2008]. Generally speaking, the matching of two images is part of an entire field of research called *transfer learning*. Among many methods, some are adapting the images by searching for a joint subspace [Arenas-Garcia et al., 2013], while other transfer the model from one image to another. The latter has been exploited for landcover classification across time [Bruzzone and Marconcini, 2009].

**Features combination** Several combinations of the multi-temporal images before their analysis is possible, as seen in Figure 2.6.

- **PCC** Post-Classification Comparison. This approach performs the classification of each single date image separately. This allows to recover the land-cover transitions that have taken place in a certain area (e.g. vegetation → road). It is prone to cumulation of classification errors. It requires the knowledge of landcover information and is adapted to supervised approaches.
- **DMC** Direct multi-date classification. This approach combines the multi-date images into a single one by stacking the spectral information of each pixel. This has the advantage of using all the information available for classifying accurately the different changes occurring. The advantage is that information about land-cover types on the unchanged regions is preserved, however it requires some knowledge and is therefore more adapted to supervised approaches.
- **DIA** Difference Image analysis: Subtracting pixel by pixel two images at two different time instants to form a new image *difference*. As well the use of indices instead of the spectral bands can provide better enhancement of the changes. For example the *Vegetation index differencing* is the difference between the Normalized

Difference Vegetation Index ($NDVI = (NIR - Red)/(NIR + Red)$) at two time instants [Singh, 1989]. Other indices also based on bands difference and ratios exist such as the normalized burn ratio (NBR) evaluating fire burns severity [Miller and Thode, 2007] or the normalized difference water index (NDWI) revealing water bodies. These approaches based on band or indices difference are the preferred solution in unsupervised change and novelty detection since all unchanged pixels collapse towards zero [Bruzzone and Serpico, 1997].

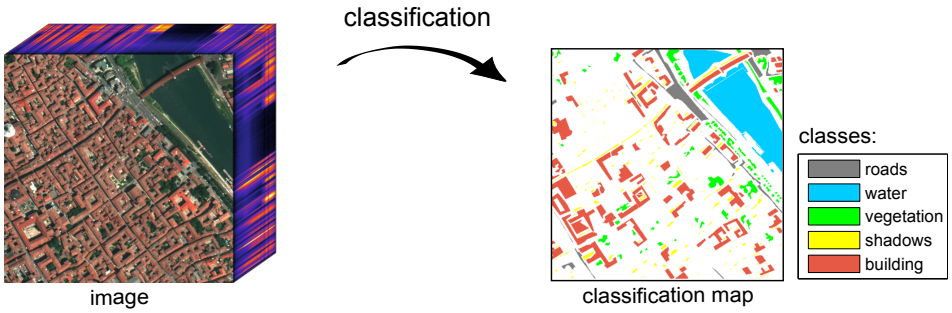
**Unknown change detection** In most situations, the lack of ground truth information remains the main challenge to overcome. The changes characteristics are often unknown beforehand, difficult to retrieve in a timely manner after the event or even difficult to model from a varying spectral signature (e.g. burnt areas, areas on fire, collapsed buildings) [Mas, 1999]. For this reason, many unsupervised change detection algorithms have been proposed in recent literature. A common approach is to transform the data in a way that highlights changed areas (DIA) and then use a threshold on the magnitude of change, like Change Vector Analysis (CVA) [Bovolo and Bruzzone, 2007b,a, Dalla Mura et al., 2008, Bovolo, 2009, Im et al., 2008, Chen et al., 2010] or Canonical Correlation Analysis (CCA) [Nielsen et al., 1998, Nielsen, 2006]. Alternatively, the difference between bi-temporal images can be represented in a feature space through the use of kernel functions [Volpi et al., 2012] or multi-scale wavelet decomposition [Celik, 2009] and $k$-means clustering performed therein.

The problem of anomalous change detection, where changes are unknown and scarce has been tackled with the use of elliptically contoured distributions [Theiler et al., 2010] or chronochrome analysis which finds patterns of spectral changes across time that are less represented [Theiler, 2008].

Novelty detection approaches, where only unchanged regions are used for training the model, have been considered for oil slick detection with SAR images using One-Class Support Vector Machines (OC-SVM) and wavelet decomposition [Mercier and Girard-Ardhuin, 2006], for landmines detection from Ground-Penetrating Radars using OC-SVM [Potin et al., 2006] or for fire detection using SVDD initialized with CVA [Bovolo et al., 2010]. More details on the methodologies behind these novelty detection approaches are exposed in Chapter 4.

### 2.3.2 Classification in remote sensing

The goal of classification in remote sensing is to derive thematic maps which are simplified representations of the Earth surface based on the image spectral content. These thematic maps are ways to summarize the complex information contained in the images in a comprehensive format for the users and are vastly used in decision-making, urban planning




or topography. The classification in hyperspectral images [Camps-Valls et al., 2014] involves complex inference and inverse problems characterized by the high spectral dimensionality, the spectral signature variations between and within classes and the small amount of labeled data (if any).

**Multi-class classification** The classification approaches have evolved together with the new types of data and the development of new machine learning methods. At first, parametric models like the Linear Discriminant Analysis (LDA) were used, but were prone to the Hughes phenomenon [Hughes, 1968] with small-size training sets and the curse of dimensionality. Non-parametric methods like $k$-NN or the decision trees could better handle more complex and non-linear distributions, before the arrival of artificial neural networks [Paola and Schowengerdt, 1995] and the supremacy of support vector machines [Camps-Valls and Bruzzone, 2009, Mountrakis et al., 2011]. The latter provided tremendous increase in performance with its robustness to noise, high-dimensionality, low number of labeled samples and intra-class variability. The new learning paradigms such as semi-supervised learning [Camps-Valls and Bruzzone, 2009],domain adaptation [Bruzzone and Prieto, 2001, Bruzzone and Marconcini, 2009] and active learning [Tuia et al., 2011] have been welcomed in the RS field with great enthusiasm. They allow to go a step further towards automatic, accurate and fast delivery of geo-spatial products.

**One-class classification** In practice, one may be interested in mapping a specific land-cover class. This land-cover should be detected independently from the other land-covers present in the image. Alternatively, one would like to have the ability of discovering unknown classes [Muñoz-Marí et al., 2007, Sanchez-Hernandez et al., 2007] or anomalies [Banerjee et al., 2009]. This allows to consider the training set as non-exhaustive and keep the possibility of having unknown output classes. One-class classifiers can be exploited either with a single-class or with multiple classes. This approach also shows the advantage of detecting outliers when a test sample is rejected by all the different classifiers. The use of the data structure can also help defining more accurate one-class classifiers [Muñoz-Marí et al., 2010]. This relates to semi-supervised learning and novelty detection presented in Section 4.3.3. A more in-depth presentation of one-class classifiers is done in Section 4.4 which reviews the state-of-the-art methods in novelty detection.

## 2.4 Summary

In this Chapter, we introduced the field of remote sensing of the Earth and the acquisition principle of passive optical sensors. We presented the processing chain of remote sensing images for classification and change detection problems. These tasks requires the help of machine learning methodologies for learning from the data with a limited interventions of human experts. The general problem of learning from the data is presented in Chapter 3.

# 3 Introduction to machine learning

*"We are drowning in information and starving for knowledge."*

— Rutherford D. Roger

In this chapter we will introduce the field of machine learning and kernel methods.

## 3.1 Learning from data

Today, the challenges are focused on how to get meaningful information from massive amounts of observations collected, while not so long ago it was already challenging to solely get observations of a single phenomenon. This new era for the data processing community is opening new possibilities and even new research fields with the unprecedented access to very diverse source of data in massive quantity. In this context, the field of *machine learning* which aims at learning information from data observations has a crucial role and should meet the expectations of this data availability. Machine learning theory started in the late fifties in an endeavour to understand brain functioning and provide machines with a degree of decision making "intelligence" [Nilsson, 1965]. The word "intelligence" has to be taken carefully here, since the machine only learns to do a desired task based on some examples. Instead of setting up specific and complex analytical models of prediction, we can let the data speaks for itself. Of course, *a priori* knowledge on the data such as the type of data features, their relationship (linear, non-linear) modeling the data structure or the type of noise are helpful to restrict the family of models describing best the data when designing our learning machine. Recalling that the goal of learning is about "separating structure from noise" [Seeger, 2001].

The field of machine learning can be divided into four main groups of learning problems: classification, clustering, regression and density estimation. These four groups are distinct in terms of their outputs (class labels, clusters, real values or probability densities), however they share the learning formalism and they may even interact. For example,

classification and clustering can rely on density estimation [Hartigan, 1975]; classification can be built on regression as in the Fisher's Linear Discriminant Analysis [Fisher, 1936]; and density estimation can be obtained from classification [Steinwart et al., 2004, Scott and Nowak, 2006]. In general, there is no advantage in using a method which has not been designed specifically for the learning task at hand: in classification, the notion of class discrimination is essential while non-existent in regression. Quoting Vapnik [Vapnik, 2000]: "When solving a given problem, try to avoid solving a more general problem as an intermediate step". As we will see in this Thesis, the specific task of novelty detection relates to density estimation and classification, and even regression models and can benefit from these approaches, which are robust to small sample size, noise and high dimensionality of the data.

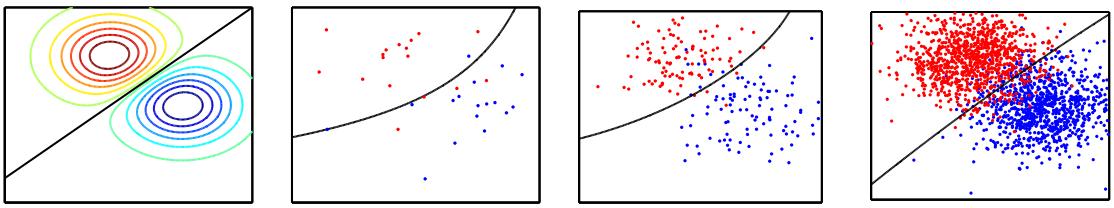
**Learning functions** Let us consider a given learning task with its associated training data pairs of input/output samples $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where the input data samples $\mathbf{x}_i \in \mathcal{X}$ are assumed to be independently and identically distributed (i.i.d) and $y_i \in \mathcal{Y}$ are the outputs. $\mathbf{x}$ and $y$ are related by an underlying joint probability distribution $P(\mathbf{x}, y)$. We would like to define a function $f : \mathcal{X} \longmapsto \mathcal{Y}$ which allows to predict the outputs from input data:

$$\begin{aligned} f_w :\ & \mathcal{X} \to \mathcal{Y} && w \in \Omega \\ & f_w(\mathbf{x}) = y \end{aligned} \tag{3.1}$$

We denote the class of functions $f_w(\mathbf{x}) \in \mathcal{F}, w \in \Omega$, where $\Omega$ is the set of parameters $w$ determining all possible functions from a chosen class and $\mathcal{F}$ is the space of all prediction functions. In principle, the class of functions $f_w$ is chosen in advance depending on the problem at hand and only a few parameters need to be determined. For a given class of functions, what we call their complexity is typically fixed. However, taking the example of the class of polynomials, the complexity will increase with the degree $d$. Other examples of classes of functions are the Gaussian mixture models (GMM), Support Vector Machines (SVM), decision trees, etc.

**Loss functions** The optimal parameters $w^*$ for the chosen class of functions $f$ can be obtained by minimizing a loss function $\mathcal{L}(f_w(\mathbf{x}), y)$ on the training data. The loss function $\mathcal{L}$ is a positive function that measures the discrepancy between the prediction $f_w(\mathbf{x})$ given by the learning machine and the true output $y$.

Different loss functions are used with different learning tasks. In classification, with $y$ being the class label $\{+1, -1\}$, the typical loss functions are the $0-1$ loss: $\mathcal{L}_{01} = \mathbf{I}\left(f_w(\mathbf{x}) \neq y\right)$ or the hinge loss: $\mathcal{L}_{hinge} = \max\left(1 - y \cdot f_w(\mathbf{x}), 0\right)$ a one-sided error function which saturates at zero for lower values. It is used in the standard SVM [Schölkopf and Smola, 2002].



(a) (b) (c) (d)

Figure 3.1: (a) True joint probability $P(\mathbf{x}, y)$. Estimation with (b) 17 samples per class, (c) 100 samples per class and (d) 1000 samples per class. The decision boundary is better approximated with a larger number of samples.

Finally, the quadratic loss: $\mathcal{L}_{quad} = (y_i - f_w(\mathbf{x}))^2$ which has the advantage of being differentiable [Rosasco et al., 2004].

The expected value of the loss is given by the following *risk functional*

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(f_w(\mathbf{x}), y)\, \mathrm{d}P(\mathbf{x}, y) \tag{3.2}$$

which is evaluated for a chosen class of function $f$. The optimal function $f^*$ with parameters $w^*$ is the one minimizing this risk functional

$$f^* = \arg\min_f R(f) \tag{3.3}$$

The desired function $f^* = f_{w^*}(\mathbf{x})$ requires the knowledge of the true joint distribution $P(\mathbf{x}, y)$.

In classification problems, the optimal classifier is the Bayes decision rule [Duda et al., 2012], which relies on the knowledge of the joint distribution $P(\mathbf{x}, y)$. It is stated as

$$f_{bayes}(\mathbf{x}, P) = \begin{cases} +1, & \text{if } P(+1|\mathbf{x}) \geq P(-1|\mathbf{x}) \\ -1, & \text{if } P(+1|\mathbf{x}) < P(-1|\mathbf{x}) \end{cases} \tag{3.4}$$

This is the theoretical optimal rule for classification [Bishop and Nasrabadi, 2006], but it is unrealistic in practice. This is because in almost all learning tasks, the joint distribution is unknown and can only be accessed via the limited training set $\{(\mathbf{x}_i, y_i)\}_i^N$ available, which is assumed to be representative of the underlying process that generated the data. An example of the optimal Bayes classifier along with the probability density functions estimated on different training set sizes are presented in Figure 3.1. The effect of the training set size is evident by observing the approximation of the two-class boundary.

**Empirical risk minimization (ERM)** Without access to the joint probability distribution, the risk $R(f)$ can only be evaluated from the training set. The *empirical risk* is

the evaluation of the average risk on the training data, i.e.,

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f_w(\mathbf{x}_i), y_i) \tag{3.5}$$

This is the so-called *training error* for the given training data set $\mathcal{T}$.

Let us take the example of density estimation, since it relates to the novelty detection paradigm studied in this Thesis. An appropriate class of function is the set of densities $p(\mathbf{x}, w)$ described by the parameters $w \in \Omega$. The probability that the data is drawn from a density $p$ with parameters $w$ is

$$P(\mathbf{x}|w) = \prod_{i=1}^{N} p(\mathbf{x}_i, w), \tag{3.6}$$

which is called the *likelihood function*. The associated loss function is

$$\mathcal{L}(p(\mathbf{x}, w)) = -\ln p(\mathbf{x}, w), \tag{3.7}$$

which corresponds to the negative log-likelihood, is broadly used in fitting probabilistic models. The *empirical risk* with this loss function becomes

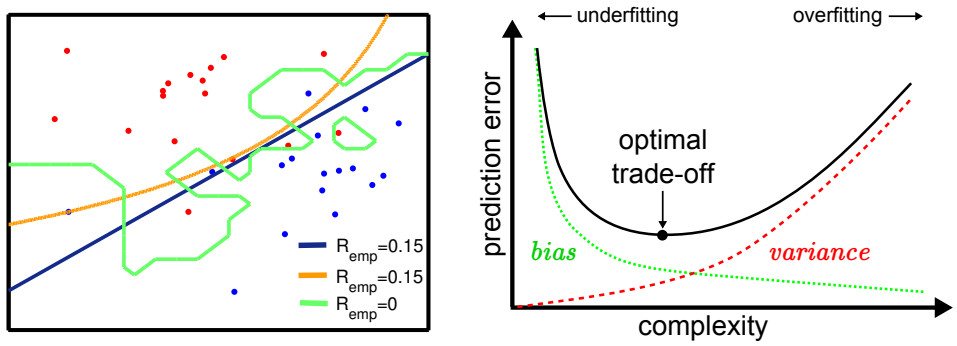
$$R_{emp}(p, w) = -\frac{1}{N}\sum_{i=1}^{N}\ln p(\mathbf{x}_i, w) \tag{3.8}$$

which is then minimized (ERM). This corresponds to the maximum likelihood estimate and provides the optimal parameters $w^*$, which completely describe the density $p(\mathbf{x}, w^*)$.

We see that the ERM principle is very general and can be adapted to any learning task where a loss function can be defined. It has the asymptotic *consistency* property, which ensures that $R(f, w^*)$ and $R_{emp}(f, w^*)$ converge to the true risk as the data set size grows to infinity. However, for small training set sizes the empirical risk is expected to be $R_{emp}(f, w^*) < R(f, w^*)$ because the ERM will lead to an exact fit of the training data. In this more realistic situation, the model obtained from the ERM is very likely to give an erroneous prediction on unseen test samples. The prediction ability of the function $f$ outside the training set, called *generalization*, needs to be controlled in some ways by introducing the notion of regularization.

## 3.2 Generalization

The ERM principle does not ensure that the model is able to generalize to unseen data. The notion of *generalization* performance for a model will be related to its complexity which can be controlled by restricting the class of functions or performing regularization




(a) (b)

Figure 3.2: (a) Trade-off between fitting the data ($R_{emp}$) and the model complexity in a classification example in $\mathbb{R}^2$. (b) The bias-variance trade-off functions of the model complexity. For complex models the bias is low while the variance is large. The optimal trade-off is found for a certain model complexity.

during training. Here are presented the different notions and approaches related to the generalization of a model.

**Occam's razor** The idea to penalize models that are too complex and more likely to be too specific and fail on a new test set can be viewed as an interpretation of the Occam's razor principle. Willam of Occam was a british philosopher and a Franciscan monk in the $13^{th}$ century who brought the idea that the simplest explanation of a phenomenon is usually the right one, provided that it fully explains the phenomenon. When different models fitting well the data exists, the simplest model is the one to be chosen, ensuring a better generalization. The underlying assumptions being that meaningful information is hidden in the data regularities whereas noise corresponds to irregularities [Rissanen, 1987]. The trade-off between goodness-of-fit and the model complexity can be observed in a small example of two-classes classification in Figure 3.2 (a). Three decision boundaries are reported with only one (light green) having a null $R_{emp}$. However, it is obvious that this boundary will not be able to predict correctly on an unseen test set and that the other boundaries should be used instead.

**Bias-variance trade-off** For a certain class of functions $f$ trained on different training sets issued from the same underlying distribution in $\mathcal{X}$, the optimal parameters $w^*$ obtained have a certain instability (variance) and provide a certain prediction accuracy (bias, equivalently $R_{emp}$). These two notions are in opposition to each other and lead to the bias-variance trade-off [Cherkassky and Mulier, 2007] in function of the complexity of the class of functions (model). The optimal complexity is found for functions showing a low bias (good prediction) and a not too high variance (complexity). The two extreme cases of model complexity are called *underfitting* and *overfitting*. As observed in Figure 3.2 (b), when

underfitting occurs, the functions $f$ are not flexible enough to appropriately characterize the data producing a high bias in prediction, even with an infinitely large amount of samples. By opposition, overfitting occurs for too flexible functions, which perfectly fit the training data but poorly predict new test examples due to the important variance in the model parameters. The problem of overfitting is more important in practice, since flexible functions are available and the training set is usually limited. This trade-off between bias and variance corresponds to the trade-off between fitness of data and model complexity.

**Curse of dimensionality** The problem of overfitting is enhanced when the dimensionality of the data increases [Duda et al., 2012, Cherkassky and Mulier, 2007]. High-dimensional data presents very particular statistical and geometrical properties. For instance, the geometry of high-dimensional data is such that hyper-dimensional objects have more surface area than volume. Statistically speaking, if a number of samples $n$ is needed to reliably estimate a function in $\mathbb{R}^1$, then the number of samples needed to maintain the data density in $d$ dimensions is $n^d$. Therefore, the sample size grows exponentially with the dimensions. Moreover, as the dimensionality increases the similarity among the samples will decrease and prevent from determining groups of similar samples. They will all look equally distant from each other since located in empty neighbourhoods of the space.

The bias-variance dilemma and curse of dimensionality can be leverage either by dimensionality reduction, or by introducing a prior restricting the choice of class of functions $f$. This regularization will help in designing functions $f$ with a lower complexity, while at the same time keeping the bias low and the variance bounded. In practice, such priors take the form of a direct penalization of the model complexity and are placed directly in the objetive of the risk to minimize.

**Regularization** As stated previously, the ERM principle will converge to the true risk $R(f)$ for an increasing training set size and for an appropriate class of function $f$ (complexity) [Vapnik, 2000]. This can be stated as

$$\lim_{N\to\infty} P\{\sup_{f\in\mathcal{F}}(R(f) - R_{emp}(f)) > \epsilon\} = 0, \quad \forall\epsilon \tag{3.9}$$

since in practice the training set size is finite and limited, the empirical risk cannot be used alone for model selection. A good generalization on new test samples is based on a trade-off between two major ingredients: the *empirical risk* and the *complexity* of the class of functions. This intuition based on Occam's razor principle and supported with the bias-variance dilemma, can be formalized as a regularization problem of the form

$$R_{reg}(f) = R_{emp}(f) + \lambda\Psi(f) \tag{3.10}$$

where the minimization of $R_{reg}(f, w)$ provides the optimal parameter $w^*$ for a certain trade-off between the fit of the training data $R_{emp}(f, w)$ and the model complexity, evaluated by a general penalization term $\Psi(f, w)$ and regulated by $\lambda$. For $\lambda = 0$, the regularization will be ignored and the risk will be the empirical risk, while for a large $\lambda$ the data will be ignored and the risk will only reflects the complexity of the model.

How can we measure the model complexity ? Occam's razor principle has found different forms in different contexts: in physics it relates to the number of constants, in signal processing to the smoothness of the functions [Unser, 1999], while in information theory it is based on the data description length, where the Minimum Description Length (MDL) principle [Rissanen, 1985] favors the most compressed data description. In the case of Bayesian inference, the penalization comprises the number of free parameters and an estimate of the noise variance. These regularized risk estimates are the well-known Bayesian information criterion (BIC) and Akaike information criterion (AIC) [Cherkassky and Mulier, 2007]. Considering the problem of the class of linear functions $f_w(\mathbf{x})$, the regularization term $\Psi$ is a norm on the coefficients $w$, giving the following regularized risk,

$$R_{reg}(f) = R_{emp}(f) + \lambda \|w\|_q \tag{3.11}$$

In the particular case of linear regression, the empirical risk is the evaluation of the Frobenius matrix-norm of the reconstruction error: $R_{emp} = \|y - \mathbf{x}w\|_F$. The $q$-norm regularization can take different forms based on different assumptions. The ridge regression is penalizing large weights using the $\ell_2$ norm [Hastie et al., 2001], related to the Tikhonov regularization for ill-posed problems [Tikhonov, 1995]. Recently, regularization norms promoting sparsity in the model ($\ell_{0<q<1}, \ell_*, \ell_{1,2}$) have been introduced. For instance the LASSO regression which sets most of the weights to zero and leads to more interpretable models [Tibshirani, 1996]. This type of regularization also helps to unveil certain structures in the data for classification [Wright et al., 2009], clustering [Elhamifar and Vidal, 2013, Liu et al., 2013] and feature extraction [Candès et al., 2011]. Finally, it has to be said that the problem of regularization is broader than just simply penalizing the complexity of a model and therefore exposed here only from a model selection point of view.

**Structural Risk Minimization (SRM)** With the rise of the Statistical Learning Theory (SLT) [Schölkopf and Smola, 2002, Bousquet et al., 2004], the notion of *capacity* of a class of functions $f$ has been introduced for model selection in learning problems. The SLT and more precisely the VC theory provide probabilistic bounds on the expected risk error for any function class [Evgeniou et al., 2002]. These bounds help in restricting the choice of functions $f$ and are based on the so-called *function capacity*, which measures

the function space complexity. These bounds are in the form of

$$R(f) < R_{emp}(f) + \phi\left(\sqrt{\frac{h}{N}}, \eta\right) \tag{3.12}$$

where $h$ is the capacity defined by Vapnik and Chervonenkis [1971] as the VC-dimension, $\phi(\cdot)$ is an increasing function of $h/N$ and $1-\eta$ is the probability under which the bound holds true. A detailed form of $\Phi$ can be found in Cherkassky and Mulier [2007]. The *structural risk minimization (SRM)* is optimizing for both the empirical risk and the structural term $\phi$. The latter will grow with the function capacity but decrease with a larger number of training samples. The VC-dimension counts the maximum number of objects that can be learned without error independently of their output values $y$ [Vapnik, 2000]. For example, the capacity of a linear classifier in $\mathbb{R}^2$ is $h = 3$ and generalizes to $h = d + 1$.

**Cross-validation** If we could have some independent test data $\{(\mathbf{x}_i, y_i)\}_i^{N_{ts}}$ for validation, the empirical risk could be evaluated as

$$R_{cv}(f, w) = \frac{1}{n_{ts}} \sum_{i=1}^{N_{ts}} \mathcal{L}(f(\mathbf{x}_i, w), y_i)) \tag{3.13}$$

However, as previously mentioned the availability of a validation set is very unlikely in practice. It is difficult to put aside some samples for validation while the training set is already limited. The idea behind cross-validation is to artificially divide the data into multiple folds, which are used alternatively as training and validation set. The learning machine is optimized with the training fold and performs predictions on the validation fold. The predictions are compared with the outputs $y$ left aside and provide an estimate of $R_{cv}$. In order to encompass all the information present in the original training set and not rely on a single partitioning of the data, each fold is alternatively used for validation. This procedure is the $k$-fold cross-validation, where $k-1$ folds are used for training and 1 fold for validation. This is repeated for all the folds combinations. Finally, the estimation of the generalization of the model is evaluated based on the average $R_{cv}$ over all the folds. In practice, the number of folds $k$ are between 5 and 10. The particular case where $k = N$ is the *leave-one-out* cross-validation. It is less used in practice since computationally intensive due to the number of folds which grows with the number of samples. Related techniques are *bootstrapping*, which creates folds by sampling the training data with replacement.

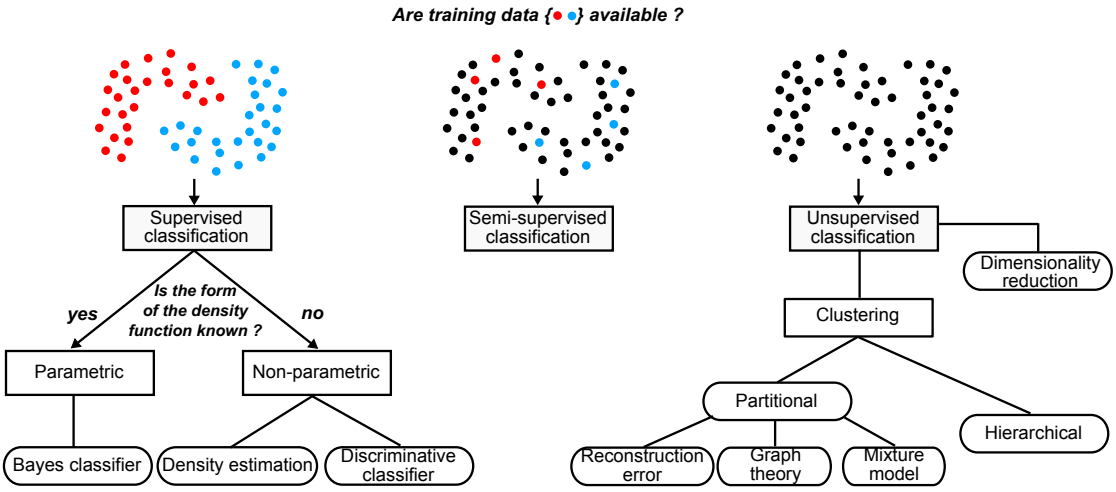



Figure 3.3: The three main learning paradigms and their sub-categories in the context of classification. Modified from [Camps-Valls and Bruzzone, 2009].

## 3.3 Learning paradigms

In the first part of the Chapter, we have seen how we can learn from data in general, for the different principal learning tasks (classification, regression, density estimation), all assuming a set of input/output data examples, the so-called *supervised learning* paradigm. Only the nature of these outputs and the respective loss functions differ between them. However, there are situations where the data available is only made of the input samples without any output examples at all, namely the *unsupervised learning*. The situations at the interface where both examples of input/output paired samples and input-only samples are available are called *semi-supervised learning*. These three paradigms are illustrated in Figure 3.3 and described in the following Sections, together with some of the state-of-the-art algorithms in the context of classification.

### 3.3.1 Supervised learning

In supervised learning, the function $f$ is trained from given example pairs of input/output data samples $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. The learned function is able to predict on unseen data. The most known supervised learning tasks are classification ($y \in \{-1, 1\}$), and regression ($y \in \mathbb{R}$) [Hastie et al., 2001], but supervised methods can also be used for feature selection [Guyon and Elisseeff, 2003] and extraction [Guyon, 2006]. In supervised classification, two main groups of methods can be identified: the *parametric* classifiers which assume a certain form of the data probability density function $P(\mathbf{x})$, and the *non-parametric* classifiers, which make no assumption on the form of the density. One typical parametric classifier is the Maximum Likelihood Classifier (MLC) which learns the mean and covariance matrix of a Gaussian distribution over each class [Scott and Symons, 1971]. If only the diagonal of the matrix is learned, we call it the Naive Bayes

classifier. The class of a test sample is inferred from the *maximum a posteriori* (MAP) over the different classes. The $k$-NN classifier is a *non-parametric* and *memory-based* classifier, which computes the distance from a test sample to all the training samples to attribute the label of the majority among the $k$ nearest neighbours. Although simple and efficient, these models fail in situations of small training sets, high dimensionality and noise. More sophisticated methods have been developed, among which the Artificial Neural networks (ANN) and the Support Vector Machines (SVM), that became extremely popular. The SVM is presented in more details in Section 3.4.4.

### 3.3.2 Unsupervised learning

In unsupervised learning, the data samples are not accompanied with their outputs, i.e. $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^N$ are the only data available and the output examples $y$ are unknown. Two different types of unsupervised methods exist: *dimensionality reduction* and *clustering* methods.

Dimensionality reduction is about finding few relevant dimensions in the data, which can be seen as a feature extraction step. The dimensionality reduction methods are learning either a subspace of the data like Principal Component Analysis (PCA) [Hotelling, 1933] and Canonical Correlation Analysis (CCA) [Hotelling, 1936], or a manifold like ISOMAP [Tenenbaum et al., 2000], Locally Linear Embedding (LLE) [Saul and Roweis, 2003] and Laplacian Eigenmaps (LEM) [Belkin and Niyogi, 2003]. It is worth mentioning that supervised dimensionality reduction methods exist such as Linear Discriminant Analysis or Partial Least Square [Hastie et al., 2001].

Clustering is about discovering structures or groups of similar samples in the data based on a similarity criterion. The definition of the similarity being the most important piece of a clustering algorithm, from which highly depends the clustering quality and usefulness for the user. Aside the definition of similarity, the data features are of prime importance when performing clustering. A semantic class defined by a user is not meant to be equivalent to a cluster in the data. This only becomes true when the data features are carefully designed for the user needs and when the similarity criterion reflects the class similarities. In situations where this assumption is granted, meaning that the clusters or groups of clusters are reflecting the classes of interests, the clustering algorithms will be able to provide the user with classes of interest. The smoothness assumption described in the semi-supervised learning Section 3.3.3 is based on the same a priori on the data features.

The problem of clustering has been widely studied in various application fields where semantic information on the data is not existing or available: grouping sequence of genes in molecular biology, segmentation of images, classification of diseases in medicine or the unsupervised classification of patterns.

The clustering algorithms can be characterized by their ability to handle different types of problems [Kononenko and Kukar, 2007], such as: i) clusters of arbitrary shapes, ii) noise, outliers and missing data, iii) *user-friendly* parameters and usage, iv) high data dimensionality, v) incremental data and vi) large scale problems. Different clustering algorithms have been proposed for the different needs emphasizing some of the aforementioned aspects. In this Thesis, we attack the first four aspects and leave the incremental and scalability aspect aside. The scalability issue is a general problem, which is addressed in the same manner for a large class of problems and benefits from the quick advances performed in fields such as data mining or computer science in general.

As depicted in Figure 3.3, the clustering algorithms can be sub-dived into two groups: i) partitional algorithms and ii) hierarchical algorithms [Jain et al., 1999]. Partitional algorithms produce a single partition, while the hierarchical algorithms produce a nested series of partitions. Among the most popular partitioning algorithms are the $k$-means [Hastie et al., 2001], ISODATA [Ball and Hall, 1965], Expectation-Maximization (EM) [Dempster et al., 1977], Self-Organizing Maps (SOM) [Kohonen, 2001], Hebbian learning based on neural networks [Sejnowski and Tesauro, 1989] and Spectral Clustering (SC) [Von Luxburg, 2007] based on data graphs. Clustering methods deriving from supervised learning methods have been also proposed such as the Support Vector Clustering (SVC) [Ben-Hur et al., 2002] or the Gaussian Process Clustering (GPC) [Kim and Lee, 2007], which are investigated in Chapter 6. Hereafter, we briefly describe the $k$-means and SC algorithms.

**$k$-means** The $k$-means algorithm is probably the most widely used partitional clustering algorithm. It has a single hyperparameter: the number of clusters $k$. $k$-means minimizes the sum-of-square distances among the cluster center and the samples attributed to that center. Given $N$ samples and $k$ cluster centers, the within-group scattering to be minimized is

$$S_w = \frac{1}{N} \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \mu_j\|^2 \tag{3.14}$$

with $C_j$ the set of samples corresponding to cluster $j$ and $\mu_j$ its center. The partition is found in an EM-like fashion by iteratively i) attributing cluster labels given cluster centers and ii) updating the centers given the labeled samples. This results in a partition of the input space into Voronoï cells defined by the centers.

**Spectral clustering (SC)** The SC algorithm is a very popular clustering method, with theoretical foundations and connection to graph-cut problems. It shows good performances and can deal with non-convex cluster shapes, where most of the standard clustering algorithms, such as the $k$-means, assume that the clusters are convex sets of samples. SC rely on a graph $\mathcal{G} = (V, E)$, where $V$ is the set of vertices representing

**Algorithm 1** Spectral Clustering (SC) [Ng et al., 2002]

**Input:** unlabeled samples: $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^{N}$, number of clusters desired $k$

1: Build a graph $\mathcal{G}$ using either

- a $k$-NN graph, which links a sample to its $k$ nearest neighbours
- an $\epsilon$-ball graph, which links a sample to neighbours within a ball of radius $\epsilon$.

2: Compute the graph edge weights $W$. Either with dummy $w_{i,j} = \{0, 1\}$ or with the heat kernel $w_{i,j} = \exp \frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}$.

3: Compute the normalized symmetric graph Laplacian $L = \mathbf{I} - D^{-1/2} W D^{-1/2}$, where $D = \sum_i W_{i,j}$ is the degree matrix and $\mathbf{I}$ the corresponding identity matrix.

4: Compute the $k$ first eigenvectors $V \in \mathbb{R}^{N \times k}$ of the Laplacian $L \approx V \Sigma V^T$

5: Normalize the rows of $V$ to unit norm: $\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}$

6: Perform $k$-means clustering on the embedded samples $\mathbf{v}_i$, the rows of $V$.

**Output:** Samples clustered in $k$ groups

the data samples and $E$ the edges linking each node. The graph can be built using the affinity of each sample and its $k$ nearest neighbours ($k$-NN graph) or its neighbours within an $\epsilon$ distance ($\epsilon$-ball graph). The graph edge weights can be simply 0 or 1 if the edge between two samples exists, or using the heat kernel [Von Luxburg, 2007]. Then, the spectral embedding of the graph is obtained by the eigenvalue decomposition of the graph Laplacian. The spectral embedding is based on the top $k$ eigenvectors that describes the major disconnected components in the graph. The clusters can be retrieved by finally performing the $k$-means algorithm in this spectral embedding. The full SC algorithm proposed by Ng et al. [2002] is described in Algorithm 1.

The SC algorithm is linked to the graph-cut problem. The minimal graph cut balanced by the number of vertices in each partition is the Ratio cut [Hagen and Kahng, 1992] and its relaxed solution is obtained by the unnormalized spectral clustering [Von Luxburg, 2007], while the minimal cut balanced by the volume (sum of edge weights) of each partition is the Normalized cut [Shi and Malik, 2000] and its relaxed solution is obtained by the normalized spectral clustering [Ng et al., 2002].

SC is extremely flexible since working with any graph built on the data. However, this brings up the open problem of building a meaningful graph on the data. The use of the $k$-NN or $\epsilon$ graph has interesting connection to the Laplace-Beltrami operator on Riemannian manifolds and kernel approaches such as the kernel $k$-means [Kulis et al., 2009] or kernel PCA [Ham et al., 2004]. A fully-connected graph with weights from a heat kernel is equivalent to a Gaussian RBF kernel and the $\epsilon$-ball graph can be seen as a truncated kernel matrix. In Chapter 9, we investigate a new approach to build a graph revealing the multi-manifold structure of certain high-dimensional data.

### 3.3.3 Semi-Supervised learning

In Semi-Supervised Learning (SSL), the training data $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^{N}$ is composed of two parts: a (small) labeled set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N_l}$ of input/output pairs, and a potentially larger set $\{\mathbf{x}_i\}_{i=N_l+1}^{N_l+N_u}$ of unlabeled samples. The rationale behind this setting is that in many applications the labeled samples are costly to gather while unlabeled samples are available at not extra cost. Thus, SSL is not a trade-off between supervised and unsupervised learning, but a wise combination aiming at performing better than supervised and unsupervised learning considered alone [Seeger, 2001, Chapelle et al., 2006].

An early attempt at characterizing the benefits from using unlabeled samples in learning was made by Valiant [1984]. At same time that SSL was developed, the related topic of constrained clustering [Basu et al., 2008] or clustering with side-information [Xing et al., 2002] was investigated in parallel. In the latter, pair-wise relationship between samples are available instead of labels. These relationships are in the form of *must-link* and *cannot-link* constraints between samples and help to guide the clustering process towards user-related clusters [Blum and Chawla, 2001, Kononenko and Kukar, 2007]. These constrained clustering approaches are investigated in Chapter 10.

The different approaches to SSL can be separated into two main categories: *transductive* and *inductive* learning. In transductive learning, one wants to predict only on the unlabeled samples in the test set. This setting is sometimes referred to as *pattern discovery*. In opposition, inductive learning aims at finding the prediction function $f$ defined over the whole input space $\mathcal{X}$.

One may want to know under which conditions the output of SSL will be more accurate than standard supervised learning, i.e. in which conditions the knowledge of $p(\mathbf{x})$ obtained from the unlabeled samples helps in the prediction of $p(y|\mathbf{x})$. These conditions under which SSL works take the form of different assumptions [Chapelle et al., 2006]. These assumptions are summarized, together with the related Parts of the Thesis exploiting them, in the following table

| Assumption | Description | In the Thesis |
|---|---|---|
| **Cluster** | if points $\mathbf{x}_1$, $\mathbf{x}_2$ are in the same cluster, they are likely to be of the same class. | Part II |
| **Low density** | the decision boundary should lie in a low-density region. | |
| **Smoothness** | if two samples $\mathbf{x}_1, \mathbf{x}_2$ in a high-density region are close, so should be the corresponding outputs $y_1, y_2$. | Part III |
| **Manifold** | the (high-dimensional) data lie (roughly) on a low-dimensional manifold. | |

One may see equivalence between these four different assumptions. The *smoothness*

*assumption* which focuses on high-density regions is similar to the *manifold assumption*, which assumes smoothness along the geodesic distances (i.e. over manifold's high-density regions). Also the *low density assumption* is implying that the boundary is not cutting high-density regions, therefore not producing different labels for two samples in the high density region: i.e. the *cluster assumption*. These different assumptions are different views on SSL and have been the inspiration of different algorithms. Notice that SSL under wrong assumptions can degrade the performances [Cozman et al., 2002, Singh et al., 2009].

The various SSL methods can be separated into two main groups of methods: i) iteratively refining the model or ii) regularizing the model from the unlabeled data.

Among the first group, the generative models [Dempster et al., 1977] can iteratively predict and incorporate unlabeled data to refine the model, as well as the Transductive Support Vector Machine (SVM) [Joachims, 1999, Bruzzone et al., 2006] which gradually labels the unlabeled data to refine its classification boundary. The co-training approach [Blum and Mitchell, 1998] is another iterative SSL approach made for multi-view datasets. It builds classifiers on different feature sets of the data and gradually incorporate the unlabeled samples. It is important to notice that these approaches, while being simple extension of existing methods, are prone to local minima.

In the second group, we find the low-density separation methods for classification [Chapelle and Zien, 2004], the cluster kernels adapting to the data manifold [Chapelle et al., 2002, Tuia and Camps-Valls, 2009, 2011], the regularization by the unlabeled data of the optimization of the SVM [Bennett and Demiriz, 1999] or neural networks [Ratle et al., 2010], the graph-based regularization methods [Belkin et al., 2006] and the label propagation approaches. Two label propagation schemes were roughly proposed at the same time, the first by Zhu et al. [2003] with Gaussian Functions and Harmonic Fields (GFHF) and the second by Zhou et al. [2004] learning with Local and Global Consistency (LLGC). They both can be seen as *random walks* on the graph, where the label of a sample $\mathbf{x}_i$ is attributed to the closest labeled sample in terms of random walk distance [Chapelle et al., 2006]. These two approaches are investigated in Chapter 10.

## 3.4 Learning with kernels

In this Section, we introduce the kernel methods which form a unified framework extending the learning capacities of many standard algorithms to general types of data, such as strings, vectors, graphs, in a transparent way. We first present the key definitions and properties of kernels and their relation to reproducing kernel Hilbert spaces (RKHS). Then, we present the support vector machine (SVM) which exploit the *kernel trick* to produce a non-linear classification function.

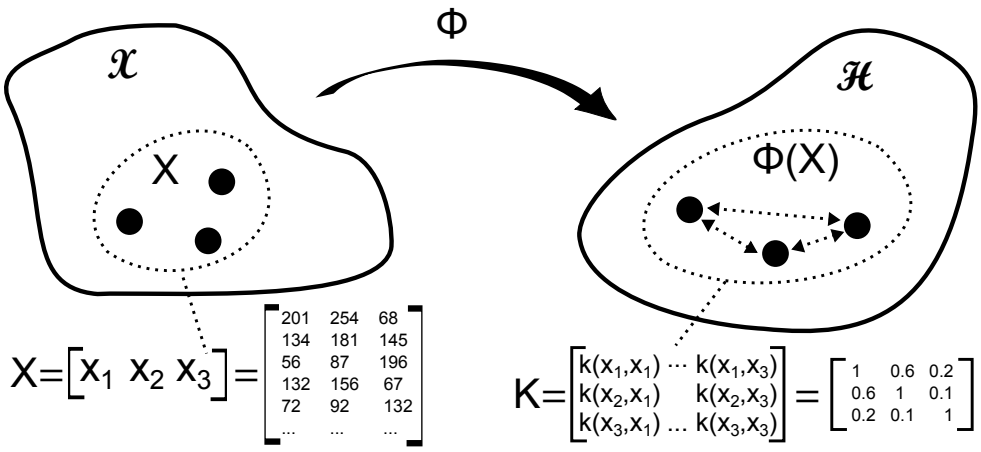



Figure 3.4: Representation of a dataset $X$ described by a number of features (e.g. spectral bands) in input space $\mathcal{X}$ and by the kernel matrix in the associated reproducing kernel Hilbert space that contains the pairwise similarity between the samples (modified from [Vert et al., 2004]).

### 3.4.1 The kernel trick

The common idea behind kernel methods is to represent the data only by pairwise comparisons instead of its individual features, the so-called *kernel trick* depicted in Figure 3.4. We will see in the following Sections that kernels are building blocks for creating powerful learning functions.

Let us consider the input dataset $\mathbf{x} \in \mathcal{X}$ and a non-linear function $\Phi$ that maps the data into the space $\mathcal{H}$.

$$\begin{aligned} \Phi: \ & \mathcal{X} \to \mathcal{H} \\ & \mathbf{x} \mapsto \Phi(\mathbf{x}) \end{aligned} \tag{3.15}$$

The space $\mathcal{H}$ is a *Hilbert space* since it has an associated dot product $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ which is a measure of similarity in this space. We can now define a kernel function as

$$\begin{aligned} K: \ & \mathcal{X} \times \mathcal{X} \to \mathbb{R} \\ & K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \quad \forall \mathbf{x}_i, \mathbf{x}_j \end{aligned} \tag{3.16}$$

As illustrated in Figure 3.4, the kernel matrix is composed of the kernel functions between pairs of samples. Each element of the kernel matrix are representing the dot product between the samples mapped in the Hilbert space. An interesting observation is that kernels are always resulting in a $N \times N$ kernel matrix in $\mathbb{R}$, whatever the type and dimensionality of the data. This will allow a learning machine based on kernels to be general for data of different sources and even combine them, as shown in Section 3.4.2.

The definition (3.16) establishes a kernel as a dot product in the feature space $\mathcal{H}$. Kernels fulfilling this condition are called *Mercer's kernels* [Mercer, 1909] or *positive definite symmetric* (PSD) kernels. A PSD kernel is defined [Mohri et al., 2012] as

**Definition 1.** *A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite symmetric (PSD) if for any $\mathbf{x} \subseteq \mathcal{X}$, the matrix $\mathbf{K}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^{N \times N}$ is symmetric positive semidefinite (SPSD).*

The matrix $\mathbf{K}$ is said SPSD if

$$\sum_{i,j}^{N} c_i c_j \mathbf{K}_{i,j} \geq 0, \quad c_i, c_j \in \mathbb{R} \tag{3.17}$$

From the kernel definition (3.16) and plugging it in the (3.17), we obtain

$$\sum_{i,j}^{N} c_i c_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}} = \langle \sum_{i}^{N} c_i \Phi(\mathbf{x}_i), \sum_{j}^{N} c_j \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}} = \| \sum_{i}^{N} c_i \Phi(\mathbf{x}_i) \|_{\mathcal{H}}^2 \geq 0, \quad c_i, c_j \in \mathbb{R} \tag{3.18}$$

which proves Definition (1). It is important to notice that any PSD kernel is associated with a unique space $\mathcal{H}$ and mapping function $\Phi$. This means that a kernel function is used to build a unique RKHS.

### 3.4.2 Mercer kernels

Let us introduce the kernel functions commonly used in applications. The choice of the kernel function will mostly depend on the type of data at hand. Among the kernels existing, we present here the main closed-form solution kernels $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$ for real-valued input vectors $\forall \mathbf{x} \in \mathbb{R}^d$ [Schölkopf and Smola, 2002].

$$\text{Linear kernel :} \quad K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.19}$$

$$\text{Polynomial kernels :} \quad K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + b)^p \tag{3.20}$$

$$\text{Sigmoid kernels :} \quad K(\mathbf{x}_i, \mathbf{x}_j) = \tanh a \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b \tag{3.21}$$

$$\text{RBF kernels :} \quad K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left( -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2} \right) \tag{3.22}$$

The linear kernel is a polynomial kernel of degree 1, which corresponds to an identity mapping $\Phi(\mathbf{x}) = \mathbf{x}$. It measures the collinearity of the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ and is zero for orthogonal vectors. The other polynomial kernels have a degree $d > 1$ and an offset parameter $b$ to tune. The sigmoid kernels have the parameters $a > 0$ and $b < 0$ and are typically used in feedforward neural networks.

The Radial Basis Function (RBF) kernel has a parameter $\sigma$, called *bandwidth*, to be tuned.

The similarity between two vectors is in relation with their Euclidean distance and quickly decreases towards zero for distances larger than the bandwidth. This kernel provides an intuitive notion of local similarity among data samples and has shown in the context of Support Vector Machines to generalize the linear kernel [Keerthi and Lin, 2003]. Notice that a full-rank matrix $\mathbf{K}$ spans a $N$-dimensional subspace of $\mathcal{H}$. This implies that the feature space is potentially of infinite dimensionality if there is no a priori restricting the number $N$ of training samples.

There are kernels dedicated to images such as the *pyramid match kernel* [Grauman and Darrell, 2007] which incorporates similarities from sub-windows at different sizes, or the *histogram kernels* based on any binary strings [Odone et al., 2005]. The *Fisher kernel* [Jaakkola et al., 1998] or the *Kullback-Leibler kernel* [Moreno et al., 2003] have been introduced for measuring similarities between distributions, the *principal angles kernel* to compare linear subspaces [Wolf and Shashua, 2003] and the *diffusion kernel* exploits the Laplacian of graph data to derive similarities between data samples [Hofmann et al., 2008]. Many other kernels have been developed and a survey can be found in [Schölkopf and Smola, 2002].

The choice of kernel functions is not restricted to predefined functions. Any combination of existing kernels that preserves the positive definiteness property is valid. The following *closure properties* are valid operations on kernels preserving the PSD [Hofmann et al., 2008]

**Definition 2.** *Let $K_1, K_2, \cdots$ be arbitrary positive definite kernels on $\mathcal{X} \times \mathcal{X}$ where $\mathcal{X}$ is a non-empty set. Then*

- *$\alpha_1 K_1 + \alpha_2 K_2$ is positive definite for $\alpha_1, \alpha_2 \geq 0$.*
- *$K$ is positive definite if $K(\mathbf{x}_i, \mathbf{x}_j) = \lim_{N \to \infty} K_N(\mathbf{x}_i, \mathbf{x}_j)$ exists for all $\mathbf{x}_i, \mathbf{x}_j$.*
- *The pointwise product $K_1 K_2$ is positive definite.*
- *The tensor product $K_1 \otimes K_2$ and the direct sum $K_1 \oplus K_2$ are positive definite kernels on $(\mathcal{X}_\infty \times \mathcal{X}_\in) \times (\mathcal{X}_\infty \times \mathcal{X}_\in)$*
- *The function $K(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i) f(\mathbf{x}_j)$ is a positive definite kernel for any function $f: \quad \mathcal{X} \to \mathbb{R}$.*

We see that convex combinations of kernel result in valid kernels, since the spectrum of their Gram matrix remains positive [Shawe-Taylor and Cristianini, 2004]. Some other operations deriving from the above mentioned elementary operations have been used in the context of *multiple kernel learning* [Bach et al., 2004] or in SSL where the kernel can be distorted using a positive definite matrix, i.e. the graph Laplacian [Belkin et al., 2006]. A survey on kernel construction is presented in [Belanche, 2013].

### 3.4.3 Reproducing kernel Hilbert spaces (RKHS)

After defining them, we study now how the kernels are linked to the learning functions. The kernels can be used to build a Hilbert space $\mathcal{F}$ containing a set of learning functions $f$. These functions are restricted to be expressed as linear combinations of the kernels, forming the *reproducing kernel Hilbert space* (RKHS):

$$\mathcal{F} = \{f(\cdot) = \sum_{i=1}^{N} \alpha_i K(\cdot, \mathbf{x}_i), \quad \alpha_i \in \mathbb{R}, \quad \mathbf{x}_i \in \mathcal{X}\}, \quad i \in \mathbb{N} \tag{3.23}$$

Notice that the space $\mathcal{F}$ is a vector space, i.e. the addition and the multiplication by a scalar are defined in that space. Let us now define the dot product between two functions $f$ and $g$ in this space as

$$\langle f, g \rangle_{\mathcal{F}} = \sum_{i=1}^{N} \sum_{j=1}^{N'} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{3.24}$$

where $g(\cdot) = \sum_{j=1}^{N'} \beta_j K(\cdot, \mathbf{x}_j), \quad \beta_j \in \mathbb{R}, \quad \mathbf{x}_j \in \mathcal{X}, \quad N' \in \mathbb{N}$.

The kernel *reproducing property* states that any function $f \in \mathcal{F}$ can be defined as a linear function of inner products in the vector space $\mathcal{F}$. If we set $g(\cdot) = K(\cdot, \mathbf{x})$, the dot product with $f$ is

$$\langle f, K(\cdot, \mathbf{x}_j) \rangle_{\mathcal{F}} = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_j) \tag{3.25}$$

and

$$\langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_j) \rangle_{\mathcal{F}} = K(\mathbf{x}_i, \mathbf{x}_j) \tag{3.26}$$

from the property (3.25), the space $\mathcal{F}$ is called a *reproducing kernel Hilbert space* (RKHS). From [Aronszajn, 1950], every positive definite kernel $K$ is associated to a unique RKHS whose kernel is K and vice versa. The corresponding mapping function $\Phi$ is called the *Aronszajn map*.

Notice that the RKHS norm of a function $\|f\|_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}} =$ is a measure of smoothness and complexity. This relates to the regularization form of (3.10), where the regularizator term is $\Psi = \|f\|_{\mathcal{F}}^2$. Since $f = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$, the regularization term can expressed as kernel evaluations: $\Psi = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$. In this situation, $\Psi$ can be expressed in a vector notation and the regularized risk only in terms of kernel evaluations

$$R_{reg}(K, \alpha) = R_{emp}(K, \alpha) + \lambda \alpha^T \mathbf{K} \alpha \tag{3.27}$$

we directly see the relation existing between the RKHS and the regularization of learning

functions. This brings us to the *Representer theorem* [Kimeldorf and Wahba, 1971, Hofmann et al., 2008] which states that the solution of an optimization in the infinite-dimensional space $\mathcal{F}$ lies in the span of the kernels centered at the training samples. Formally, we have

**Definition 3.** ***Representer Theorem.*** *Denote by $\Psi : [0, \infty) \to \mathbb{R}$ a strictly monotonic increasing function, by $\mathcal{X}$ a set, and by $\mathcal{L} : \mathcal{X} \to \mathbb{R} \cup +\infty$ an arbitrary loss function. Then each minimizer $f \in \mathcal{F}$ of the regularized risk functional*

$$\mathcal{L}(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)) + \Psi(\|f\|_{\mathcal{F}}^2) \tag{3.28}$$

*admits a solution of the form $f^*(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x})$.*

Now that the functions in $\mathcal{F}$ have been defined as linear combinations of kernel evaluations, we will see that some operations in the space $\mathcal{F}$ can be computed in the same way.

The first operation is the *translation* in the feature space: $\tilde{\Phi}(\mathbf{x}) = \Phi(\mathbf{x}) + \Gamma, \quad \Gamma \in \mathcal{H}$. By restricting the translation $\Gamma$ to lie in the span of $\Phi(\mathbf{x})$, we can compute the translated dot product between samples $\mathbf{x}$ and a test sample $\mathbf{x}_t$ as

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_t) \rangle = K(\mathbf{x}, \mathbf{x}_t) + \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_t, \mathbf{x}_i) + \sum_{i,j=1}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{3.29}$$

The translation by the mean of the data $\mu_\Phi = \frac{1}{N} \sum_{i=1}^{N} \Phi(\mathbf{x}_i) = -\Gamma$, corresponds to a *centering in the feature space* of the data.

The *normalization in the feature space* of the data results in the following dot product

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \frac{\Phi(\mathbf{x}_i)}{\|\Phi(\mathbf{x}_i)\|}, \frac{\Phi(\mathbf{x}_j)}{\|\Phi(\mathbf{x}_j)\|} \rangle \tag{3.30}$$

which is the cosine of the angle in the feature space.

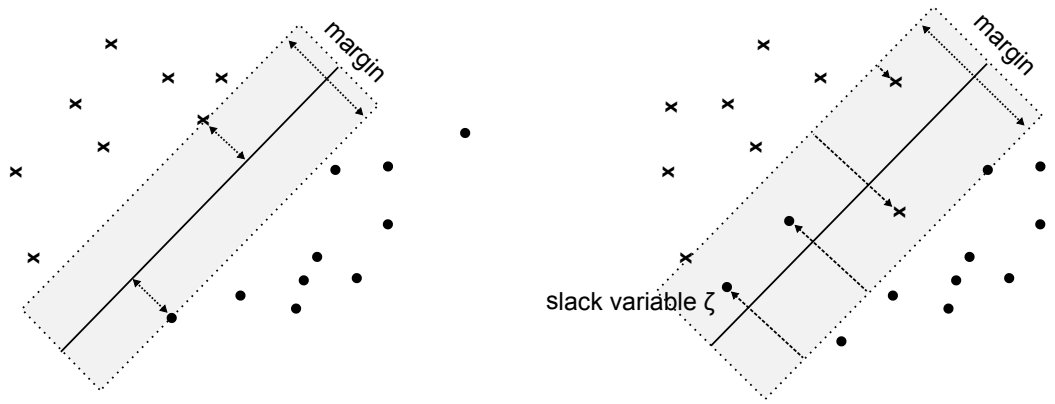
Finally, the computation of *distances in the feature space* is directly related to the notion of similarity in the RKHS and can be expressed in terms of kernel evaluations as

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|_{\mathcal{H}} = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_i, \mathbf{x}_j) + K(\mathbf{x}_j, \mathbf{x}_j)} \tag{3.31}$$

### 3.4.4 Support Vector Machines (SVM)

We can now introduce the *Support Vector Machine* (SVM), which is one of the most successful kernel methods.

The SVM is a very efficient machine learning tool for the problem of classification. It




(a) (b)

Figure 3.5: Schematic of the SVM decision boundary. (a) The separable case with the maximum margin separation. (b) The non-separable (overlap) case. Some of the samples are on the other side of the decision boundary. These samples on the wrong side are penalized by their distance $\xi$ to their margin. The boundary tries to maximize the margin while minimizing the samples on the wrong side.

has been used in a wide range of applications with great success [Osuna et al., 1997, Schölkopf and Smola, 2002, Camps-Valls and Bruzzone, 2009]. The SVM is directly connected to the ERM principle. Vapnik's original idea was to design an algorithm aiming at finding the minimal empirical risk for a set of training data and a fixed class of learning functions $f$ [Vapnik, 2000]. Let us assume a set of training samples with associated labels $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$. The SVM defines the hyperplane separating the two classes as

$$\langle w, \mathbf{x} \rangle + b = 0, \quad w \in \mathbb{R}^d, b \in \mathbb{R} \tag{3.32}$$

The margin between the two classes is $\frac{2}{\|w\|}$ which gives a measure of the separation between the two classes, as represented in Figure 3.5 (a). Among all the hyperplanes separating the two classes, there is only one unique hyperplane which separate them with maximum margin. With this optimal hyperplane, the minimal empirical risk will be achieved.

The SVM solves an optimization problem which finds the hyperplane separating two classes with a maximum margin. Maximizing the separation is equivalent to minimizing $\frac{\|w\|}{2}$. This way we avoid any trouble for $w \sim 0$. Furthermore, the minimization of the convex form $\frac{\|w\|^2}{2}$ is preferred in terms of optimization.

In the case of two classes which are not perfectly separable, we can allow for some errors of classification. This can be done by introducing the slack variables $\xi_i$ that represent the distance from the margin to the misclassified samples, as represented in Figure 3.5 (b). The optimization will then search for a trade-off between maximizing the margin and

minimizing the misclassification distances.

**Primal SVM** We are now able to formulate the SVM objective function to be optimized. The so-called *primal* problem is

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ &\text{s.t.} \ \ y_i(\langle w, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{3.33}$$

with $C$ weighting the penalty term. This quadratic optimization problem under constraints is solved by introducing Lagrangian multipliers

$$L_p(w, b, \xi, \alpha, \gamma) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i[y_i(\langle w, \mathbf{x}_i\rangle + b) - 1 + \xi_i] - \sum_i \gamma_i \xi_i \tag{3.34}$$

The optimal solution of the objective function is obtained at the minima of the Lagrangian. The derivative with respect to $w$, $b$ and $\xi$ are

$$\begin{aligned} \tfrac{\partial L}{\partial w}: &\quad w - \textstyle\sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \rightarrow \qquad w = \sum_i \alpha_i \\ \tfrac{\partial L}{\partial b}: &\qquad \sum_i \alpha_i y_i = 0 \\ \tfrac{\partial L}{\partial \xi_i}: &\quad C - \alpha_i - \gamma_i = 0 \quad \rightarrow \quad \alpha_i = C - \gamma_i, \quad \forall i \end{aligned} \tag{3.35}$$

The Karush-Kuhn-Tucker (KKT) conditions are added since the constraint of (3.33) are inequalities and allows to uniquely relate the following dual problem with the primal problem.

**Dual SVM** By substituting partial derivatives (3.35) into (3.34), we obtain the *dual* problem

$$\begin{aligned} &\max_{\alpha} \ \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \\ &\text{s.t.} \ \ 0 \leq \alpha_i \leq C; \quad i = 1, ..., N, \quad \sum_i \alpha_i y_i = 0 \end{aligned} \tag{3.36}$$

that ensures an optimal solution in the form of a linear combination of the training samples. This form of solutions is similar to the statement of the Representer's theorem (Definition 3) for kernels.

**Non-linear Support Vector Machines** The standard SVM is solving the classification problem by finding an optimal plane separating the two classes. However in many

situations, the optimal separation is not a linear hyperplane. In such situations, the data can be mapped into a higher dimension space where it could become linearly separable. This leads us back to the notion of kernels and RKHS. When replacing the training set $\mathbf{x}_i$ by its mapped version $\Phi(\mathbf{x}_i)$ in the SVM formulation, we end up with only dot products of the mapped data. The mapping function dot products thus represent kernel functions $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$.

Replacing the dot products with kernels back into the SVM dual formulation of (3.36), we get the following non-linear dual formulation

$$\begin{aligned} \max_{\alpha} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, ..., N, \quad \sum_i \alpha_i y_i = 0 \end{aligned} \tag{3.37}$$

The dual optimization problem is a standard quadratic programming (QP) problem. Many algorithms can solve QP problems efficiently by exploiting decomposition methods (e.g. Sequential Minimal Optimization [Platt, 1998]).

The class label of a test sample $\mathbf{x}_t$ is obtained from the sign of the decision function

$$f(\mathbf{x}_t) = \sum_i \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_t) + b \tag{3.38}$$

where $\alpha_i^*$ are the support vector coefficients solutions of (3.37). This decision function $f(\mathbf{x})$ can be interpreted as a distance to the boundary in the kernel induced space, becoming null for samples lying on it.

### 3.4.5 Cost-Sensitive SVM (CS-SVM)

In the standard SVM, the cost parameter $C$ penalizes equally the errors done on the two classes. This is usually desirable, but in some specific situations, like unbalanced problems, the two classes should be penalized with different strength. The Cost-Sensitive SVM (CS-SVM) has two costs $C_+$ and $C_-$, one for each class, combined into the cost asymmetry parameter $\gamma = \frac{C_+}{C_+ + C_-}$ and the total amount of regularization $\lambda = \frac{1}{C_+ + C_-}$ [Bach et al., 2006].

Assuming no bias term $b$, the CS-SVM primal optimization problem is

$$\begin{aligned} \min_{w,\xi} \quad & \frac{\lambda}{2} \|w\|^2 + \gamma \sum_{i \in I_+} \xi_i + (1 - \gamma) \sum_{i \in I_-} \xi_i \\ \text{s.t.} \quad & y_i \langle w, \Phi(\mathbf{x}_i) \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \end{aligned} \tag{3.39}$$

The cost asymmetry $\gamma$ controls the trade-off between false positive and false negative

rates. The entire set of classifiers along the Receiver Operator Curve (ROC) are obtained for $\gamma$ ranging from 0 to 1. For $\gamma = 0.5$, the algorithm reduces to the standard SVM (Cost-Insensitive SVM).

The optimization problem (3.39) is solved through its Lagrangian dual formulation, given by

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2\lambda}\sum_{i,j}\alpha_{i,\gamma}\alpha_{j,\gamma}y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) - \sum_i \alpha_{i,\gamma} \\ \text{s.t.} \quad & 0 \leq \alpha_{i,\gamma} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma,\ \forall i \end{aligned} \tag{3.40}$$

where $K(\mathbf{x}_i,\mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ is the kernel representing the dot product in the high-dimensional space induced by $\Phi$. The indicator function $\mathbf{1}_{\{y_i<0\}}$ returns 1 for $y_i = -1$ and 0 otherwise. It is important to notice that the solution of (3.40) is obtained for a single and fixed cost asymmetry $\gamma$, and thus it becomes necessary to solve an additional optimization problem for each new cost asymmetry considered.

Similarly to the standard SVM, the class label of a test sample $\mathbf{x}_t$ is obtained from the sign of the decision function

$$f_{\gamma,\lambda}(\mathbf{x}_t) = \frac{1}{\lambda}\sum_i \alpha^*_{i,\gamma} y_i K(\mathbf{x}_i,\mathbf{x}_t) \tag{3.41}$$

where $\alpha^*_{i,\gamma}$ are the support vector coefficients solutions of (3.40).

The CS-SVM has the drawback of having an additional parameter $\gamma$ compared to the original SVM. Efficient ways of computing the solutions for a range of $\gamma$ and $\lambda$ values are presented in Chapter 5.

## 3.5 Summary

In this Chapter, we have introduced the statistical learning framework allowing a machine to learn from examples. Its extension with the use of kernel methods allows to consider new problems where the data is potentially non-linear and of any type, as long as pairwise similarities can be computed. The kernels methods have been considered in the remote sensing community with great interest [Camps-Valls and Bruzzone, 2009], since they allow to address efficiently many of the issues inherent to the complex data characteristics in the context of remote sensing image processing [Camps-Valls et al., 2011].

The main notions and methods for learning from data having been presented, we can focus on the particular learning situation of novelty detection in Chapter 4.

# 4 Novelty detection in machine learning

The definition of novelty detection is not unique and the appropriate methods will drastically change depending on the type of data, the definition of the *normality* and the context of the problem. Novelty detection aims at detecting unknown, anomalous or different data samples. This links novelty detection to one-class classification, which is focused on the *normal* class instead of the novelties. Novelty detection is also closely related to anomaly, outlier and target detection, as well as the classification problems in situations of extremely imbalanced classes. The distinctions, even if not strict, are described in Section 4.1.

## 4.1 The novelty, the outlier and the anomaly

Here we review the similarities and differences between novelty, anomaly, outlier and target detection, which are schematized in Figure 4.1. We also relate novelty detection to clustering and learning from positive and unlabeled (LPU) examples settings. These different topics are misleading when we start comparing them. At first sight they look equivalent since all look for observations which deviates significantly from the normal observations. In the literature, these different terms are employed mostly depending on their related applications and without a narrow and precise definition. In the end, these problems can be recasted as different instances of novelty detection or one-class classification, as illustrated in Figure 4.1.

**Anomaly detection** The detection of anomalies [Chandola et al., 2009] is defined in a similar way to novelty detection. It is mainly distinguished by the degree of rareness of the novelty. While novelty detection does not say if the novelties are rare, anomaly detection usually makes this assumption. Therefore, anomalies can be detected as insignificant after learning a mixture model [Eskin, 2000], or as badly reconstructed after performing robust PCA [Kwitt and Hofmann, 2007]. Semi-supervised anomaly detection

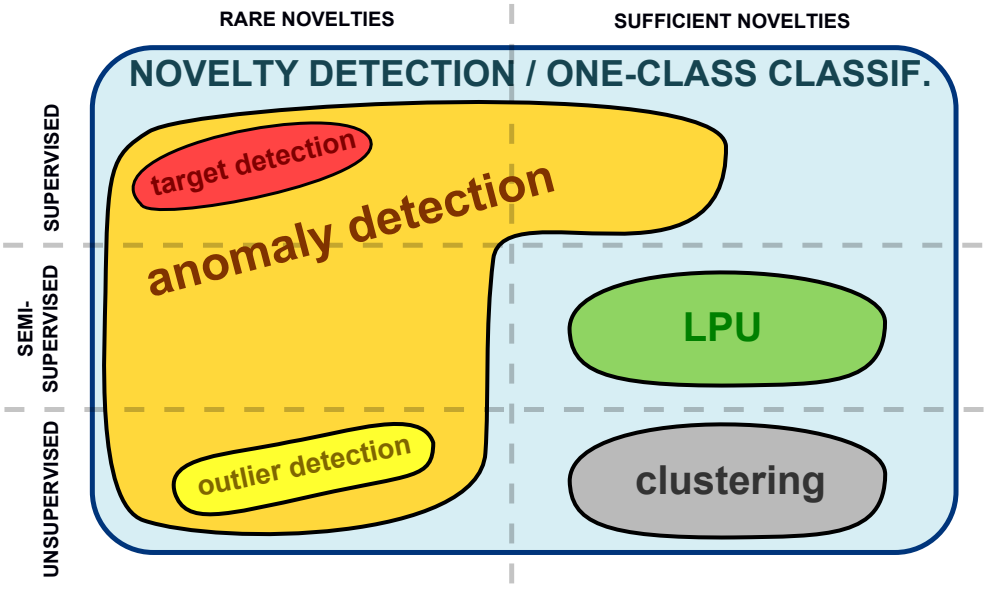



Figure 4.1: Taxonomy of the different fields related to novelty detection depending on the frequency of novelties (rare or not) and on the level of supervision (supervised, semi-supervised, unsupervised).

based on incorporating labeled and unlabeled samples in a convex formulation [Goernitz et al., 2013] and active learning for anomalies [Pelleg and Moore, 2004] are equivalent to the semi-supervised novelty detection problem presented in Section 4.3.3. Anomalous change detection [Theiler et al., 2010] is referring to anomaly and novelty detection in remote sensing change detection problems where the changes are assumed unknown and anomalous. Finally, with the assumption that anomalies are scarce, anomaly detection is directly connected to the field of outlier detection in statistics. As an example, the data can be mapped into a feature space where outlier detection is performed to retrieve anomlies in the data [Eskin et al., 2002].

**Outlier detection** The analysis of outliers is an important topic which has received a lot of interest in statistics, and continues to attract new researchers [Barnett and Lewis, 1994]. If we recall the definition made by Hawkins in 1980:

> *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*

The notion of deviation can be seen as a definition of anomaly or novelty. In outlier detection, all the data is available and comparisons between data samples are performed in order to decide if a sample is an outlier or not. However, the difference lies in that no supervision is available. The outliers are lying among the data as noise present in a

training dataset. Therefore outlier detection can be seen as the particular case of novelty and anomaly detection which is unsupervised with rare novelties.

**Target detection** The field of target detection is falling right into the supervised part of novelty and anomaly detection. In target detection, it is generally assumed that training samples or information on the target is available and that targets are sporadic spatially. The main approaches are based on the RX algorithm (see Section 4.4.2) and other matched-filters [Nasrabadi, 2014], which can be seen as supervised one-class classifiers integrating spatial information.

Finally, novelty detection with novelties that are not rare can be seen as an instance of **clustering** (see Section 4.3.2) or of **learning from positive and unlabeled (LPU)** examples (see Section 4.3.3). Novelty detection as a clustering problem is investigated in Chapters 6 and 9 and as a LPU problem in Chapters 7 and 10. In these *transductive* situations, unlabeled data containing both normal samples and novelties is available. Before establishing the general formulation of a novelty detection problem in Section 4.3, we review the main applications in which novelty detection is used in Section 4.2.

## 4.2 Novelty detection in applications

We here review the principal fields of interest and applications making use of novelty detection.

Novelty detection is most often related to the monitoring of a certain behaviour like critical mechanical systems where unusual events are the potential sign of a default or damage. It has been used for the automatic detection of jet engine default [Tarassenko et al., 1999, Hayton et al., 2000, Sundaram et al., 2009], atypical machine vibrations [Miljković, 2008] and milling tools [Dasgupta and Forrest, 1996]. In telecommunications, it has been used for fraud detection in e-commerce, credit card transactions and mobile phones [Bolton and Hand, 2002], for the detection of intrusion in computer systems of suspicious trades in the equity markets [Ryan et al., 1998, Yeung and Chow, 2002], and spatio-temporal changes in time-evolving graphs [Chan et al., 2008].

The ability to detect unusual events and objects on the way of mobile robots and autonomous vehicles are of prime importance [Sofman et al., 2010, Munroe and Madden, 2005]. For example, such robots can be used for the detection of rope wire defects in unreachable places [Wacker and Denzler, 2013].

More recently, the analysis of ego-centric videos could benefit from novelty detection for discovering important objects or persons [Aghazadeh et al., 2011, Lee et al., 2012], as well as new situations with respect to routine activities [Blanke and Schiele, 2009]. The idea relies on using on-body wearable devices recording the daily activities of a

subject, the so-called life logging [Doherty and Smeaton, 2010]. Memory assistance via a wearable camera could have a great potential in helping patient with the Alzheimer's disease [Hodges et al., 2006]. Daily recordings would provide normal data and events could be evaluated as novel or not, replacing's brain capacity of novelty detection [Borisyuk and Kazanovich, 2004].

In biomedical, novelty detection has been used for the detection of masses in mammograms [Tarassenko et al., 1995], of protein interactions [Deris et al., 2006] and of epileptic seizures in electroencephalograms [Roberts and Tarassenko, 1994, Gardner et al., 2006]. In biometrics, it allows the recognition of a known faces and the detection of unauthorized persons [Seo and Ko, 2004, Wang et al., 2004, Fernandes et al., 2011].

It finds also applications in audio with the detection of abrupt audio changes [Davy and Godsill, 2002], in document classification [Manevitz and Yousef, 2002], and in forensic data to determine whether a similar case exists and if further chemical investigations are needed [Ratle et al., 2007].

Finally, novelty detection has been investigated in remote sensing for anomaly detection [Banerjee et al., 2009], target detection [Nasrabadi, 2014], one-class classification [Muñoz-Marí et al., 2007, Theiler et al., 2010] and multi-temporal change detection [Mercier and Girard-Ardhuin, 2006, Bovolo et al., 2010, Volpi et al., 2012].

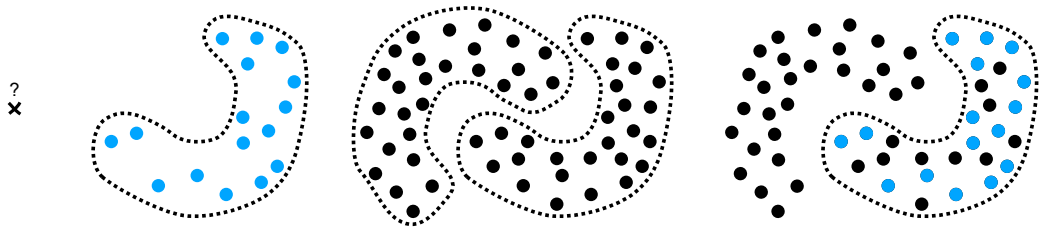
## 4.3 General formulation

Here, we introduce a general formulation for the novelty detection problem, which boils down to anomaly, outlier and one-class classification under appropriate assumptions. Let us denote the distribution of the *normal*, also called *nominal*, *unchanged* or even *target*, data $p_0$ and the distribution of novelties $p_1$. Considering a set of $N$ input data vectors, $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^d \subseteq \mathcal{X}$, we can assume that the data follows the contamination model

$$\mathbf{x}_i \sim p = \pi p_1 + (1 - \pi) p_0$$

with $\pi$ the unknown fraction of novelties.

The goal of novelty detection is to learn a decision function $f(\mathbf{x})$, which equals to $-1$ for a novelty and 1 for a normal sample. Therefore, the positive values of $f$ will refer to the description of the domain of the *normal* class. The knowledge of both distributions $p_0, p_1$ would lead to the optimal Bayes decision rule as in supervised classification (c.f. equation (3.1)). However, this is not realistic in practice since we have access only to a limited set of data samples not representative of the total distribution $p$. If novelties are present in the training data, they are either too few to correctly estimate $p_1$ or not labeled.




(a) (b) (c)

Figure 4.2: The three novelty detection paradigms considered in this Thesis: (a) supervised novelty detection where only labels from one-class are available, (b) unsupervised novelty detection where no labels are available, (c) semi-supervised novelty detection where both labeled and unlabeled samples are available.

This ill-posed situation requires to make certain assumptions on the distributions. Often it is assumed that novelties are uniformly distributed. This leads to a density estimation problem where $f$ is defined as a level set of the probability distribution $p_0$ [Scott and Nowak, 2006]:

$$f(\mathbf{x}, p_0) = \begin{cases} +1, & \text{if } \; p_0(\mathbf{x}) > \theta \\ -1, & \text{if } \; p_0(\mathbf{x}) \leq \theta \end{cases} \tag{4.1}$$

where the decision function $f(\mathbf{x})$ is positive on density regions $p(\mathbf{x})$ higher than $\theta$. The level set $\theta$ directly relates to the fraction $\pi$ of the uniform distribution of novelties. We will see in Section 4.4 that a large number of methods aims at finding the level set function $f$ directly without passing through the estimation of the distribution $p_0(\mathbf{x})$. This follows Occam's razor principle, since the real distribution of the data may not be of interest in detection problems.

In anomaly detection, $\pi$ will be generally assumed low, since novelties are rare. The same assumption holds for outlier detection.

In one-class classification, or supervised novelty detection, the training set $\mathcal{T}$ is assumed to be drawn from a distribution $p$ which is novelty-free ($\pi \approx 0$ and $p \sim p_0$), while the test samples are derived from a distribution $p'$ where $\pi' \in [0, 1]$.

The main paradigm of novelty detection is that information is available on the *normal* distribution that may be contaminated, but not on the novelty distribution. However, depending on the availability of labeled and unlabeled samples at training, different novelty detection paradigms can be identified. The three main situations, namely *supervised novelty detection*, *unsupervised novelty detection* and *semi-supervised novelty detection* are presented in the three following Sections and illustrated in Figure 4.2.

### 4.3.1 Supervised novelty detection

The supervised novelty detection is the standard setting. Only samples from the *normal* distribution are available without any other information. In this supervised setting, all the training data is assumed drawn from $p_0(\mathbf{x})$ where $\pi \sim 0$. This can be summarized as finding the support or level set of the data probability density function. If few labeled novelties are also available, they can be used to better restrict the support of the *normal* distribution. This relates to unbalanced classification problems where we have very few labeled samples for one of the two classes. In the case of wrongly labeled examples or noisy samples present in the training data, the distribution will be robustly estimated at a certain level set and assuming $\pi > 0$.

### 4.3.2 Unsupervised novelty detection

Unsupervised novelty detection corresponds to the situation where a set of unlabeled observations, composed of an unknown combination of normal data and novelties, is available. This situation is also called the *missing label problem*. The two distributions $p_0$ and $p_1$ are estimated by either assuming the normal samples to be "contaminated" by novelties without knowing the contamination level $\pi$ [Scott and Kolaczyk, 2007] (*outlier detection* problem), or by finding large structures in the data (*clustering* problem). Prior knowledge on $p_0$ and $p_1$ is definitely crucial to define the type of algorithms needed. In this Thesis we assume that novelties are not rare and can therefore be retrieved as clusters in the data distribution.

### 4.3.3 Semi-supervised novelty detection

Semi-supervised novelty detection (SSND) corresponds to the situation where both unlabeled and labeled *normal* samples are available. In comparison to the standard novelty detection setting, which is inductive, SSND is often transductive and can highly benefit from the additional unlabeled samples in this challenging context [Scott and Blanchard, 2009].

The semi-supervised approaches to novelty detection are based on the same assumption than standard SSL and range from simple extension of one-class classifiers to approaches very specific to this particular setting, as in the case of LPU.

The Semi-Supervised One-Class SVM ($S^2$OCSVM) [Muñoz-Marí et al., 2010] is an extension of the One-Class SVM which distorts the kernel via the graph Laplacian to adapt to the data manifold geometry. The method is further presented in Section 4.4.3 and used in Chapters 7 and 10.

The LPU [Liu et al., 2003, Li and Liu, 2003] is equivalent to SSND. First studied

theoretically by Muggleton [1997] who assumed known probability density functions of positive and unlabeled data, LPU moved one step ahead in viewing it as a two-class classification problem [Liu et al., 2002]. Methods dedicated to the LPU setting have been designed such as C4.5 trees [De Comité et al., 1999, Letouzey et al., 2000] where the method learns the fraction of novelties $\pi$, weighted logistic regression [Lee and Liu, 2003], neural networks [Skabar, 2003], co-training [Denis et al., 2003], labeling heuristic [Fung et al., 2006], a SVM iteratively discovering novelties [Yu, 2005] and finally a cost-sensitive SVM, also called biased SVM, penalizing less the errors on unlabeled samples [Steinwart et al., 2004, Elkan and Noto, 2008, Li et al., 2011b]. The reduction of the LPU (or SSND), problem to a standard two-class classification problem has proved to lead to the optimal classifier without any assumption made on the distribution of the novelties [Blanchard et al., 2010]. This approach is further investigated in Chapter 7.

Other SSND approaches have been developed for certain particular applications: the detection of novelties in time series using a semi-supervised kernel principal component analysis [Thompson et al., 2012], the detection of anomalies solved by incorporating labeled and unlabeled samples in a convex formulation [Goernitz et al., 2013], semi-supervised target detection relying on the kernel distorted by the graph Laplacian [Capobianco et al., 2009], as in the $S^2$OCSVM, novelty detection in multi-modal or multi-view data with the semi-supervised CCA [Blaschko et al., 2008, Kimura et al., 2010] or as a *multiple instance learning* (MIL) [Maron and Lozano-Pérez, 1998] where the labels are available only on bags of samples, called instances, and are set positive if at least one sample is positive in the bag [Maron and Lozano-Pérez, 1998, Bolton and Gader, 2010].

Finally, the SSND can also be formulated as the estimation of the ratio [Sugiyama et al., 2012] or difference [Sugiyama et al., 2013] of the labeled and unlabeled probability density functions: $p_0$ and $p$, respectively.

## 4.4 State-of-the-art in novelty detection

A large number of methods have been developed for the particular problem of novelty detection and one-class classification. Here are reviewed the principal approaches with a particular focus on kernel methods and the approaches used in remote sensing.

### 4.4.1 Density estimation approaches

**Gaussian Data Description (GDD)** The simplest Bayesian model for a class of interest is a Gaussian probability distribution in $\mathbb{R}^d$ expressed as

$$p_{\mathcal{N}}(\mathbf{x}, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\} \tag{4.2}$$

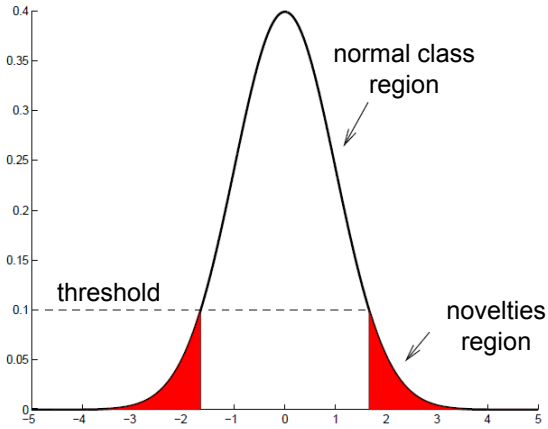



Figure 4.3: One dimensional Gaussian distribution and a threshold defining the class boundaries and the region of novelties (modified from Tax [2001]).

where the mean $\mu$ and covariance $\Sigma$ can be directly inferred from the training data. Notice that the inversion of the covariance matrix may be problematic and one often resorts to the pseudo-inverse or impose a diagonal covariance matrix. The novelties are detected in the regions of low density, as illustrated in Figure 4.3. However, this model is often too simplistic and cannot fit the data distribution correctly, as illustrated in Figure 4.4 (a).

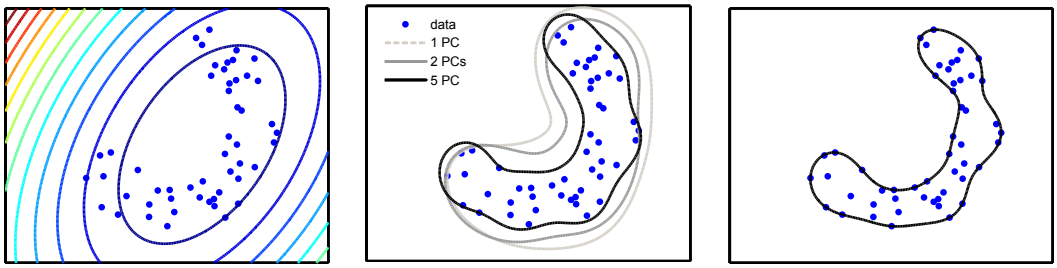
**Mixture of Gaussian Data Description (MoGDD)** Normal class distribution that may have multiple modes or that does not fulfil the Gaussian assumption can be modeled by a mixture of Gaussian models

$$p_{MoG}(\mathbf{x}, \mu, \Sigma) = \frac{1}{k_{MoG}} \sum_j \alpha_j p_{\mathcal{N}}(\mathbf{x}, \mu_j, \Sigma_j) \tag{4.3}$$

where $\alpha_j$ are the mixing coefficients. The mixture of Gaussian has an additional parameter $k_{MoG}$ compared to the single Gaussian, which is the number of Gaussian components in the mixture. This more complex model requires a larger number of training data samples to be reliably estimated [Markou and Singh, 2003a, Muñoz-Marí et al., 2007].

These parametric models have the drawback of being often too restrictive in their assumptions regarding the data complexity. Moreover, they require an important number of training samples in order to accurately estimate the parameters of the model, without even considering the problems arising with the curse of dimensionality, described earlier in Section 3.2.

The assumption of data Gaussianity in the input space is most of the time erroneous leading to sub-optimal detection of novelties. To alleviate this assumption, Laparra et al. [2011] proposed to look for a succession of rotations and whitenings of the data which transforms it into a Gaussian distribution. Therein, novelty detection can easily be




(a) (b) (c)

Figure 4.4: Novelty detection/One-class classifiers: (a) Gaussian model, (b) kernel PCA with various number of principal components used and (c) SVDD with the same kernel bandwidth.

performed with the method above, illustrated in Figure 4.3. The idea of simplifying the distribution prior to detection has been exploited via a one hidden-layer perceptron which transforms the training data in a Gaussian distribution. Novelties are easily detected in the low-density regions of the Gaussian [Parra et al., 1996].

Non-parametric approaches have been proposed to approximate the density of the data using the distance to the $k^{th}$ neighbour on a $k$-NN graph [Tax, 2001, Hautamaki et al., 2004, Harmeling et al., 2006]. A test sample is declared novel if its $k^{th}$ distance exceeds a certain threshold. This procedure is generalized in Chen et al. [2008], where a kernelized spatial depth is introduced to measure the novelty of samples with respect to a given dataset. Alternatives based on a local estimation of the density are the local outlying factor (LOF) [Breunig et al., 2000], the local correlation integral (LOCI) [Papadimitriou et al., 2003], genetic algorithms for high dimensional data [Aggarwal and Yu, 2001] or the local outlier probability (LoOP) [Kriegel et al., 2009]. The Prototype-based Domain Description [Angiulli, 2012] introduces a statistical test for outlier detection, proposes a subset selection method to encompass large datasets and generalizes the $k$-NN approach. Finally, decision trees like the Minimum Spanning Tree (MST) also describe local relationships among the data samples and novelties can be detected based on the closest edge distance [Juszczak et al., 2009].

The human immune system has inspired methods which treats novelties in the same manner our body protects from bacteria and viruses, in retaining unique patterns from the data and comparing the test data with this collection [Dasgupta and Forrest, 1996].

Also it is worth mentioning the branch of statistic devoted to model extreme and rare events relying on the Extreme Value Theory (EVT). It allows to study the outlying region of a known distribution and therefore better detect novelties [Roberts, 1999].

### 4.4.2 Reconstruction-based approaches

Among the simplest models based on the reconstruction of the data is the $k$-centers [Ypma and Duin, 1998]. The maximum distance of all minimum distances between training objects and the centers is used as a measure of novelty. This produces a description of the data made of hyperspheres centered on the $k$ centers. Hereafter are presented other reconstruction methods all trying to recover the subspace on which lies the normal data.

**PCA and Orthogonal Subspace Projections** The principal component analysis (PCA) has been extensively used for the analysis of high-dimensional data. It has been exploited in applications such as data compression, data visualization or network intrusion [Wang and Qin, 2002]. More recently, it has been used for change detection in remote sensing [Wu et al., 2013]. The data is assumed to be composed of the normal class $\mathbf{X}_{normal}$ lying onto the principal dimensions related to high variance and the novelties $\mathbf{X}_{novelty}$ are found in the other dimensions.

$$\mathbf{X} = \mathbf{X}_{normal} + \mathbf{X}_{novelty} \tag{4.4}$$

The data projections onto the target and novelty subspace respectively are

$$\hat{\mathbf{X}}_{normal} = PP^T\mathbf{X} \tag{4.5}$$

$$\hat{\mathbf{X}}_{novelty} = (\mathbf{I} - PP^T)\mathbf{X} \tag{4.6}$$

where $P = U_1$ the $d \times k$ matrix containing the $k$ principal component vectors obtained from a set of labeled normal samples $\mathbf{X}_l \subseteq \mathbf{X}_{normal} \in \mathbb{R}^{d \times N}$. It is computed as $U\Sigma U = \text{eig}(\text{cov}(\mathbf{X}_l))$, where $U_1 : \sigma_i,\ \forall i \leq k$ and $U_2 : \sigma_i,\ \forall i > k$. The novelty detection of a new unseen sample $\mathbf{x}_t$ is based on the norm of the projection onto the novelty subspace $d_{novelty}(\mathbf{x}_t) = \|\hat{\mathbf{x}}_{t,novelty}\| = \|(\mathbf{I} - PP^T)\mathbf{x}_t\|$ where a novelty is detected if it has a larger projection norm than the set of samples $\mathbf{X}_l$: $\mathbf{y}_i = d_{novelty}(\mathbf{x}_i) > \max(d_{novelty}(\mathbf{X}_l))$.

If the training samples $\mathbf{X}_l$ cannot be assumed free of anomaly/outliers, the subspace projector $PP^T$ has to be estimated either by retaining only the largest eigenvalues to form $P$ or by considering robust estimators [Kwitt and Hofmann, 2007].

**The RX algorithm** The RX algorithm [Reed and Yu, 1990] is a popular anomaly detection algorithm that was first developed for multi-spectral imagery, then transferred to hyperspectral and non-linear data distributions. The RX algorithm is closely related to the previous idea of orthogonal subspace projections, while being specifically dedicated to image data and requiring no training data, at least in its original formulation [Chang, 2002]. The RX algorithm computes a Mahalanobis distance to measure the degree of anomaly of a pixel with respect to its background. The background is modeled over a window covering $M$ pixels $\mathbf{X}_b$ around the pixel of interest $\mathbf{x} \in \mathbb{R}^d$, $d$ being the number of

spectral bands. It assumes that standard and anomaly pixels share the same covariance matrix, but have a different mean. The RX algorithm is widely used in remote sensing and has evolved together with the advances in machine learning and signal processing. It has been non-linearized [Kwon and Nasrabadi, 2005] to perform the detection in the mapped feature space through the use of kernel functions and it has been adapted to work with data acquired under a compressive sampling scheme [Fowler and Du, 2012].

**Kernel PCA** Assuming that the *normal* data lies on a manifold and not a linear subspace, the non-linear extension to PCA [Schölkopf et al., 1998] can be used for estimating the manifold and detecting novelties when the reconstruction error is high [Hoffmann, 2007]. In Figure 4.4 (b), the different boundaries obtained using KPCA with various number of principal components are illustrated. Note that the KPCA transformation with the training kernel matrix centered is also called the Kernel Empirical Space (KES) [Xiong et al., 2005]. The KES is an Euclidean embedding of the training data, which preserves the geometrical properties of the feature space. It involves also the eigen-decomposition of the data kernel matrix and leads to the same formulation as KPCA. As for the linear case, when the training samples are corrupted by outliers, a more robust estimation can be obtained using the $\ell_1$ norm instead of the standard $\ell_2$ norm [Xiao et al., 2013].

Finally, KPCA can be used as a pre-processing step which extracts non-linear features. Linear algorithms such as the RX anomaly detector algorithm [Gu et al., 2008] can then be used.

**Manifold perturbation** Another approach relying on the data manifold is called *PERTurbo* [Courty et al., 2011, Chapel et al., 2012]. Perturbo evaluates the novelty of a test sample as the perturbation it induces in the description of the manifold when it is added to it. It is interesting to notice that for an equivalent description of the manifold the formulation of the perturbation measure is equivalent to the KPCA reconstruction error: i.e. an $\epsilon$-ball graph with the heat kernel weights for PERTurbo and an RBF kernel for KPCA, both with similar bandwidths. These two measures are equivalent to a Mahalanobis distance in the feature space.

### 4.4.3 Support Vector Machines and boundary approaches

#### Support Vector Data Description (SVDD)

The *Support Vector Data Description* (SVDD) has been introduced by Tax and Duin [1999] and fits an hypersphere around the class of interest. The radius $R$ and the center $\mathbf{a}$ of the sphere are optimized to obtain the minimum volume containing most of the data. In Figure 4.4 (c), the boundary of the SVD are illustrated. As previously, some slack variables are added to allow for some data outside the sphere. The primal SVDD

optimization problem is

$$\begin{aligned} \min_{R,\mathbf{a},\xi} \quad & R^2 + C\sum_i \xi_i \\ & \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{4.7}$$

with $C$ weighting the penalty term. As for the SVM, Lagrangian multipliers are introduced for the constraints and leads to the following Lagrangian

$$L_p(R,\mathbf{a},\xi,\alpha,\gamma) = R^{\mathbf{2}} + \mathbf{C}\sum_{\mathbf{i}}\xi_{\mathbf{i}} - \sum_{\mathbf{i}}\alpha_{\mathbf{i}}[R^{\mathbf{2}} + \xi_{\mathbf{i}} - (\mathbf{x_i}\cdot\mathbf{x_i} - \mathbf{2a}\cdot\mathbf{x_i} + \mathbf{a}\cdot\mathbf{a})] - \sum_{\mathbf{i}}\gamma_{\mathbf{i}}\xi_{\mathbf{i}} \tag{4.8}$$

The optimal solution is found for the Lagrangian minima where its derivatives are null. The derivative of $L_p$ with respect to $R$, $\mathbf{a}$ and $\xi$ are

$$\begin{aligned} \tfrac{\partial L}{\partial R}: \quad & 2R - 2\sum_i \alpha_i R = 0 \quad \rightarrow \quad \sum_i \alpha_i = 1 \\ \tfrac{\partial L}{\partial \mathbf{a}}: \quad & \sum_i \alpha_i(-2\mathbf{x}_i + 2\mathbf{a}) = \mathbf{0} \quad \rightarrow \quad \mathbf{a} = \frac{\sum_{\mathbf{i}}\alpha_{\mathbf{i}}\mathbf{x}_{\mathbf{i}}}{\sum_{\mathbf{i}}\alpha_{\mathbf{i}}} = \sum_{\mathbf{i}}\alpha_{\mathbf{i}}\mathbf{x}_{\mathbf{i}} \\ \tfrac{\partial L}{\partial \xi_i}: \quad & C - \alpha_i - \gamma_i = 0 \quad \rightarrow \quad \alpha_i = C - \gamma_i \end{aligned} \tag{4.9}$$

The results obtained from the partial derivatives (4.9) replaced into the primal formulation (4.7) gives the *dual* problem

$$\begin{aligned} \max_{\alpha} & \sum_i \alpha_i K(\mathbf{x}_i,\mathbf{x}_i) - \sum_j\sum_i \alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) \\ & 0 \leq \alpha_i \leq C, \quad i = 1,...,N, \quad \sum_i \alpha_i = 1 \end{aligned} \tag{4.10}$$

This QP optimization problem is very close to the SVM dual QP of (3.37). Once solved, the optimal solutions $\alpha^*$ can be used to compute the sphere center $\mathbf{a}$ from (4.9). To evaluate if a test samples lies inside the sphere or not, we need to compute the sphere radius using the distance between the center and any support vector on the boundary: $R^2 = K(\mathbf{x}_{sv},\mathbf{x}_{sv}) - 2\sum_i \alpha_i^* K(\mathbf{x}_i,\mathbf{x}_{sv}) + \sum_{i,j}\alpha_i^*\alpha_j^* K(\mathbf{x}_i,\mathbf{x}_j)$. The decision function of the SVDD compares the distance between a test sample and the center with the radius of the sphere.

$$f(\mathbf{x}_t) = \underbrace{R^2}_{\text{radius}} - \underbrace{K(\mathbf{x}_t,\mathbf{x}_t) - 2\sum_i \alpha_i^* K(\mathbf{x}_i,\mathbf{x}_t) + \sum_{i,j}\alpha_i^*\alpha_j^* K(\mathbf{x}_i,\mathbf{x}_j)}_{\text{distance to center}} \tag{4.11}$$

where the test sample $\mathbf{x}_t$ is declared novel, anomalous or outlier if $f(\mathbf{x}_t) < 0$, meaning that it lies outside the sphere.

The use of SVDD for a novelty detection task is interesting since it does not require any *a priori* on the data distribution and relies for testing only on a few number of training

samples, i.e. the Support Vectors which suffice to describe the boundary. However, hyperparameters related to the kernel function or the outlier rejection level have to be set prior to the training of the SVDD using cross-validation and a fixed false detection rate [Tax and Duin, 2004] minimizing the enclosed volume [Scott and Nowak, 2006, Lee and Scott, 2007].

### One-class Support Vector Machine (OC-SVM)

The *One-Class Support Vector Machine* (OC-SVM) was proposed by B. Schölkopf around the same period as the SVDD. Instead of using a hypersphere to enclose the class of interest, here a hyperplane is used to separate the origin and the class. OC-SVM tries to maximize the margin between the data and the origin in a high-dimensional Hilbert feature space where the data has been mapped by a (non-linear) function $\Phi$ [Scholkopf et al., 2000, Schölkopf et al., 2001]. Compared to the standard SVM, OC-SVM assumes the negative class being at the origin in the feature space. Similarly to SVM, slack variables $\xi_i$ are introduced for better generalization. The primal OC-SVM optimization problem is

$$\begin{aligned} \min_{w,\rho,\xi} \quad & \frac{1}{2}\|w\|^2 - \rho + C\sum_i \xi_i \\ \text{s.t.} \quad & \langle w, \Phi(\mathbf{x}_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{4.12}$$

where the parameter $C$ controls the trade-off between accepting all the data in the class description and having a generalized solution performing well with test data and the quantity $\frac{\rho}{\|w\|}$ is the margin, i.e. the distance between the origin and the hyperplane.

The OC-SVM *dual* problem, derived from the Lagrangian of (4.12) and its partial derivatives, is

$$\begin{aligned} \min_{\alpha} & \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, ..., N, \quad \sum_i \alpha_i = 1 \end{aligned} \tag{4.13}$$

The decision function, based on the optimal solutions $\alpha^*$ of (4.13), is

$$f(\mathbf{x}_t) = \sum_i \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_t) - \rho \tag{4.14}$$

where the non-null $\alpha_i^*$ are called the support vectors.

### Semi-supervised OC-SVM and SVDD

The semi-supervised extension of the OC-SVM and SVDD has been proposed in [Muñoz-Marí et al., 2010]. It is based on the graph-based semi-supervised approach regularizing the decision function accordingly to the data manifold [Belkin et al., 2006]. This regularization which forces the learned function to be smooth on the manifold takes the form of a kernel "distortion". The pairwise similarities among the labeled data, in the form of a kernel, are distorted using a graph built on the labeled and unlabeled data which reflects the underlying data distribution structure.

More formally, we want the learned decision function $f$ to be smooth over the data manifold. The RKHS will be transformed to account for the unlabeled data structure and will provide a training kernel with distorted pairwise similarities. Considering the kernel matrix $K(\mathbf{x}_i, \mathbf{x}_j)$ used for training the OC-SVM or SVDD and the Laplacian $L$ of a graph built on the data, the regularization term for learning the function $f$ is: $\mathbf{f^T M f}$, where the symmetric semi-definite matrix $\mathbf{M} = \varepsilon \mathbf{L}$ and $\mathbf{f} = [\mathbf{f}(\mathbf{x_1}), ..., \mathbf{f}(\mathbf{x_N})]$. This results in the distortion of the kernel as

$$\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) - \mathbf{K}^T_{\mathbf{x}_i}(\mathbf{I} + \mathbf{MK})^{-\mathbf{1}}\mathbf{MK}_{\mathbf{x_j}} \tag{4.15}$$

where $\mathbf{K}$ is the kernel matrix formed by all the data samples (labeled and unlabeled) and $\mathbf{K}_{\mathbf{x}_i} = [K(\mathbf{x}_1, \mathbf{x}_i), ..., K(\mathbf{x}_N, \mathbf{x}_i)]^T$ and $\mathbf{I}$ the identity matrix.

Using the kernel $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j)$ accounting for the data geometry, the decision functions of OC-SVM and SVDD can become semi-supervised and provide tighter boundaries. Notice that the distortion of the kernel requires the inversion of a large matrix of the size of the number of labeled and unlabeled samples. Nyström methods or incomplete Cholesky factorization can be used to alleviate the computational burden [Muñoz-Marí et al., 2010]. Moreover, the new distortion parameter $\varepsilon$ will have to be tuned and a graph built on the data.

The semi-supervised SVDD ($S^3$VDD) is compared with our proposed approaches in Chapter 7, and further investigated with different types of graph in Chapter 10.

**Alternatives and improvements** In Banerjee et al. [2009], the OC-SVM is exploited as an RX detector to estimate the background distribution and detect anomalies in hyperspectral images. The OC-SVM and SVDD have shown to be consistent estimators of density level sets using their regularization parameters for outlier rejection and an appropriate RBF kernel bandwidth [Vert and Vert, 2006]. The observation that density level sets are nested can be imposed to the boundaries at different outlier rejection levels to better match the density distribution [Lee and Scott, 2007, 2010] (see Chapters 5 and 6). An approach that better scales with the size of the dataset than the SVDD and OC-SVM is introduced by Campbell and Bennett [2001]. Their algorithm relies on

minimizing the mean of a decision function similar to the OC-SVM using a simple linear programming problem.

As a matter of fact, SVDD tends to produce loose decision boundaries even with an appropriately tuned RBF kernel or with the help of some labeled novelties. Several approaches addressing this issue have been proposed. In Guo et al. [2009], the post-processing of the boundary allows to tighten it by reducing its distance with its nearest samples. The inclusion of the data density in the objective [Lee et al., 2005, 2007, Yin and Huang], or the minimization of the distance between the sphere center and the center of gravity [Guo et al., 2009] have helped solving this issue.

The problem of clustering using SVDD has been tackled in [Camastra and Verri, 2005] and also in Wang et al. [2006], where the global data structure is retrieved using agglomerative clustering and imposed in the optimization of the class boundary. This is closely related to the hierarchical clustering algorithms investigated in Chapter 6.

**One-class Gaussian Processes** Gaussian Processes (GP) are powerful models for probabilistic regression and classification [Rasmussen and Williams, 2006]. Recently, they have been used in the context of one-class classification [Kemmler et al., 2011]. The predictive function $f$ is not issued from a fixed parametric family of functions, but assumed to be drawn from a probability density function $p(f|\mathbf{x})$, i.e. a GP defines a distribution over functions $f : \mathcal{X} \to \mathbb{R}$. The GP prior stipulates that the values of all functions are jointly normally distributed $p(f|\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$. In other words, the functions $f$ are described by a mean function $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and a covariance $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ function, a kernel function.

In GP, classification problems can be tackled by regressing the labeled values as in logistic regression for classification. For the one-class classification problem, the GP prior is assumed with a zero-mean and the labeled samples are set to $y = 1$. The GP function $f$ will predict 1 on and around training samples and decrease to 0 when moving away from the samples. Additionally, the variance of the prediction will behave in opposition: it will be null on the training samples and large when away from them. These are indicators that can be used to define the class boundary [Kemmler et al., 2011]. Based on the same observations, GP have been used for clustering [Kim and Lee, 2007]: this interpretation is exploited in Chapter 6.

### 4.4.4 Neural network approaches

Bishop [1994] has first proposed to assess the novelty of a sample from the output of an Artificial Neural Network (ANN).

Since then, ANN have been declined into various models handling the information of a sin-

gle class: auto-encoder ANN [Japkowicz et al., 1995] and auto-associative ANN [Japkowicz, 1999, Fernandes et al., 2011], adaptive resonance theory which exploits auto-regressive model parameters to detect novelties [Carpenter et al., 1991], constrained second-order ANN [Moya and Hush, 1996], single-layer recurrent ANN [Rowland and Maida, 2004] or oscillatory models, where a resonance amplification mechanism responds to familiar stimuli [Borisyuk et al., 2000, Borisyuk and Kazanovich, 2004]. All these models are established in the same way the hippocampus region of our brain is behaving for novelty detection. They all try to predict the *normal* behaviour and detect novelties as an erroneous prediction or an important adaptation of the model to the current data. It is worth mentioning the special case of image classification based on image receptive fields ANN which directly works with the images as input and does not need any feature extraction. It detects novelties by analyzing the randomness of the ANN response [Smagghe et al., 2013].

### 4.4.5 Information theoretic approaches

The information-theoretic measures like entropy [Lee and Xiang, 2001] or Kullback-Leibler divergence between distributions [Filippone and Sanguinetti, 2011] have been exploited to describe the normal class and set a threshold to detect anomalies.

The Maximum Mean Discrepancy (MMD) is a measure evaluating if two sets of samples are issued from the same distribution, assuming the sets as random variables. Therefore, MMD can be used as an outlier detector where a new sample is tested against the normal set of samples [Gretton et al., 2012].

## 4.5 Summary

In this Chapter, we have introduced the field of novelty detection and its co-related definitions of anomaly, outlier or target detection. We reviewed the main novelty detection approaches and emphasized on the kernel methods, which are the common point across all the Thesis. The different paradigms of novelty detection span a large area of machine learning methods from clustering (see Chapters 6 and 9) to imbalanced classification (see Chapter 7) and one-class classification (see Chapter 10).

# Part II Novelty detection in clusters

# 5 Background

In this Chapter, we describe the cluster assumption exploited in the rest of this Part of the Thesis and present some background notions on more advanced algorithms for obtaining SVM solutions. These algorithms will be used in the two next Chapters for clustering (Chapter 6) and semi-supervised novelty detection (Chapter 7).

## 5.1 Data made of overlapping clusters

The definition of the data structure is of prime importance when learning with low level of supervision. In this Part of the Thesis, we are assuming that the data structure is made of overlapping clusters of arbitrary shapes. This structure is motivated by the physical properties of optical remote sensing data, as presented in Figure 5.1. The image pixels can be represented as a mixture of elementary material components. Among these pixels, some will be representative of a certain material and the rest will be a mixture of several materials. Most of the methods tackling this type of problems fall in the category of *unmixing techniques* and try to find *pure* pixels representing the sources and the mixing coefficients by assuming linear or non-linear mixing of the sources [Keshava and Mustard, 2002, Dobigeon et al., 2014].

In remote sensing imagery, unmixing resorts to finding the materials that compose each pixels, where a pure pixel, representing a single material, is called an *endmember* and the unmixing procedure corresponds both to the *endmember extraction* and the computation of the mixing coefficients for each pixel [Parente and Plaza, 2010, Plaza et al., 2004, Chang et al., 2011]. This situation happens most often when dealing with low spatial resolution images, where the pixels contain different materials mixed linearly or non-linearly [Villa et al., 2013]. However, we may be interested only in the *dominant* class label of each pixel for classification. In this case, we do not need to resort to unmixing, the estimation of the number of sources and the computation of mixture coefficients. We can simply retrieve groups of clusters in the data by using classification or clustering algorithms. Clustering

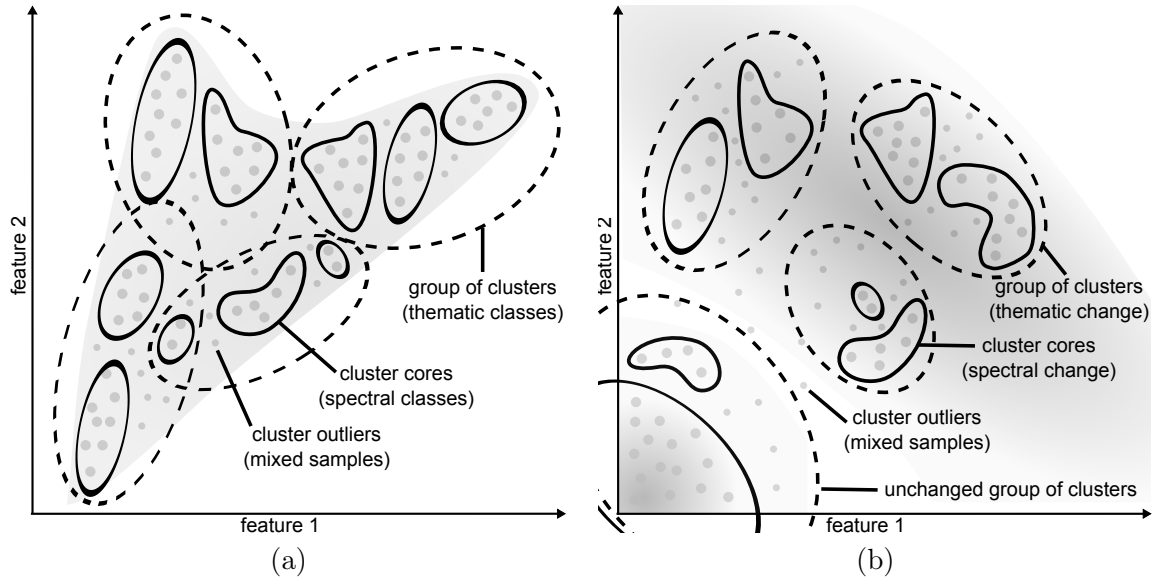



Figure 5.1: Example of data structure with groups of overlapping clusters. (a) The data can be decomposed into groups of clusters, cores and outliers. (b) In a change detection situation, the thematic classes corresponds to different types of changes (or non-change).

and classification approaches appropriate for this type of data structure are presented in Chapter 6 and 7, respectively.

These methods are adapted to the data structure and can retrieve the clusters that may overlap by exploiting the regularization of different forms of SVM to find the appropriate classes boundaries. We present in the following Section 5.2 the support vector clustering algorithm which is well adapted to this type of data. Then in Section 5.3, we present the algorithms that allow to find sets of boundaries for all possible regularization at a low computational cost.

## 5.2 Support Vector Clustering

In this Section, we present the Support Vector Clustering (SVC) algorithm, which can deal with overlapping clusters and which is the foundation for one of the proposed clustering approaches in Chapter 6.

### 5.2.1 Describing cluster boundaries with support vectors

The idea behind the SVC is the following: the data samples are mapped from the input space to a high dimensional feature space using a kernel (typically a Gaussian RBF). In the feature space, the linear boundary that separates most of the data with maximum margin from the origin or the smallest sphere that encloses most of the data is retrieved.

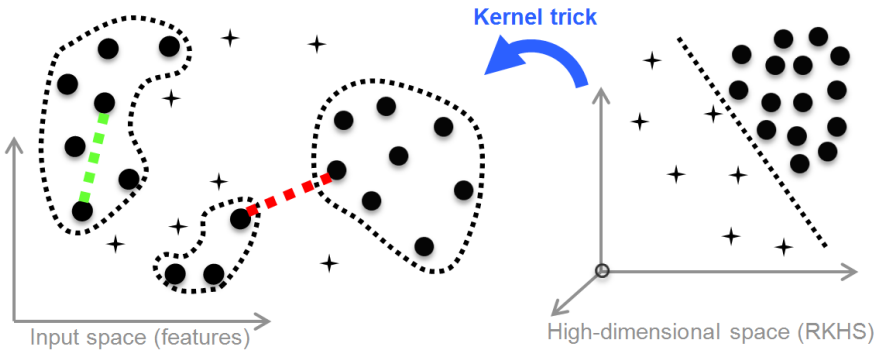



Figure 5.2: Principle of the support vector clustering. The data is linearly separated in a mapped feature space, which defines (multiple) contours in the input space that may be disconnected. If the path linking two inlier samples is going outside the cluster boundaries it indicates that they are from two different clusters.

The first corresponds to the one-class support vector machine (OC-SVM) [Scholkopf et al., 2000, Schölkopf et al., 2001], while the second corresponds to the support vector data description (SVDD) [Tax and Duin, 1999, 2004]. As illustrated in Figure 5.2, this boundary mapped back to data space forms a set of contours which enclose the data samples. These contours can be interpreted as cluster boundaries, where samples enclosed by each separate contour are associated with the same cluster. Actually, depending on the kernel bandwidth and the fraction of rejected outlier samples the contours can be interpreted as a density level set of the underlying probability distribution of the data. When the bandwidth parameter of the RBF kernel is decreased, the number of disconnected contours in data space increases, leading to an increasing number of clusters.

### 5.2.2 Cluster assignment

The problem of assigning a cluster label to each sample has to be addressed. The OC-SVM or SVDD are only defining inlier and outlier samples from the decision boundary which defines the contour of the clusters. However, two inlier samples may be from two different clusters. The idea to obtain the different cluster labels is to look at the line linking every pair of inlier samples in the input space, as illustrated in Figure 5.2. Two samples that have their link between them going outside the cluster boundaries are from different clusters, while two samples that have a link remaining inside the cluster are from the same cluster.

An adjacency matrix $A$ can be established by checking along the link between each pair of inliers.

$$A(i,j) = \left\{ \begin{array}{ll} 1 & \text{if the link between i and j remains inside} \\ 0 & \text{otherwise} \end{array} \right\} \tag{5.1}$$

In practice, the paths are sampled using a certain number of points and the clusters are defined as the connected components of the graph defined by the adjacency matrix $A$.

## 5.3 Entire solution path algorithms

It can be observed that the SVM solutions along different values of the regularization parameters penalizing the training errors are piecewise linear [Hastie et al., 2004, Bach et al., 2006] contrarily to the kernel parameter [Wang et al., 2007]. These solutions can therefore be obtained by *tracing the entire solution paths*. Actually, not only the SVM has piecewise linear solutions but any model with an $\ell_1$ constraint and a quadratic, piecewise quadratic, piecewise linear, or mixed quadratic and linear loss function [Hastie et al., 2001]. The sets of solutions being piecewise linear, the *path* is fully described with the solutions at the interface between two linear pieces, since the other solutions can be linearly interpolated in between. In Figure 5.5, the piecewise linear behaviour of the solutions can be observed for two different types of SVM. This particular property has the advantage of simplifying the description of all the solutions. They can be described by their breakpoints between two linear segments which allows to obtain all the solutions at the same computational cost than a single SVM optimization. The solution path can be constrained to be nested, which provides more coherent classification boundaries along the path [Lee and Scott, 2010]. This way, all the boundaries are included in each other, meaning that the predicted class of a sample changes only once along the path and that a certain hierarchy can be established among the samples. Hereafter, we introduce the solution path algorithms for the OC-SVM and CS-SVM and present the algorithms deriving the nested paths for both of them. These algorithms will be used in Chapters 6 and 7.

### 5.3.1 OC-SVM entire solution path

The entire solution path of the OC-SVM is called the *regularization path* since it provides the set of solutions for all different level of regularization, i.e. the different outlier rejection levels. An alternative formulation replacing the usual regularization parameter $C$ (setting the level of outlier rejection) penalizing the errors by a regularization parameter $\lambda_m$ penalizing the margin has been proposed by Lee and Scott [2007] to allow the derivation of the entire *regularization path*. Given a dataset $\{\mathbf{x}_i\}_{i=1}^N$, its support is found based on the following optimization problem

$$\begin{aligned} \min_{w,\xi} \quad & \frac{\lambda_m}{2}\|w\|^2 + \frac{1}{N}\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & \langle w, \Phi(\mathbf{x}_i)\rangle \geq 1 - \xi_i, \xi_i \geq 0, \forall i \end{aligned} \tag{5.2}$$

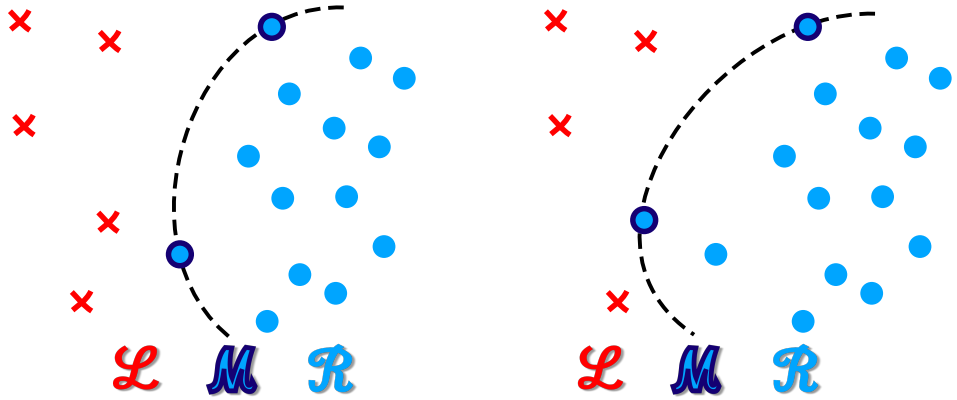



Figure 5.3: The boundary and the active sets of samples ($\mathcal{L}$, $\mathcal{M}$ and $\mathcal{R}$) for two different breakpoints on the solution path of the OC-SVM or cost-asymmetry path of the CS-SVM.

The problem (5.2) is solved through its Lagrangian dual leading to the function

$$f_{\lambda_m}(\mathbf{x}) = \frac{1}{\lambda_m}\sum_i \alpha_{i,m} K(\mathbf{x}_i, \mathbf{x}) \tag{5.3}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function and $m$ is a point on the *regularization path*. The corresponding decision function is

$$g_m(\mathbf{x}) = \text{sign}(f_{\lambda_m}(\mathbf{x}) - 1) \tag{5.4}$$

which is negative for samples inside the support and positive for those outside.

Consider the three *active sets*: $\mathcal{L}$, $\mathcal{R}$ and $\mathcal{M}$ corresponding to inliers, outliers and boundary samples respectively. They are represented in an example in Figure 5.3 and defined as

$$\begin{array}{lll} \text{Boundary:} & \mathcal{M} & = \{i, \quad f_{\lambda_m}(\mathbf{x}_i) = 1\} \\ \text{Inliers:} & \mathcal{R} & = \{i, \quad f_{\lambda_m}(\mathbf{x}_i) < 1\} \\ \text{Outliers:} & \mathcal{L} & = \{i, \quad f_{\lambda_m}(\mathbf{x}_i) > 1\} \end{array}$$

The entire regularization path of the OC-SVM is a continuous piecewise linear function that can be retrieved by starting where all the samples are considered outliers. This happens when $\lambda_m$ is maximum where $\forall i: \; f_{\lambda_m}(\mathbf{x}_i) > 1$. By replacing $f_{\lambda_m}(\mathbf{x})$ using (5.3) and isolating $\lambda_m$ we get $\lambda_{max} = \max_i \frac{1}{N}\sum_j K(\mathbf{x}_i, \mathbf{x}_j)$.

From this starting solution, the $\lambda_m$ should be decreased until a change in the different sets occurs, meaning a sample entering or leaving the boundary, where a new solution $\alpha_m$ is obtained. This is repeated until $\lambda_m = 0$ to get the entire regularization path with solutions for all the $m$ breakpoints retrieved [Lee and Scott, 2007]. All the solutions $\alpha_{i,\cdot}$ in between two $\lambda_m$ can be obtained from a linear interpolation between the lower and upper solutions at $\lambda_m$ and $\lambda_{m+1}$.

### 5.3.2 CS-SVM entire solution path

Among the different parameters of the CS-SVM, see Section 3.4.5, the global regularization parameter $\lambda$ and the cost asymmetry $\gamma$ produce solutions that are continuous piecewise linear [Hastie et al., 2004, Bach et al., 2006]. These are called the *solution paths* of the CS-SVM, with the *regularization path* and the *cost asymmetry path* respectively. Given some labels $\{y\}_{i=1}^{N}$, let us split here again the samples $\mathbf{x}_i$ into three *active sets*, respecting the convention used in [Bach et al., 2006]

$$\begin{array}{lll} \text{Margin:} & \mathcal{M} & = \{i, \quad y_i f_{\gamma,\lambda}(\mathbf{x}_i) = 1\} \\ \text{Left of margin:} & \mathcal{L} & = \{i, \quad y_i f_{\gamma,\lambda}(\mathbf{x}_i) < 1\} \\ \text{Right of margin:} & \mathcal{R} & = \{i, \quad y_i f_{\gamma,\lambda}(\mathbf{x}_i) > 1\} \end{array}$$

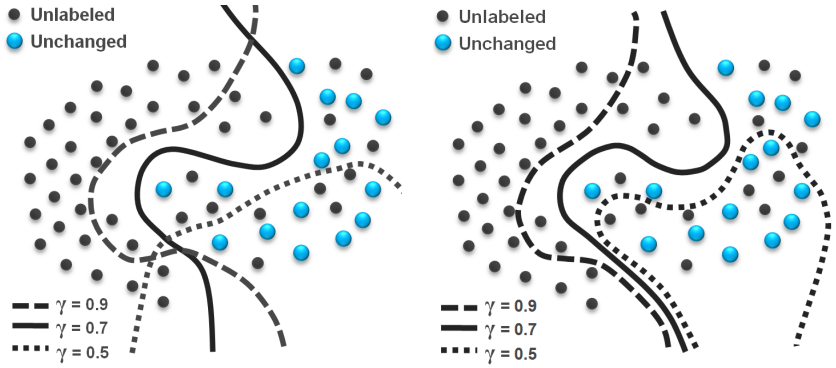
In Figure 5.3 (a)-(b), the three active sets are represented, with the non-null $\alpha_i$ inside ($i \in \mathcal{L}$) and on the margin ($i \in \mathcal{M}$), at two different cost asymmetries $\gamma$. The corresponding $\alpha_i$ for the three different sets are: $\forall i \in \mathcal{M} : 0 \leq \alpha_{i,\gamma} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma$, $\forall i \in \mathcal{L} : \alpha_{i,\gamma} = \mathbf{1}_{\{y_i<0\}} + y_i\gamma$, $\forall i \in \mathcal{R} : \alpha_{i,\gamma} = 0$.

For known active sets $\mathcal{M}$, $\mathcal{L}$ and $\mathcal{R}$, the optimal solutions $\alpha_{i,\gamma}$ can be derived from a linear system of equations. Actually, only the $\alpha_{i,\gamma} \in \mathcal{M}$ are really unknown, since $\alpha_{i,\gamma} \in \mathcal{L} \cup \mathcal{R}$ are either fixed at 0 or $\mathbf{1}_{\{y_i<0\}} + y_i\gamma$. The active sets remain unchanged on a certain range of the parameter resulting in a linear "segment" on the path where the $\alpha_i$ are linearly related to the parameter. The value at which the active sets change produces a breakpoint. The breakpoints can be incrementally computed by tracking the next parameter value for which the KKT conditions are no more satisfied (e.g. the events when samples enter or leave the margin). Figure 5.5 represents the piecewise linear $\boldsymbol{\alpha}$ along the cost asymmetry path.

This algorithm, called CS-SVM PATH, starts considering the path in the situation where all the samples are inside the margin ($\forall i, i \in \mathcal{L}$, $\mathcal{M} \cup \mathcal{R} \in \emptyset$). In this situation, all the $\alpha_{i,\gamma} = \mathbf{1}_{\{y_i<0\}} + y_i\gamma$. This is achieved by the largest regularization enforcing the maximum margin, meaning a null penalization of the errors: $C_+ + C_- = 0 \Rightarrow \lambda = \infty$. However, this solution is reached already at a certain regularization parameter $\lambda_{max}$ above which the solutions stop changing: $\forall \lambda \in [\lambda_{max}, \infty]$. This maximum regularization parameter can be obtained directly from the kernel matrix: since all the samples are inside and on the margin: $y_i f_{\gamma,\lambda}(\mathbf{x}_i) \leq 1$, $\forall i$, by replacing $f_{\gamma,\lambda}(\mathbf{x}_i)$ using (3.41) and isolating $\lambda$, we end up with

$$\lambda_{max} = \max \left[ \gamma \sum_{j \in I_+} y_i y_j K_{i,j} + (1-\gamma) \sum_{j \in I_-} y_i y_j K_{i,j} \right] \tag{5.5}$$

See [Lee and Scott, 2010] for more details on the derivation of (5.5) and on the implementation of the CS-SVM PATH algorithm based on the *SVMPath* toolbox [Hastie et al., 2004].




(a) (b)

Figure 5.4: Example of CS-SVM decision boundaries along (a) the solution path and (b) the nested solution path (NCS-SVM). Enforcing the nested constraint provides a hierarchical set of solutions which are more stable.

## 5.4 Nested decision boundaries

In practice, it can be desired that the solutions along the path respect a certain hierarchy. The solution path, even if piece-wise linear, does not ensure a smooth change at the boundaries, as illustrated in Figure 5.4. In this section, we review how to impose a constraint to nest the boundaries and obtain boundaries that change smoothly along the solution path.

### 5.4.1 Nested OC-SVM solution path

In the nested OC-SVM (NOC-SVM), a nestedness constraint is added to the derivation of the OC-SVM solution path [Lee and Scott, 2010]. Adding a constraint to the optimization problem enforces the generated boundaries of the OC-SVM to be included in each other or *nested*. This produces a set of boundaries for different rejection of outliers that can approximate density level sets with an appropriate RBF kernel bandwidth. Assuming $M$ different level of interest on the path $\lambda_1 > \lambda_2 > ... > \lambda_M > 0$, the optimization problem leading to a family of $m$ nested solutions is

$$\begin{aligned} \min_{\boldsymbol{a}_1,\cdots,\boldsymbol{a}_M} \quad & \sum_{m=1}^{M} \left[ \frac{1}{2\lambda_m} \sum_{i,j} \alpha_{i,m}\alpha_{j,m} K(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \alpha_{i,m} \right] \\ \text{s.t.} \quad & 0 \leq \alpha_{i,m} \leq \frac{1}{N}, \forall i, m \\ & \frac{\alpha_{i,1}}{\lambda_1} \leq \frac{\alpha_{i,2}}{\lambda_2} \leq ... \leq \frac{\alpha_{i,M}}{\lambda_M} \end{aligned} \tag{5.6}$$

with $\alpha_{i,m}$ the Lagrangian multiplier corresponding to the sample $\mathbf{x}_i$ at level set $\lambda_m$. All solutions between two points $m$ on the path are obtained by linear interpolation. The

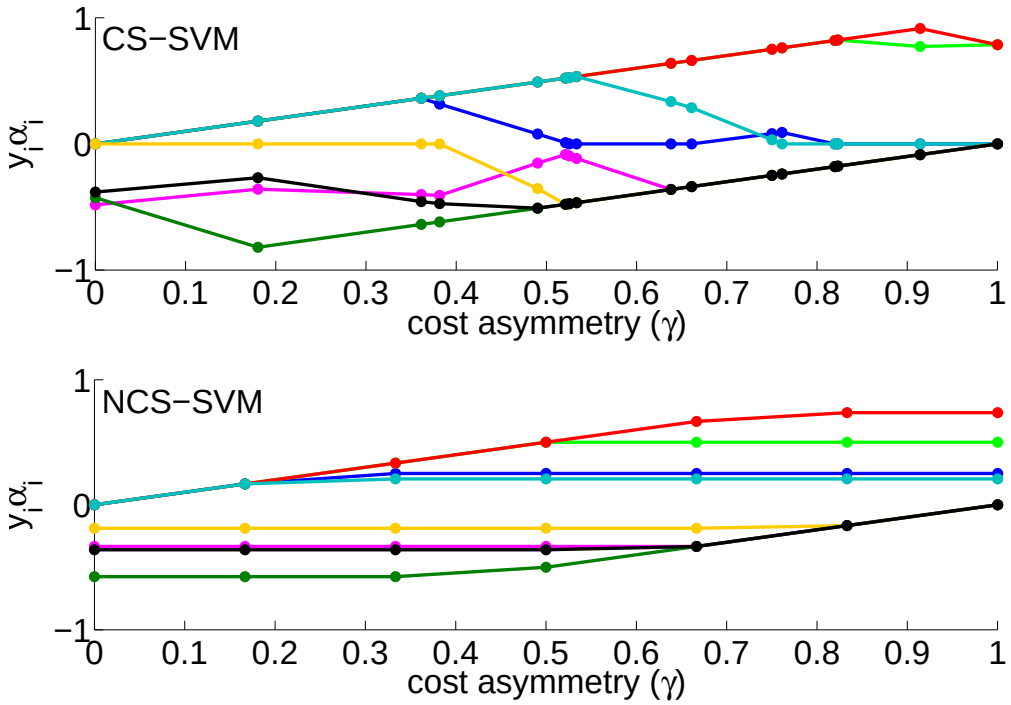



Figure 5.5: The piecewise linear solutions $\alpha_i$ along the path of cost asymmetries $\gamma$ for the CS-SVM (top) and NCS-SVM (bottom) for 8 samples (toy example of two overlapping classes with 4 samples each). Each color represents a different sample and each dot represents a breakpoint where a solution has been obtained. The number of breakpoints of the CS-SVM is usually around $2 \approx 3$ times the number of samples whereas the $M$ breakpoints of the NCS-SVM are manually preset (here $M = 7$).

optimization problem is iteratively solved by splitting (5.6) into sub-problems solving the path for a single sample $\mathbf{x}$ at each iteration, see [Lee and Scott, 2010] for details.

### 5.4.2 Nested CS-SVM solution path

The nested Cost-Sensitive SVM (NCS-SVM) is another formulation for computing the entire cost asymmetry path of the CS-SVM proposed in [Lee and Scott, 2010]. The NCS-SVM forces the boundaries obtained at different cost asymmetries to be nested (an example of nested boundaries can be seen in Chapter 7). Let $G_{\gamma_m}$ be the set of samples in the positive class for the cost asymmetry $\gamma_m$. For a $\gamma_k > \gamma_m$, $G_{\gamma_m} \subseteq G_{\gamma_k}$. A sample on the positive side of the boundary for a certain $\gamma_m$ will remain on the positive side for all $\gamma$ larger than $\gamma_m$. This ensures a coherence along the path of cost asymmetries and should provide less sensitivity to the free parameters (regularization and kernel parameters) and to the noise in training data.

Nested solution paths are also continuous piecewise linear function of the cost asymmetry

parameter. The nested solution paths are monotonic due to the nestedness constraint, as it can be observed in Figure 5.5. The breakpoints along the cost asymmetry path are pre-defined by the user and not derived from the data as in the CS-SVM PATH algorithm (see section 5.3.2). The monotonicity of the path allows pre-defining only a small number of breakpoints. In this setting, the NCS-SVM dual formulation is

$$\min_{\alpha_{i,1},\dots,\alpha_{i,M}} \sum_{m=1}^{M} \left[ \frac{1}{2\lambda} \sum_{i,j} \alpha_{i,m}\alpha_{j,m} y_i y_j K_{i,j} - \sum_i \alpha_{i,m} \right] \tag{5.7}$$

$$\text{s.t. } 0 \leq \alpha_{i,m} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma_m, \ \forall i,m \tag{5.8}$$

$$y_i\alpha_{i,1} \leq \dots \leq y_i\alpha_{i,M} \ \forall i \tag{5.9}$$

This dual formulation is extremely close to the dual formulation of the CS-SVM in (3.40), see Section 3.4.5. The differences lie in i) the simultaneous consideration of all the breakpoints and in ii) the additional constraints in (5.9) enforcing the nesting of the boundaries. The solutions $\alpha_{i,m+\eta}$ at an intermediate cost asymmetry $\gamma_{m+\eta}$ ($0<\eta<1$) can be obtained by linear interpolation.

The optimization problem of (5.7) is complex since the constraints are imposed on the samples (5.8) and the cost asymmetries (5.9) resulting in $M \times N$ variables ($N$ being the number of labeled samples). A decomposition algorithm is proposed to split the QP problem into smaller subproblems. The sample $\mathbf{x}_k$ violating the KKT conditions the most is selected and its solutions $\alpha_{k,m}$ are optimized along the breakpoints while the other solutions remain fixed. The objective function of (5.7) is thus rewritten as

$$\sum_{m=1}^{M} \left[ \frac{1}{2\lambda} \sum_{i,j} \alpha_{i,m}\alpha_{j,m} y_i y_j K_{i,j} - \sum_i \alpha_{i,m} \right] = \sum_{m=1}^{M} \left[ \frac{1}{2\lambda} \alpha_{k,m}^2 K_{k,k} + \alpha_{k,m} \left( \sum_{j \neq k} \alpha_{j,m} y_j y_k K_{k,j} - \lambda \right) \right] + Cst$$

where $Cst$ contains the terms not related to $\alpha_{k,m}$. Samples are selected iteratively to update their solutions until the optimality condition error is under a predefined value[1].

[1] For more details on the sample selection and termination criterion of the NCS-SVM, please refer to Lee and Scott [2010]

## 5.5 Summary

In this Chapter, we have introduced the problematic of retrieving the cluster structure of the data composed of overlapping clusters of arbitrary shapes. We have described extensions of the OC-SVM and CS-SVM providing the entire set of solutions related to the regularization parameters and their nested counterpart providing hierarchical solutions. These will be further exploited in next Chapters for clustering (Chapter 6) and semi-supervised novelty detection (Chapter 7).

# 6 Unsupervised change detection via hierarchical clustering

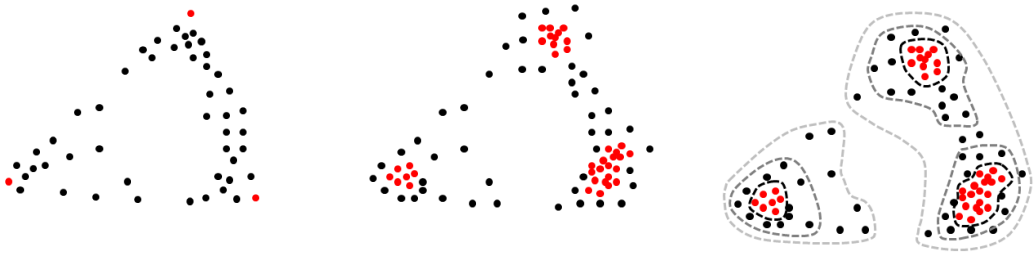
In this Chapter[1], we first formally introduce the structure assumption made on the data already presented in Section 5.1. Then we present two hierarchical clustering algorithms, the first exploiting the OC-SVM entire solution path algorithms in Section 6.2.1, the other based on Gaussian Process in Section 6.2.2. We propose a cluster validity measure for hyperparameter tuning in Section 6.3 and a merging system for unsupervised change detection in Section 6.4. Finally in Section 6.5, we present experiments on two multispectral datasets to confirm the appropriateness of the proposed system in change detection situations.

## 6.1 Clustering and outlier hierarchy

The more general setting assuming overlapping clusters is less restrictive than standard unmixing where unique pure sources are considered [Villa et al., 2013]. The samples representative of the clusters form the cluster cores which are surrounded by the mixed samples or noisy samples from the data inherent variability. These mixed samples preserve the information on the relations between the clusters: clusters strongly mixed between each other are more likely to be merged in the same group of clusters. Whereas clusters with only a few mixed samples are less likely to be grouped together.

Let us consider a dataset $\mathbf{X} \in \mathbb{R}^{N \times d}$ made of $N$ samples $\mathbf{x}_i, i = [1, .., N]$ and $d$ dimensions. The aim is to attribute labels $y_i$ corresponding to a certain cluster $C_k$. If the sample is too far away from the support of its cluster (i.e. in a low density region) or in the ambiguous region between two clusters, it will most probably be considered as an outlier or a mixed sample. Thus, the dataset can be decomposed in cluster cores and outliers around them, which might contain mixed samples. An example of cluster cores and outliers on a synthetic dataset is presented in Figure 6.1 along with the typical *unmixing* setting. The hierarchical levels represent the likelihood of being an outlier and corresponds

[1] This Chapter is based on [de Morsier et al., 2012b, 2014b]



(a) (b) (C)

Figure 6.1: Two dimensional examples of (a) standard unmixing problem with 3 endmembers (red dots) and mixed samples (black dots), (b) the more general case with 3 clusters (red dots) surrounded by noisy and mixed samples (black dots) and (c) the corresponding outlier hierarchy levels.

in our case to the "representativity" of the sample regarding its cluster core.

Let us define properly the different components of a dataset as following:

- Core: a cluster core is the homogeneous group of similar samples representing the cluster and its natural variability.
- Outliers: outliers are least representative samples of a certain cluster. They may have been corrupted by noise or be a mixture of several cluster cores.
- Cluster: a cluster is made of an homogeneous core of samples possibly surrounded by outliers (noisy samples) or mixture of clusters.
- Class: a class is a thematic (user-defined) group of similar samples usually composed of one or several clusters, but not necessary.

This representation is interesting since it groups most representative samples inside cores and less representative mixed samples as outliers. The definition of the cluster is cleaner and no more perturbed by mixed or noisy samples. The classes, made of one or several cluster cores, can be semantically defined by the user interests or naturally discovered depending on the relative structure of the cores and of their outliers. The rationale behind it being that two cores sharing most of their outliers have a higher probability of having mixed samples issue from their combination and should naturally belong to a common class or group of clusters.

Such a description of the data with an appropriate separation of core samples and outliers can be difficult to obtain directly. It is often easier to obtain a complete hierarchical representation of the data with various levels of outlier rejection, as presented in Figure 6.1 (c). At the bottom level, the few most similar samples represent the cluster cores and

all the other samples are considered outliers. Going up in the level of outlier hierarchy, cluster cores grow as more samples are included. This reduces the number of outliers until reaching the top of the hierarchy where all the samples are inside cluster cores and there are no outliers left. The validity measure presented in the next section allows to select the appropriate separation between cluster cores and outliers.

## 6.2 Hierarchical clustering algorithms

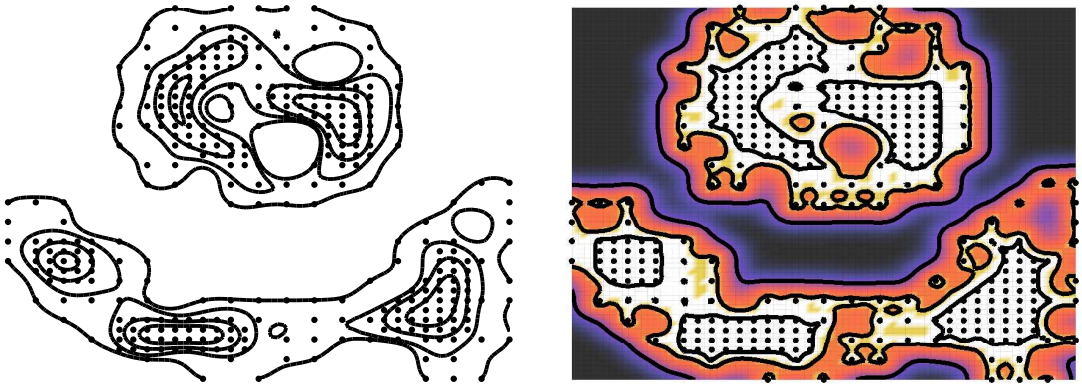
In this section, we introduce two algorithms for clustering with hierarchical outlier rejection. The first one is a variant of Support Vector Clustering leading to a set of truly hierarchical clustering solutions [de Morsier et al., 2012b]. The second one is the Gaussian Processes Clustering [Kim and Lee, 2007] from which a set of hierarchical solutions is derived. These two algorithms will be tuned using a cluster validity measure (CVM) presented in Section 6.3 and groups of clusters will be obtained using a merging system (CMS) presented in Section 6.4. Finally, two experiments with multi-temporal multi-spectral images are presented in Section 6.5.

### 6.2.1 Hierarchical support vector clustering

The Support Vector Clustering (SVC), introduced in Section 5.2, is closely related to the One-Class Support Vector Machine (OC-SVM). Let us recall that the OC-SVM estimates a level set (usually the support) of a distribution in a high dimensional space induced by a mapping function $\Phi$ [Schölkopf et al., 2001]. The OC-SVM separates the data with a hyperplane at maximum distance from the origin. Coming back in the input space, the level set boundary usually forms disconnected "islands", which correspond to different modes of the data distribution, as it can be observed in Figure 6.2. In novelty detection, these clusters may represent the normal and novelties distribution $p_0$ and $p_1$. These distinct modes, representing regions of highly similar samples, form different clusters. The OC-SVM has the ability of rejecting the outliers and produce clusters with inliers inside the cluster boundary and outliers left outside the cluster boundary. This property is extremely useful in the case of overlapping clusters where an important number of mixed pixels is present. If some of the samples are not rejected, the clusters will fuse together and will not be resolved as distinct clusters.

In order to obtain a set of boundaries with different level of outlier rejection, the *entire regularization path* of the OC-SVM, presented in Section 5.3.1, can be exploited [Hansen et al., 2007]. This produces a pseudo-hierarchical set of solutions, since the OC-SVM entire solution path does not provide boundaries that are included in each other which ensures a hierarchy. An example of nested boundaries is presented in Figure 6.2.

In order to derive a truly-hierarchical set of SVC solutions, the Nested OC-SVM (NSVC)



(a) (b)

Figure 6.2: Nested boundaries of the (a) hierarchical support vector clustering and (b) Gaussian Process clustering algorithms at different level of outlier rejection for the example of Figure 6.3. The negative predicted variance of the Gaussian Processes clustering algorithm is overlayed.

presented in Chapter 5 is used to generate solutions from which clusters are discovered. The decision function of the NSVC is

$$g_m(\mathbf{x}) = \text{sign}(f_{\lambda_m}(\mathbf{x}) - 1) \tag{6.1}$$

which is negative for samples inside the support (inliers) and positive for those outside (outliers). The recovery of the cluster labels from the OC-SVM decision function is further explain in Section 6.2.3 after having presented the Gaussian process clustering, which assigns cluster labels in the same way.

### 6.2.2 Hierarchical Gaussian process clustering

The GP, originally introduced for regression, has also been applied to clustering [Kim and Lee, 2007]. As it can be observed in Figure 6.2 (b), the predicted variance of a GP learned on a dataset $\mathbf{X}$ is small in dense regions and large in regions with few samples. Therefore the variance characterizes the data probability density function and can be exploited for clustering in a similar way than the SVC model. The disconnected "islands" formed by level set of the predicted variance represent the different clusters. It is important to note that the true data probability density function can be retrieved from OC-SVM or GPs only for a specific set of hyperparameters. On the contrary to Kim and Lee [2007] who learn one set of hyperparameters, we do not restrict the clustering solutions to be density level set estimations. We select the hyperparameters based on a cluster validity index presented in Section 6.3.

The predicted variance of a GP is derived as follows: the output labels $y$, equal to 1 for all the samples in the support and 0 otherwise, are assumed to be generated from a latent

function $f : \mathcal{X} \to \mathbb{R}$. The function $f$ is assumed normally distributed, which corresponds to putting a Gaussian prior on the distribution of functions $f$. The function values are therefore jointly normally distributed

$$\mathbf{f}|\mathbf{X} \sim \mathcal{N}(\mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})), \tag{6.2}$$

which can be explicitly expressed as

$$p(\mathbf{f}|\mathbf{X}, \Theta) = \frac{1}{(2\pi)^{N/2}|K(\mathbf{X}, \mathbf{X})|^{1/2}} \cdot \exp -\frac{1}{2}(\mathbf{f} - \mu)K(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f} - \mu) \tag{6.3}$$

The mean $\mu$ of the GP prior is usually set to 0 and an additive Gaussian noise with variance $\sigma_n^2$ is also considered. The predictive distribution for new samples $\mathbf{x}$ is defined by the following mean and variance

$$\mu(\mathbf{x}) = K(\mathbf{X}, \mathbf{x})^T (K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \tag{6.4}$$

$$\sigma^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) - K(\mathbf{X}, \mathbf{x})^T (K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} K(\mathbf{X}, \mathbf{x}) \tag{6.5}$$

The covariance function $K$ defines the way samples generalize to neighbours and is selected as: $K(\mathbf{x}_i, \mathbf{x}_j) = s \cdot \exp\{-\frac{1}{2\sigma}\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$. The hyperparameters $s$ and $\sigma$ can be selected by maximizing the pseudo-likelihood on the data or, as in our case, using the CVM. The cluster cores of the data are estimated based on the negative variance of the prediction: $f(\mathbf{x}) = -\sigma^2(\mathbf{x})$, where low variance correlates with higher density and vice-versa [Kim and Lee, 2007].

We propose to find a hierarchical set of solutions by considering different level sets of the negative predicted variance, denoted HGPC. A Gaussian Process clustering solution at level $m$ is defined by

$$g_m(\mathbf{x}) = \underbrace{-\sigma^2(\mathbf{x})}_{f(\mathbf{x})} + \underbrace{(\Theta_{min} + \frac{(m-1)}{M} \cdot (\Theta_{max} - \Theta_{min}))}_{\text{the negative variance rescaled at level set } m} \tag{6.6}$$

with $m = \{1, .., M\}$, $\Theta_{max} = \max_{\mathbf{x} \in \mathbf{X}} \left(-\sigma^2(\mathbf{x})\right)$ and $\Theta_{min} = -s$. Samples with $g_m(\mathbf{x}) \geq 0$ are inliers and samples with $g_m(\mathbf{x}) < 0$ are outliers.

### 6.2.3 Cluster assignment from hierarchical adjacency matrices

The hierarchical sets of solutions obtained from either the NSVC or HGPC are not yet clustering solutions. The decision function $g_m(\mathbf{x})$ defines only if a sample $\mathbf{x}$ is an inlier (inside a cluster core) or an outlier (outside cluster core) but it does not indicate if two inliers belong to the same cluster. To assign cluster labels to all inliers, we use the following strategy: the path linking a pair of inlier samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the input space will cross an outlier region ($g_m(\mathbf{x}_{\text{onpath}(i,j)}) < 0$) if they are from two different clusters, as

by analogy we would have to swim to pass from an "island" to another one. Following this intuition, an adjacency matrix $A_m$ for each specific level $m$ can be built by checking the $g_m$ function along the path between each pairs of inliers.

$$A_m(i,j) = \left\{ \begin{array}{ll} 1 & \text{if } g_m(\mathbf{x}_i + \epsilon(\mathbf{x}_j - \mathbf{x}_i)) \geq 0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{6.7}$$

with $\epsilon \in [0;1]$. In practice, the paths are sampled using around 10 different values for $\epsilon$. Several methods based on graphs have been proposed to accelerate the computation of the matrix $A$. Delaunay diagram, Minimum Spanning Tree or $k$-NN graphs can be built on the data and allow to reduce the number of sample pairs checked [Lee and Lee, 2005].

The cluster memberships are retrieved from the adjacency matrix $A_m$ by following the existing links between inlier samples and propagating cluster labels. For outliers, memberships are assigned to the closest cluster using the Euclidean distance. Further acceleration is obtained by starting with the lowest outlier rejection level and setting $A_m(i,j) = 0$ for all the next levels if $A_m(i,j) = 0$ for the current level. This is justified by the hierarchical structure adopted, which makes it impossible for two samples being in two different clusters to be found in a common cluster for higher rejection levels.

## 6.3 Validity measure for hierarchical clustering

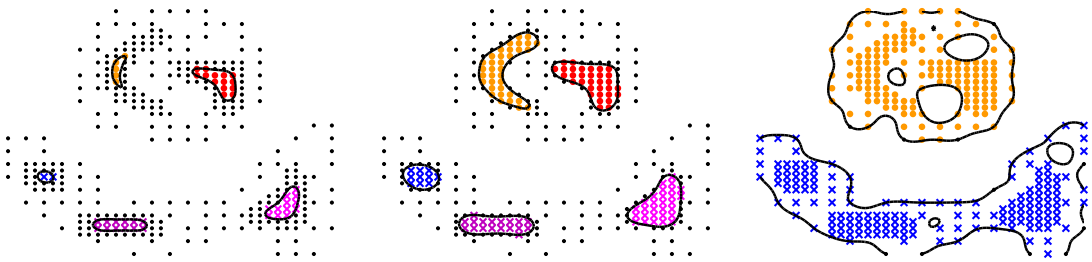
Many cluster validity indices are based on notion of compactness and separability of the clusters introduced by Dunn [Dunn, 1974]. Transferring the notion of compactness to a hierarchy of cluster cores can be dangerous since this would always select hierarchical levels having the most compact cluster cores, e.g. the small and dense clusters in Figure 6.3 (a). Instead of compactness, the cluster cores sizes could be maximized together with their separation. This leads to cluster cores covering the majority of the samples, but leaving out samples in the overlapped region between clusters.

Consider a set of $M$ solutions from a hierarchical outlier rejection method, where at $m = 1$ most of the samples are rejected as outliers and at $m = M$ all samples are inside cluster cores. Among the $M$ different levels of clustering, the solution in the hierarchy fitting the most our representation with homogeneous clusters core surrounded by outliers should be selected.

We will consider the following measures to build up our validity index handling clusters of arbitrary shape having an homogeneous core and outliers.

**Definition 4.** *The core size is the maximal distance among two of the cluster core samples.*

$$d_{in}(k) = \max_{i,j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \tag{6.8}$$



(a) CVM=2.5 (b) CVM=4.6 (c) CVM=1.8

Figure 6.3: Different boundaries obtained by running a hierarchical support vector clustering algorithm with different level of outlier rejection on a synthetic dataset. Their corresponding Cluster Validity Measure (CVM) are indicated at the bottom. The kernel bandwidth $\sigma$ corresponds to $1 \cdot \sigma_0$ (i.e. mean distance among dataset samples) and the level of outlier rejection are $m = 4, 6, 10$. Too small cluster cores or too large inhomogeneous cluster cores result in small CVM.

The core size is defined by the distance $d_{\text{in}}(k)$ which is the maximal size of the core of cluster $k$. $C_k$ is the set of inliers in the $k$-th cluster.

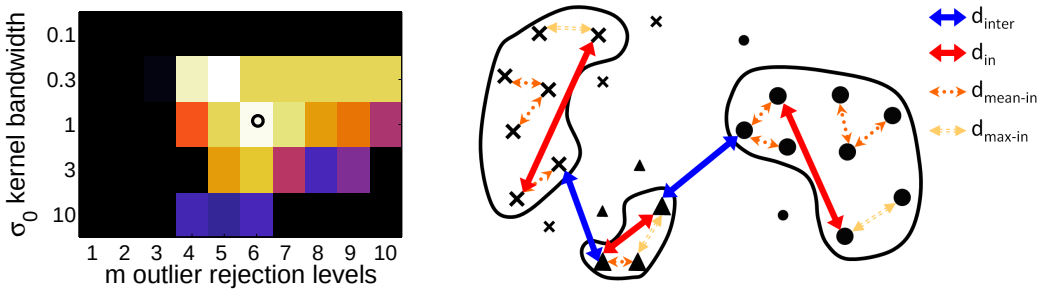
**Definition 5.** *The core separability is the minimum distance between a core sample to the closest sample from another cluster core. It is defined as*

$$d_{inter}(k) = \min_{i \in C_k, j \in C_l, k \neq l} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \tag{6.9}$$

The distance $d_{\text{inter}}(k)$ is the minimal distance between a sample of cluster core $k$ to a sample of any other cluster core $l$. This measure qualifies the smallest separation existing with any other cluster. This measure is equivalent to the single linkage criterion for merging two groups of samples [Hartigan, 1975].

Based on these definitions of core size and separability, we can design a cluster validity measure that is high for large and separated clusters. The usual notion of compactness cannot be used with a hierarchical set of solutions, since it would select solutions on top of the hierarchy where almost all the samples are considered outliers apart from the cluster centers. Therefore the different cluster sizes $d_{\text{in}}$ should be maximized at the same time that the separability of the clusters $d_{\text{inter}}$.

However, clustering solutions including noisy samples in the cluster core description, such as the external boundaries in Figure 6.1 (c), should also be penalized. We define a measure of core homogeneity that rescales the core size accordingly. In Wang and Chiang [2008], a homogeneity criterion is used to fix the regularization parameter of the Support Vector Clustering algorithm before the application of a standard validity measure. On the contrary, the homogeneity ratio we propose is directly included in the CVM and can be therefore applied with any clustering algorithms. Moreover, the criterion does not




(a) (b)

Figure 6.4: (a) CVM values for different solutions obtained by support vector clustering with different level of outlier rejection and kernel bandwidth on the data illustrated in Figure 6.3. (b) Example of distances involved in the CVM proposed.

represent a global density of the cluster cores, since different cores can have different densities. The homogeneity of each cluster core is maximized independently from the others.

**Definition 6.** *The core inhomogeneity is the ratio between the maximum and the average smallest inner distances (nearest neighbour) inside the cluster core. It is defined as*

$$R_{in}(k) = \frac{d_{max\text{-}in}(k)}{d_{mean\text{-}in}(k)} = \frac{\max\left(\min_{i,j\in C_k, i\neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)}{\frac{1}{|C_k|}\sum_{j\in C_k}\left(\min_{i\in C_k, i\neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)} \tag{6.10}$$

The ratio $R_{\text{in}}$ reflects the density inhomogeneity inside a cluster core. The larger the ratio, the more important the inhomogeneity of the core. A ratio close to 1 indicates a very homogeneous core.

In Figure 6.3, the CVM values are computed on three different clustering solutions obtained on a synthetic dataset made of five cores surrounded by noisy samples. In Figure 6.4 (a), the CVM values for various clustering solutions obtained from different parameters on the same synthetic data are presented. Finally in Figure 6.4 (b), the distances involved in the CVM are presented.

Summing up, the proposed CVM maximizes the *rescaled* cluster core sizes $d_{\text{in}}$ by $1/R_{\text{in}}$ and the cluster cores separability. This way solutions with homogeneous cluster cores well separated are favourized while the ones containing outliers inside cluster cores are avoided, since they lead to small rescaled cluster core sizes with low homogeneity ($R_{\text{in}} << 1$). The

CVM can be expressed as follows

$$\text{CVM} = \arg\max_{\lambda,\sigma} \sum_{k=1}^{M} \frac{d_{\text{in}}(k)}{R_{\text{in}}(k)} \cdot \sum_{k=1}^{M} d_{\text{inter}}(k) \tag{6.11}$$

The CVM becomes small for solutions with cluster cores being too small, too inhomogeneous (including too many outliers) or too close. This translates directly the trade-off between cluster core size and outlier rejection.

## 6.4 Cluster merging criteria

Finding the natural cluster cores will most probably over-partition the data with respect to the user needs or the scale at which the data is investigated. The user is interested in a certain number of classes depending on the targeted application. However, and independently from the number of classes desired, the general correspondence between classes and clusters will be that a group of several cluster cores matches a certain class. For example in change detection analysis, the final number of classes of interest is two: changed and unchanged classes. These classes represent an ensemble of "changed" and an ensemble of "unchanged" cluster cores. However, special care should be taken when additional changes are present which are not of interest for the user. An example of such situations is the detection of clouds as changes and cloud-free as unchanged, however additional changes will be detected from the cloud shadows. These are the type of situations were it is useful not to have a pre-fixed number of clusters and avoid the merging.

In the following, we propose a merging criterion allowing to group the cores one after the other ending up with the desired number of cluster ensembles. The cluster merging criterion could be application-dependent, since each application implies different underlying assumptions. We present here the main idea, which is valid for all situations, as well as an additional criterion that can be easily added or removed for specific cases where a very large cluster core would absorb all small surrounding cores, as it is often for change detection applications. In this specific application, we expect the merged solution to be extremely asymmetric (a group of large cluster cores and a group of very small cluster cores corresponding to the changed areas). This situation is depicted in Figure 6.5 (b).

### 6.4.1 Outlier dispersion as merging mechanism

The merging criterion proposed in [Wang and Chiang, 2008] assumes Gaussian distributed clusters and tests the overlap between the projection of clusters probability density functions on the different variables. This parametric approach lacks of flexibility and tends to evaluates the distance between cluster centers considering their respective

sizes. Our merging criterion is indirectly related to the distance between cluster cores by exploiting the outlier dispersion in between them. Two cluster cores have a high probability of being part of the same ensemble if they are close to each other and have shared outliers between them.

The outlier dispersion between two clusters $C_k$ and $C_l$ should be qualified only by the subset of outliers lying approximately in between the two cores $k$ and $l$ and not with outliers elsewhere, which are potentially far away in the input space. The dispersion between outliers can be reflected by the minimal distances between the set of outliers from cluster $k$ to cluster $l$. A large distance will underline the existence of a gap between the two groups of outliers and will tend to the distance between the two cluster cores; on the contrary a small distance will show the continuity between the two groups of outliers, which belong to an unified group of mixed samples from the two clusters.

**Definition 7.** *The average minimal distance between the outliers of each cluster, considering only the outliers in between the two cluster cores.*

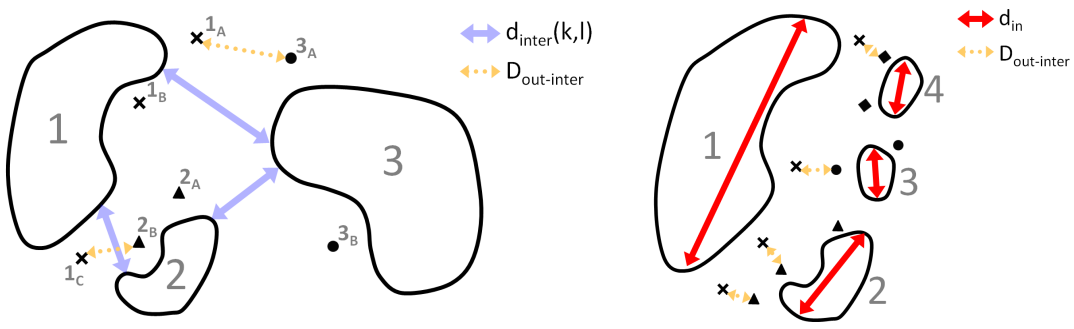
$$d_{out\text{-}inter}(k,l) = \frac{1}{|O_k^*|} \sum_{i \in O_k^*} \min_{j \in O_l, l \neq k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \tag{6.12}$$

where $O_l$ is the set of outliers of the l-th cluster and $O_k^* = \{i, \min_{j \in O_l, l \neq k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq d_{inter}(k,l)\}$ defines the set of the outliers located approximately between the two cluster cores, i.e. the outliers of the $k$-th cluster at a distance of the closest outlier of the $l$-th cluster no larger than the distance between the two cluster cores. Notice that the distance between clusters $d_{\text{inter}}(k,l)$ is evaluated between two arbitrary clusters $k$ and $l$, instead of (6.9) where the distance is between $k$ and its closest neighbour. This restricted set $O_k^*$ allows to exclude outliers that are far apart from the two cluster cores and to avoid involving them in the distance qualifying the outliers dispersion between the two cores. If no outliers are present in between the cores, $d_{\text{out-inter}}(k,l)$ is set equal to $d_{\text{inter}}(k,l)$.

The first cluster merging score (CMS A) proposed to merge pairs of clusters is actually based on the outliers dispersion of the two cluster cores. The first cluster cores to merge will be those sharing outliers close to each other, intrinsically meaning that they belong to the same cluster ensemble.

$$\text{CMS A} = \arg\min_{k,l,k \neq l} \left(d_{\text{out-inter}}(k,l) + d_{\text{out-inter}}(l,k)\right) \tag{6.13}$$

The merging order of clusters is obtained by listing the couple of clusters $k$ and $l$ in ascending CMS values. The first couple of clusters $k$ and $l$ to be merged are the ones having the minimal CMS value.



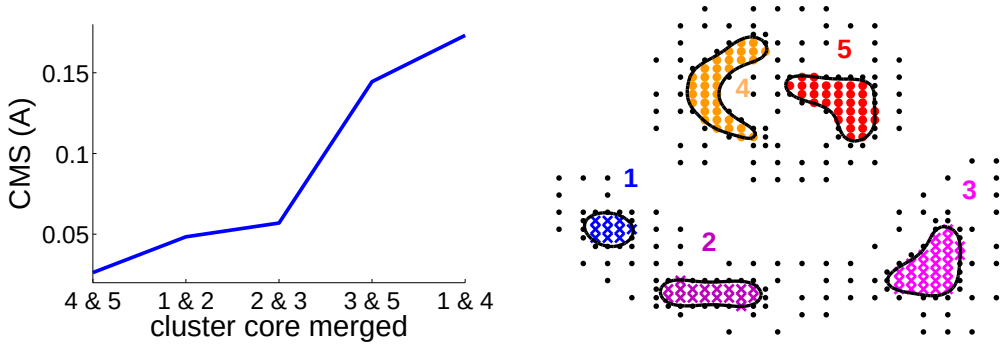

(a) (b)

Figure 6.5: (a) Examples of distances involved in the CMS criteria A & B. Distances between outliers lying in between two cluster cores, thus contributing to $d_{out-inter}(k,l)$, are depicted by dashed arrows. The distance between $1_A$ and $3_A$ is smaller than $d_{inter}(1,3)$ and is therefore kept as a distance representing the outlier dispersion between the cluster cores 1 and 3. In opposition, the distance between $1_A$ and $3_B$ is not considered since larger than $d_{inter}(1,3)$. (b) Example of a large cluster with outliers that would merge with all the smaller surrounding cores. To avoid this type of situation, the cluster core sizes difference is taken into account in the CMS B criterion. The merging of a larger cluster core with a much smaller one is penalized by their sizes difference.

### 6.4.2 Penalizing different core sizes

In some applications, we expect the cluster ensembles to be very asymmetric, as presented in Figure 6.5 (b) where a very large cluster core is surrounded by smaller ones. The risk with the previous merging criterion is that the small clusters have few outliers resulting in merging score close to their distance of separation, while the large cluster may have more outliers potentially situated close to the outliers of the small clusters. Therefore, the difference in core sizes should be penalized to avoid the small cluster core being merged one after the other with the larger one. This would result in an unbalanced situation in terms of cluster sizes which is usually not desirable. In the specific case of change detection, the number of ensemble expected is two: changed and unchanged areas. However, the two ensemble are often of different size, e.g. small and dense cluster cores for well defined changes and large for the more generic unchanged class containing variations that are not considered as changes but still enlarging the class cluster ensemble. These considerations lead to propose a second merging score (CMS B), which penalizes different core sizes by adding their absolute size difference.

$$\text{CMS B} = \text{CMS A} + |d_{\text{in}}(k) - d_{\text{in}}(l)| \tag{6.14}$$




(a) (b)

Figure 6.6: (a) CMS values in ascending order for the *5cores* synthetic data and (b) its optimal CVM solution. Cluster numbers start from bottom-left to bottom-right (cores 1 to 3) then top-left to top-right (cores 4 and 5).

### 6.4.3 Setting the merging level

Both merging criteria provide an order for cluster merging but no direct information on how many cluster should be merged. The merging level, i.e. how many cluster cores are grouped before stopping, is left to the user or driven by the application. In most cases, the user is interested in a specific number of cluster ensembles corresponding to the final number of classes. Therefore, the user needs only to merge the cluster cores until the desired number of ensembles is reached. In cases where the number of cluster ensembles is not known *a priori*, the CMS score can be exploited to discover a natural merging level. One could investigate the values of the CMS in ascending order and merge clusters up to a knee. Such a breakpoint in the CMS score indicates that the merging of clusters changes in scale. The CMS values for the synthetic data of Figure 6.3 are reported in ascending order in Figure 6.6.

### 6.4.4 Computational considerations

The previously presented validity and merging criteria are computationally intensive. They imply computations on many different pairs of samples in the training dataset: the validity measure requires all the pairs of cluster core samples and the merging criterion all the pairs of outliers. Several approaches based on graphs have been introduced to decrease the computational burden of other validity measures [Pal and Biswas, 1997]. Minimum Spanning Tree (MST) or Relative Neighbourhood graph (RNG) among others, allow to consider only certain pairs of samples being the graph edges. For example, the homogeneity of a cluster core could be approximated by the ratio between the maximum edge and the average of edges inside a cluster core. However, the size of a cluster core cannot be efficiently computed from the graph edges. It can be replaced by the maximum edge contained in the cluster core as proposed in [Yousri et al., 2008] and the

core separability can be computed as the minimum edge between the cluster core and another core. This approximation of the cluster validity measure would not preserve the homogeneity of the cluster cores, the ratio of the core size and homogeneity resulting only in the average of edges inside a cluster core. Concerning the cluster merging criterion, graph edges between the outliers could be constructed and used as dispersion measure. These approximations are too important to be considered in our experiments and we leave for further research the study of the impact of these approximations on the clustering results.

## 6.5 Experiments

We present in this section experiments on unsupervised change detection in two different natural disasters settings, wildfire and floods detection. The automatic detection of changes in multi-temporal multispectral images can be performed by finding the two groups of pixels corresponding to changed and unchanged pixels. These two groups are retrieved by the clustering algorithms and the CVM/CMS system proposed.

### 6.5.1 Experimental setup

The two multitemporal remote sensing datasets used are the *Gloucester floods* and the *Bastrop fires*. They are from different sensors and contain different types of changes: a river flood and a forest fire. Both datasets are presented in Appendix B.1.2 and B.1.1.

In this unsupervised setting and with bi-temporal pair of images, the DIA is an appropriate choice. Thus the features are the stack of the difference image and we also add the *NDVI* indices at the two different acquisition periods.

The experiments involve the SVC, its nested equivalent (NSVC) and the Hierarchical Gaussian Process Clustering (HGPC) with both $M = 10$ different outlier rejection levels. The kernel bandwidth parameter, $\sigma = [1/10, .., 2] \cdot \sigma_0$ with $\sigma_0$ the median distance between the samples, and the outlier rejection (hierarchy level) $m$ are selected using the proposed CVM. The CMS is then applied to merge clusters into two classes if more than two clusters are retrieved. Cores having less than 1% samples are considered outliers.

The experiments are performed using a subset of the samples to accelerate the computations ($N = 1000$ samples) and the remaining samples are labeled based on their nearest neighbour clustered core sample. The averaged Cohen's $\kappa$, false alarm rate (FAR) and missed alarm rate (MAR) are reported for 10 random runs and the three best runs in order to emphasize the stability of the different algorithms.

Table 6.1: Mean (standard deviation) statistics over different realizations using 1000 random pixels from *Gloucester floods*.

| | $\kappa$ (10 runs) | FAR | MAR | $\kappa$ (3 best runs) | FAR | MAR |
|---|---|---|---|---|---|---|
| $k$-means | .556 (.01) | 11.8 (.8) | 17.3 (1.5) | .568 (.00) | 11.1 (.1) | 18 (.2) |
| kernel $k$-means | .602 (.02) | 9.68 (1.1) | 17.5 (1.2) | .616 (.02) | 8.90 (.8) | 18 (.8) |
| linkage (average) | .265 (.35) | 1.76 (2.3) | 72.2 (36) | .681 (.03) | 2.43 (2.5) | 58 (36) |
| linkage (complete) | .468 (.34) | 14.1 (13) | 34.0 (31) | .675 (.03) | 3.36 (3.0) | 54 (40) |
| SVC merging A | .553 (.33) | 5.96 (13) | 44.3 (25) | .720 (.00) | 5.68 (13) | 45 (24) |
| SVC merging B | .306 (.39) | 18.7 (20) | 52.3 (32) | .717 (.00) | 2.07 (.8) | 32 (3.1) |
| NSVC merging A | .628 (.26) | 5.07 (11) | 39.5 (16) | .728 (.01) | 5.12 (11) | 39 (17) |
| NSVC merging B | .496 (.24) | 11.9 (10) | 33.1 (19) | .693 (.04) | 3.64 (3.3) | 28 (7.5) |
| HGPC merging A | .643 (.24) | 1.48 (.6) | 4.7 (21) | .733 (.01) | 1.55 (.6) | 40 (21) |
| HGPC merging B | .572 (.22) | 5.05 (3.9) | 36.1 (24) | .720 (.02) | 3.46 (2.4) | 38 (22) |

### 6.5.2 Results

The results are presented in Tables 6.1 and 6.2. Averaged change maps over the three best runs can be visualized in Figure 6.7.

Gloucester floods (see Table 6.1): The $k$-means and its non-linear version never capture properly the data structure due to the significant amount of mixed samples behaving as outliers. The linkage algorithms are not robust to outliers and show to be very dependent on the sub-sampled set of pixels. The SVC, NSVC and HGPC have the outlier rejection ability, which accommodates better the data structure. For all of them, CMS A is more appropriate than CMS B. Probably, because the unchanged and the changed ensembles of clusters contain both cores of various sizes. The SVC, which is not truly hierarchical, is less stable through the ten different random runs and has higher missed detection (MAR) than its nested counterpart (NSVC) and the HGPC. The HGPC provides the best results in terms of $\kappa$ as well as the lowest false detection of changes (FAR). The latter is more important since some of the ambiguous changes labeled changed are missed by all the algorithms and therefore increases the MAR, as it can be observed in the change maps in Figure 6.7.

Bastrop fires (see Table 6.2): The $k$-means and the linkage algorithms fail to accommodate the important number of mixed samples (as in the previous case of the Gloucester floods). They provided very unbalanced solutions with the group of changes being only a tiny region of a certain type of changes and all the rest being grouped as unchanged. The kernel $k$-means is better in terms of $\kappa$, but produces a high number of false detection (FAR) compared to the others. It does not separate the two groups well, due to the mixed pixels considered as outliers. The two merging systems give nearly similar results for SVC and equivalent results for NSVC since only two clusters were obtained requiring no merging. Both give the best results for the ten random runs and the three best runs, respectively. The HGPC is less stable and only captures the data structure for the three best runs using the merging system B, penalizing different core sizes.

Table 6.2: Mean (standard deviation) statistics over different realizations using 1000 random pixels from *Bastrop fires*.

| | $\kappa$ (10 runs) | FAR | MAR | $\kappa$ (3 best runs) | FAR | MAR |
|---|---|---|---|---|---|---|
| $k$-means | .243 (.17) | 5.0 (14) | 76 (18) | .437 (.21) | .63 (.50) | 65 (19) |
| kernel $k$-means | .522 (.11) | 15 (6.5) | 24 (1.1) | .658 (.05) | 7.15 (2.5) | 25 (.5) |
| linkage (average) | .081 (.07) | .02 (.01) | 95 (4.8) | .159 (.01) | .03 (.01) | 90 (1.0) |
| linkage (complete) | .145 (.09) | .16 (.30) | 90 (6.5) | .240 (.08) | .49 (.50) | 83 (6.8) |
| SVC merging A | .699 (.17) | .07 (.04)) | 40 (18) | .821 (.09) | .1 (.03) | 26 (12) |
| SVC merging B | .691 (.16) | .06 (.02) | 41 (17) | .799 (.05) | .07 (.00) | 29 (6.7) |
| NSVC merging A | .467 (.37) | 3.9 (11) | 57 (30) | .831 (.08) | .11 (.05) | 24 (11) |
| NSVC merging B | .467 (.37) | 3.9 (11) | 57 (30) | .831 (.08) | .11 (.05) | 24 (11) |
| HGPC merging A | .261 (.20) | .04 (.02) | 81 (18) | .500 (.23) | .03 (.01) | 60 (23) |
| HGPC merging B | .408 (.29) | .04 (.02) | 67 (28) | .811 (.05) | .04 (.00) | 27 (6.4) |

In both experiments, we observed that the $k$-means and its non-linear version do not perform well in situations having an important number of mixed pixels, as reported in [Volpi et al., 2012], where standard clustering could be performed only after removing ambiguous samples located between unchanged and changed clusters. Both SVC methods provide more stable solutions able to handle these mixed pixels. The true hierarchy of NSVC is more coherent with the proposed CVM, however the nestedness constraint could regularize excessively the solutions and a larger number of outlier rejection levels $m$ should be taken to avoid such situations. The HGPC presents a nice flexibility for representing the clusters but has a tendency to form clusters of very different sizes, as it can be observed in Figure 6.2. It benefits from the second merging system proposed (CMS B) penalizing different core sizes to merge. The obtained change maps show very similar results for SVC, NSVC and HGPC. The SVC performs better in detecting the bottom-left patch of fire in the Bastrop fires images, whereas HGPC is missing it, as well as other changes.

### Computational costs

The computational costs of the CVM and CMS, as well as the clustering algorithms are presented in Table 6.3. The $k$-means and the linkage algorithms are the fastest and could be applied to the entire image. The kernel extension of $k$-means is heavier and already requires to be applied on a subset of the image. The SVC, NSVC and HGPC algorithms generate boundaries at $M = 10$ hierarchical levels for a fixed kernel bandwidth (boundaries) and build a 5-NN graph for accelerating the cluster assignment (clustering) from the obtained boundaries. We observe the gain obtained by using *entire solution path* algorithms for the SVC and NSVC. The HGPC has a larger cost for recovering the cluster from its adjacency matrix, while the SVC have larger cost for obtaining their cluster boundaries. Despite the number of distances computed for the CVM and CMS, the computational costs are of an order of magnitude lower than these algorithms. We still observe an important gap between these algorithms and the standard $k$-means or

Table 6.3: Computational times (seconds) for different steps in the clustering algorithms with a subset $N = 1000$ of *Gloucester floods* on an Intel Quad core Q9400 @2.66$Ghz$.

| | $k$-means | ker. $k$-means | linkage | SVC | SVC (path) | N-SVC | HGP |
|---|---|---|---|---|---|---|---|
| boundaries | × | × | × | 103.162 | 17.450 | 26.891 | 0.844 |
| clustering | 0.492 | 1.075 | 0.167 | 6.329 | 3.161 | 4.122 | 93.055 |
| CVM+CMS | × | × | × | 0.110 | 0.241 | 0.132 | 2.144 |
| total | 0.492 | 1.075 | 0.167 | 109.601 | 20.852 | 31.145 | 96.043 |

linkage algorithms. Notice that this is emphasized by the differences in implementation, the $k$-means and linkage using the optimized mex compiled version in matlab while the others are uncompiled matlab implementations.

## 6.6 Conclusions

In this chapter, we presented two clustering algorithms handling overlapping clusters and providing hierarchical sets of solutions. We proposed a cluster validity measure allowing to qualify the quality of the clustering solutions of these hierarchical clustering algorithms. It allows to perform parameter selection within a set of solutions at various levels of outlier rejection and with arbitrary shaped clusters. We also provided a merging mechanism to group clusters that overlaps since sharing a certain number of samples as outliers. The support vector clustering algorithms are improved in terms of computational costs, however they still remain the bottleneck compared to the cluster validity measure and merging system which are more efficiently computed. Experiments on unsupervised change detection in multi-temporal multispectral remote sensing imagery show the accuracy of the proposed methods and the benefit from their intrinsic regularization producing less false detections.

These approaches to unsupervised change detection could have a potential impact on early warning systems in order to provide change maps in an automatic way. However, some examples of unchanged regions might be also available in a timely manner after a natural disaster. This additional information could help better detecting the changes and is investigated in the next Chapter.

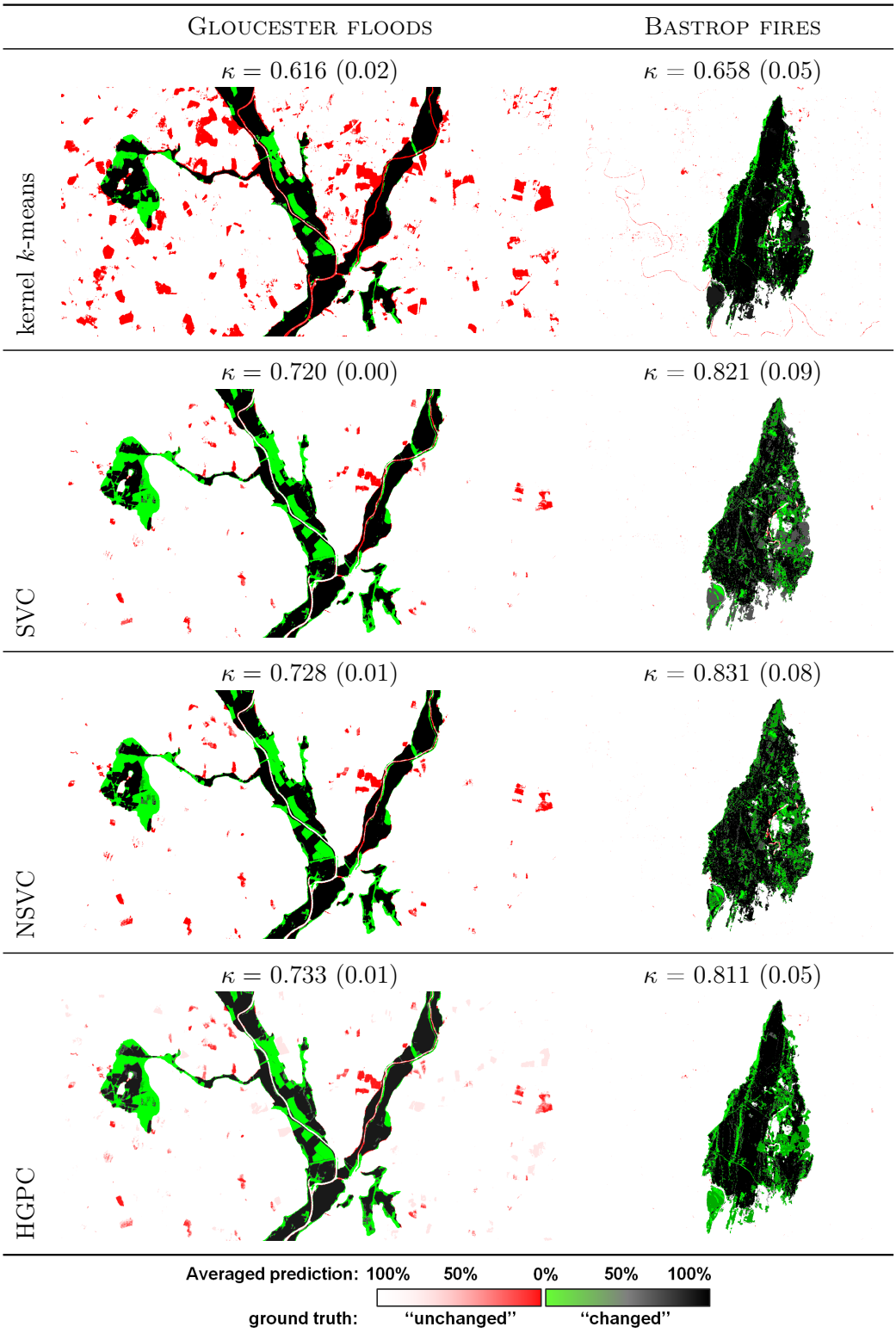



Figure 6.7: Averaged change maps for different realizations using 1000 random pixels. The CMS A is used except the HGPC in *Bastrop fires* using CMS B.

Color legend: white= 100% detected "unchanged", black=100% detected "changed", red=false detection, green=missed detection and gradient levels for percentages in between.

# 7 Semi-supervised novelty detection via SVM entire solution path

In this Chapter[1], an approach to semi-supervised novelty detection (SSND) that relies on an unbalanced classification between labeled and unlabeled samples is presented. The methodology based on the cost-sensitive SVM (CS-SVM) avoids the heavy supervised parameter selection thanks to entire solution paths algorithm, already presented in Section 5.3 and to a low density criterion for unsupervised parameter selection that we introduce in Section 7.3. The experiments on two multitemporal change detection datasets (flood and fire detection) are presented in Section 7.5 and finally discussed in Section 7.5.

## 7.1 Detecting unknown changes under limited supervision

In this Chapter, we investigate the problem of SSND as an unbalanced classification problem, as described in Section 7.2, and study the impact of using entire solution paths algorithms in this particular setting. Moreover, we propose a low density criterion for selecting the hyperparameters automatically.

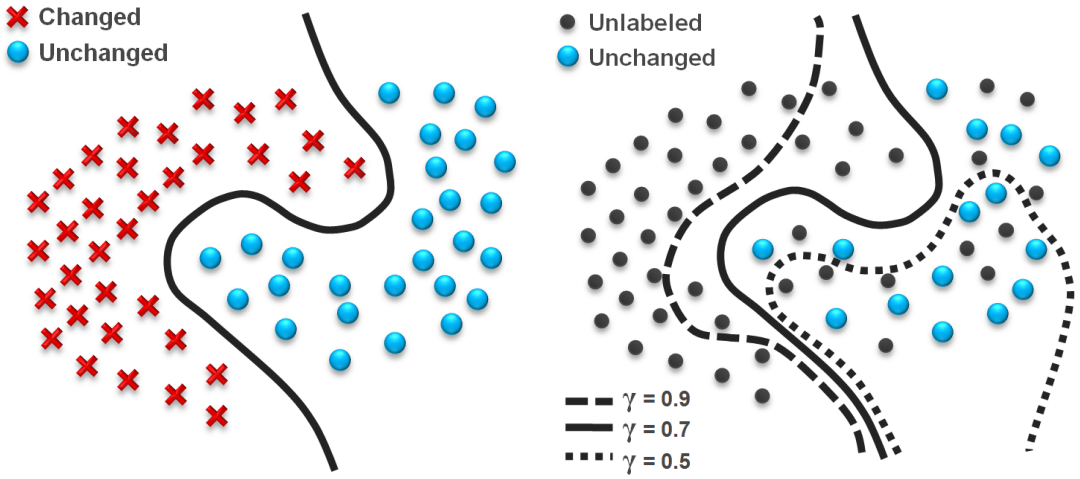
If information about the change is often difficult to obtain, information on the nature of the unchanged areas is easier to have beforehand. Considering these initial conditions (ignorance about the nature of the change and knowledge about some "unchanged" areas), the change detection problem can be reformulated as a novelty detection problem where only samples from "unchanged" areas are available to detect "changed" regions, referred to as *novelties*. Since labeling is extremely costly, samples of "unchanged" areas are assumed to be available only in a limited amount.

However, one can expect to have access to a large set of unlabeled samples at no extra costs, contrarily to labeled data requiring expert's labeling time or expensive ground surveys. Semi-supervised techniques, exploiting unlabeled data, have shown great improvements

[1]This Chapter is based on [de Morsier et al., 2012a, 2013c]

for classification methods under appropriate assumptions on the data distributions [Zhu and Goldberg, 2009]. In remote sensing data classification, semi-supervised learning has driven a strong current of research, where methods exploiting graphs on manifolds [Camps-Valls et al., 2007, Gómez-Chova et al., 2008], low density areas [Bruzzone et al., 2006, Bruzzone and Persello, 2009], clustering of data [Tuia and Camps-Valls, 2009, Gómez-Chova et al., 2010] have shown to be strongly beneficial for the classification performance. In [Bovolo et al., 2008], semi-supervised change detection is performed using the Semi-Supervised SVM (S3VM), which labels progressively the unlabeled samples from an initial classification requiring both "unchanged" and "changed" labels. A context-sensitive semi-supervised SVM has been proposed to be robust to mislabeled examples in the training set [Bruzzone and Persello, 2009] and a semi-supervised SVM using composite kernels based on clustering and mean map kernel allowed to overcome the problem of sample selection bias [Gómez-Chova et al., 2010]. Finally in [Capobianco et al., 2009], semi-supervised kernel orthogonal space projection is proposed to perform target detection without knowledge of the outlier class. In SSND, two approaches were presented and compared in [Muñoz-Marí et al., 2010]. The first is the standard CS-SVM with labeled and unlabeled data, while the second is the SVDD with a kernel distorted by a graph Laplacian built on the unlabeled samples (S2OCSVM). CS-SVM was very efficient in difficult change detection scenarios (e.g. cloud vs. snow). In [Li et al., 2011b], the authors proposed a SSND approach aiming at learning the conditional probabilities of the "unchanged" class by training on labeled "unchanged" and unlabeled examples. The retrieved probabilities were finally normalized by a constant factor. This approach showed promising results in remote sensing one-class classification under the assumption that labeled samples are selected completely at random. Note that in such a setting (which is also the one proposed in this Chapter), the natural unbalance of the novelty detection problem is further increased, since we consider labeled pixels, that are "unchanged", and confront them against unlabeled pixels, that are a mixture of "changed" and (mostly) "unchanged" areas. As a consequence, the unbalance between "changed " and "unchanged" regions is even stronger.

The SSND approaches considered in remote sensing so far show three main weaknesses: (i) they are all extremely time consuming, in particular when they reduce to an unbalanced two-classes classification problem which requires a heavy parameter selection to find the optimal cost asymmetry [Elkan and Noto, 2008, Muñoz-Marí et al., 2010]. In [Yao et al., 2009], the entire regularization path (along $\lambda$) of the standard SVM was assessed in SSND situations but without exploiting the cost asymmetry. (ii) Very often, the approaches are based on modeling the changes, thus requiring labeled "changed" samples that are difficult to obtain in sufficient quantity to be representative. (iii) Finally, even the approaches based only on "unchanged" regions (Novelty Detection approaches) select the optimal boundary and the free parameters through cross-validation, thus again using labels from both "unchanged" and "changed" classes.




(a) (b)

Figure 7.1: (a) Supervised two-class classification ("unchanged" vs. "changed") and (b) Semi-Supervised novelty detection by classifying labeled vs. unlabeled data with different cost asymmetries $\gamma$.

## 7.2 Semi-supervised novelty detection as a two-class problem

In this Chapter, the following principle is used: the positive samples are classified against all the unlabeled samples with different penalization on the respective errors [Liu et al., 2003]. The asymmetry in the classes cost allows to penalize more the errors done on the labeled samples and less those on the unlabeled samples (since they contain both "unchanged" and "changed" samples). For an appropriate (but unknown) asymmetric cost, the "unchanged" and "changed" are optimally classified. This approach reduces SSND to a binary unbalanced classification problem (labeled vs. unlabeled) and has been proven to be very effective and general, since no assumptions have to be done on the distributions and on the proportion of novelties [Blanchard et al., 2010]. This principle is illustrated in Figure 7.1 (b).

More formally, we define the SSND setting with $I_+ = \{i : y_i = +1\}$ the set of labeled "unchanged" samples and $I_- = \{i : y_i = -1\}$ the set of unlabeled samples containing a mixture of the two classes ("unchanged" and "changed"). The CS-SVM primal optimization problem is defined as

$$\begin{aligned} \min_{w,\xi} \quad & \frac{\lambda}{2}\|w\|^2 + \gamma \sum_{i \in I_+} \xi_i + (1-\gamma) \sum_{i \in I_-} \xi_i \\ \text{s.t.} \quad & y_i \langle w, \Phi(\mathbf{x}_i) \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \end{aligned} \tag{7.1}$$

where the cost asymmetry $\gamma$ controls the trade-off between penalizing the error on the

labeled and on the unlabeled samples.

We propose to tackle this asymmetric classification problem, where the labeled samples are classified against the unlabeled samples, using the entire solution path of the CS-SVM, which is the CS-SVM PATH presented in Section 5.3.2. This provides us with all possible cost asymmetries in a computationally efficient way. We can further impose a strict hierarchy in the set of solutions by imposing nested boundaries, which is the NCS-SVM presented in Section 5.4.2. In this Chapter, we aim at demonstrating that algorithms providing the entire solution path for CS-SVM result in faster and coherent classifiers for SSND tasks, which are complex unbalanced situations because of the important overlap between labeled and unlabeled samples.

From this set of solutions, only one penalizes correctly the labeled and unlabeled samples and retrieves the optimal boundary. In the following Section, we present a way of identifying this optimal solution among those in the entire path.

## 7.3 Unsupervised parameter selection based on the cluster assumption

In the SSND settings presented above, no assumptions are made on the distribution of the novelties. There is a trade-off between restricting to certain assumptions and performing the parameter selection: either no assumptions are made but then a parameter selection based on cross-validation becomes necessary (thus requiring labeled pixels corresponding to "changed" areas), or an assumption is made on the two distributions and the parameter selection can be done in an unsupervised way.

We propose an approach for selecting the optimal cost asymmetry and the other free parameters based on the *cluster assumption*, an extensively used assumption in semi-supervised learning [Chapelle et al., 2006, Zhu and Goldberg, 2009]. This assumption states that the two classes ("changed" and "unchanged") are clustered in the input space. Therefore the boundary of the optimal classifier should not pass through the clusters but in the region of low density between them. The Transductive SVM (TSVM) exploits the cluster assumption by iteratively labeling the unlabeled samples and retraining with the augmented labeled set of samples [Vapnik, 2000], while the Semi-Supervised SVM (S3VM) penalizes the unlabeled samples lying inside the margin of the SVM directly in its objective function [Bennett and Demiriz, 1999, Bovolo et al., 2008].

In our SSND context, unlabeled data are already involved in the training process to help obtaining a better discrimination between the "unchanged" and "changed" classes. The only remaining step is the selection of the optimal solution along the cost asymmetry path. This selection is usually done by cross-validation, training on a subset of the data and testing on a separate subset. It is however impossible to get a reliable accuracy since

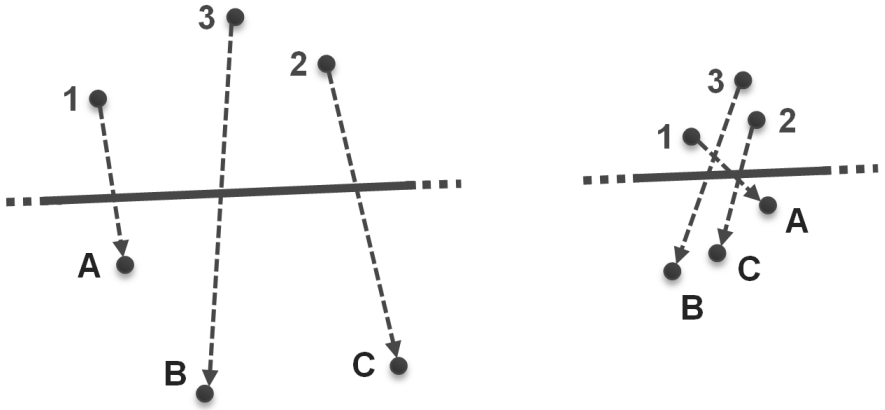



Figure 7.2: Low-density criterion principle for $k = 3$ for low density (left) and high density (right). 3 unique pairs of samples with minimum distance across the boundary are formed. First the closest sample (1) is paired with its closest sample on the other side (A). Then sample (2) is linked with (C) since the closest sample (A) is already linked. Finally sample (3) is paired with (B) the closest remaining sample.

no labels are available for the "changed" areas. To overcome this issue, the boundary selection can be performed based on the local density around the boundary. The distances between the closest samples across the boundary are inversely related to the density in these regions: the larger the distances, the lower the density. The other parameters influencing the boundary (kernel and regularization parameters) can be selected in the same way by maximizing the distances across the boundary.

Let us define $H_{O+}$ and $H_{O-}$, the sets of samples ordered by their distance to the boundary ($|f_{\gamma,\lambda}(\mathbf{x}_i)|$) on the positive and negative side respectively. A set of $k$ unique pairs of samples are built across the boundary (each sample is linked only once). The first sample of $H_{O+}$ is linked with the sample at minimum distance from $H_{O-}$. The second closest sample from $H_{O+}$ is linked with the closest of the $k-1$ remaining samples on the other side of the boundary. This linkage goes on until the $k^{th}$ sample has been linked. This principle is illustrated in Figure 7.2. The procedure is then repeated considering the other side of the boundary, and finally provides $2 \cdot k$ pairs of samples across the boundary. The set of Euclidean distances between the pairs of samples is denoted by $D_{pair}(k, \gamma)$.

The estimated average distance across the boundary is obtained using the median of the paired distances to avoid possible biases related to isolated samples linked with remaining distant samples. The density criterion across the boundary is defined as $DC(k, \gamma) = median\{D_{pair}(k, \gamma)\}$. A small $DC$ indicates that boundary passes through high density regions, while a large $DC$ indicates that boundary passes through low density region. The maximum $DC$ value indicates the optimal boundary passing through lowest density regions . For this reason, the final criterion LDC is defined as $LDC(k) = \max_\gamma(DC(k, \gamma))$.

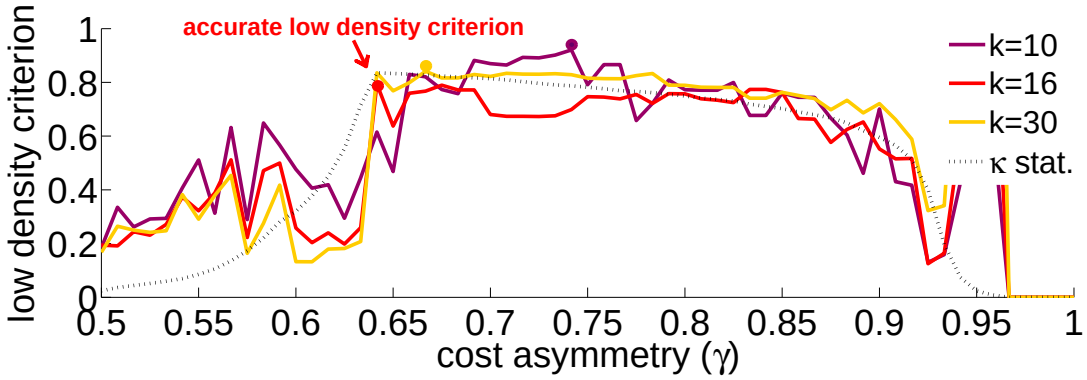



Figure 7.3: Low density criterions along the cost asymmetries $\gamma$ for different number of pairs of samples across the boundary ($k = 10, 16, 30$). The minimal maximum peak indicates the appropriate $k$ for the low density criterion (here $k = 16$), since $k = 30$ adds pairs of samples far from the boundary and $k = 10$ misses samples along the boundary. The low density criterions have been rescaled for visualization purposes with the overlayed $\kappa$ statistic. The maximum $\kappa$ is obtained for $\gamma \approx 0.64$, matching with the low density criterion peak at $k = 16$.

Even if a robust average is obtained using the median, the choice of the parameter $k$ influences the extent of the density measure (from local to global for increasing $k$). A too large $k$ could add pairs of samples actually far from the boundary (inside the clusters), whereas a too small $k$ could miss pairs close to the boundary. Both situations could result into an overestimated median distance across the boundary. To avoid such situations, the most accurate distance across the boundary is searched through a range of $k$ values ($k_{range}$), see Figure 7.3. The value minimizing LDC over $k_{range}$ is retained as the most robust value, $k^*$. Using this optimal $k^*$, the boundary passing through the lowest density region is localized at $\gamma^* = \arg\max_\gamma(DC(k^*, \gamma))$. The full procedure is summarized in Algorithm 2.

## 7.4 Computational complexity

Solving the QP problem for the standard SVM requires approximately $\mathcal{O}(s^2N + s^3)$ with $s$ the number of samples inside the margin (e.g. support vectors). The complexity is dependent on the regularization parameter $\lambda$ and end up between $\mathcal{O}(N^2)$ and $\mathcal{O}(N^3)$ for high and low regularization respectively [Bordes et al., 2005]. Therefore, if $P$ cost asymmetries $\gamma$ are considered along the path, the maximal computational complexity would be $\mathcal{O}(PN^3)$.

The CS-SVM PATH algorithm has a complexity of $\mathcal{O}(m^2N + N^2m)$ for an entire solution path, with $m$ the maximum number of samples on the margin $\mathcal{M}$ along the path [Hastie et al., 2004]. The maximum value for $m$ is $N$ leading to a computational cost equivalent to two standard SVM. Therefore the CS-SVM PATH algorithm has already a lower

**Algorithm 2** Low density boundary selection

**Input:** (N)CS-SVM solutions at $M$ cost asymmetries $\gamma_m$
**for each** $k \in k_{range}$ **do**
  **for each** $\gamma \in \gamma_m$, $m = 1, ..., M$ **do**
    Get $H_{O+} = sort(f_{\gamma,\lambda}(\mathbf{x}_i))$ with $\{i, f_{\gamma,\lambda}(\mathbf{x}_i) > 0\}$
    Get $H_{O-} = sort(|f_{\gamma,\lambda}(\mathbf{x}_i)|)$ with $\{i, f_{\gamma,\lambda}(\mathbf{x}_i) < 0\}$
    Build unique pairs of samples (from $H_{O+}$ and $H_{O-}$)
    $D_{pair}(k, \gamma) \leftarrow$ set of distances between the pairs
    $DC(k, \gamma) \leftarrow median\{D_{pair}(k, \gamma)\}$ {median distance across boundary}
  **end for**
  $LDC(k) \leftarrow \max_\gamma(DC(k, \gamma))$ {low density criterion}
**end for**
$k^* = \arg\min_k(LDC(k))$ {optimal $k$ not overestimating}
**Output:** $\gamma^* \leftarrow \arg\max_\gamma(DC(k^*, \gamma))$ {optimal boundary}

complexity than the standard CS-SVM when more than two cost asymmetries are solved.

The NCS-SVM has a complexity linked to the number of iterations required to converge which is proportional to the number of samples $\mathcal{O}(N)$. The size of the QP subproblem solved at each iteration is proportional to the number of breakpoints $\mathcal{O}(M^2)$. Moreover checking the KKT conditions requires $\mathcal{O}(NM)$, leading to a total complexity of approximately: $\mathcal{O}(M^2N + N^2M)$. Experimentally, $M \approx 10$ and the iterations required are $\approx 5 \times N$ for the NCS-SVM [Lee and Scott, 2010]. Thus the NCS-SVM will have a lower computational cost than the standard CS-SVM and CS-SVM PATH algorithms for datasets larger than $\approx 50$ samples.

## 7.5 Experiments

This Section presents experiments on two multitemporal change detection datasets (flood and fire detection) and discuss the results obtained with the different methods presented. The two multitemporal datasets used to compare and validate the proposed approaches are A. *Gloucester floods*: SPOT images (3 bands) before and after a flood near Gloucester in the U.K. B. *Bastrop fires* 2010/11: Landsat images (6 bands) before and after a forest fire in Texas, USA.

The *Bastrop fires* 2010/11 dataset is particularly challenging, since the unchanged areas show radiometric differences related to seasonal vegetation changes. Therefore, without any possibilities of matching the two images the traditional unsupervised methods based on image differencing are exposed to return a large amount of false alarms.

### 7.5.1 Experimental setup

The results for the CS-SVM PATH and NCS-SVM methods are compared with the standard CS-SVM, the semi-supervised and fully-supervised SVDD and the unsupervised CVA [Chen et al., 2010]. Moreover, the proposed methods are evaluated both using cross-validation (CV) and low density (LD) parameter selection criteria.

In a change detection scenario, either the difference or ratio image are often used [Camps-Valls et al., 2006]. Here we consider the difference of the spectral bands of the two images. Another set of features more physically inspired is a stack of the NDVI obtained separately at each acquisition. It is important in such scenarios to choose the set of features, which is best adapted to the situation [Singh, 1989]. The floods over non-urban areas impact the vegetation which is reflected by a change of the response of flooded vegetation, which is better visible in the NDVI than in the difference or stacked images. For this reason, the stacked NDVI will be used with the *Gloucester floods* dataset. For the Bastrop dataset, the NDVI features could be ambiguous for regions having dried grass or cut crops. This type of land, being well spread among the unburnt areas, would avoid detecting burnt areas. Therefore the difference image is used in the *Bastrop fires* experiments. The datasets have been centered (zero-mean) and normalized by their standard deviation on each feature independently.

Ground truth have been established for the whole images both by visual interpretation and using documents assessing the range of damages. Some small ambiguous zones ($\approx 3\%$) have been left aside from the *Bastrop fires* ground truth. The training set is composed of $N_{labeled}$ samples randomly selected from the "unchanged" class and of $N_{unlabeled}$ samples randomly selected among all the remaining samples. The number of labeled and unlabeled samples are varied in the range: $[50, 100, 200, 300, 400, 500]$.

The validation set, composed of 10000 samples, allows to select the parameters $\sigma$ (bandwidth of RBF kernel) and $\lambda$ (regularization) returning the highest averaged Cohen's kappa statistic ($\kappa$) [Foody and Mathur, 2004] (CV hereafter). Our proposed alternative to cross-validation selects the optimal parameters by finding the maximum low density criterion (LD hereafter). The average over three different random training and validation sets is used for both CV and LD parameter selection. An initial guess on the bandwidth of the RBF kernel $\sigma_0$ is obtained from the median standard deviation among 1000 random "unchanged" samples. The search is then realized over 15 values: $\sigma = [0.1 \times \sigma_0, ..., 1.5 \times \sigma_0]$. The regularization parameter $\lambda$, which balances the importance between maximum margin and error penalization, is searched among different values lower than the maximum regularization parameter $\lambda_{max}$: $\lambda = [0.01 \times \lambda_{max}, 0.1 \times \lambda_{max}, \lambda_{max}]$.

The Cohen's $\kappa$, overall accuracy, F1-score [Meng and Kerekes, 2012], False alarm rate (False changes detected/total "unchanged") and Missed alarm rate (Missed changes/total "changed") are assessed using the best parameters ($\sigma$,$\lambda$,$\gamma^*$,$k^*$). Ten independent training

Table 7.1: Mean (standard deviation) statistics over 10 different realizations using 500 pixels for each class.

| Methods | | $\kappa$ | $\kappa$ diff. | OA | F1-score | FAR | MAR |
|---|---|---|---|---|---|---|---|
| Gloucester floods | | | | | | | |
| CVA | | .407 (—) | -.414 | 80.18 (—) | 0.51 (—) | 20.0 (—) | 18.35 (—) |
| $S^3$VDD | | .725 (.03) | -.096 | 94.01 (0.58) | 0.76 (0.03) | 3.25 (1.03) | 24.85 (7.34) |
| SVDD | | .738 (.03) | -.083 | 94.40 (1.15) | 0.88 (0.03) | 4.09 (1.97) | 17.58 (6.50) |
| CS-SVM | | .821 (.02) | / | 96.58 (0.42) | 0.84 (0.02) | 1.28 (0.51) | 20.27 (3.89) |
| CS-SVM | CV | .808 (.03) | -.013 | 95.98 (0.61) | 0.83 (0.02) | 1.37 (0.64) | 22.30 (1.54) |
| PATH | LD | .752 (.08) | -.069 | 94.44 (2.99) | 0.78 (0.07) | 2.83 (4.35) | 24.40 (10.0) |
| NCS-SVM | CV | .819 (.02) | **-.002** | 96.21 (0.36) | 0.84 (0.02) | 1.30 (0.31) | 20.99 (1.81) |
| | LD | .791 (.04) | -.030 | 95.86 (0.65) | 0.81 (0.04) | 0.76 (0.40) | 27.55 (7.02) |
| Bastrop fires | | | | | | | |
| CVA | | -.023 (—) | -.897 | 38.44 (—) | 0.23 (—) | 65.50 (—) | 39.80 (—) |
| $S^3$VDD | | .620 (.07) | -.263 | 90.59 (2.46) | 0.67 (0.06) | 6.82 (3.01) | 26.55 (7.80) |
| SVDD | | .819 (.04) | -.064 | 95.785 (1.13) | 0.84 (0.04) | 2.77 (1.26) | 13.85 (4.35) |
| CS-SVM | | .883 (.02) | / | 97.412 (0.50) | 0.90 (0.02) | 1.13 (0.257) | 12.30 (3.35) |
| CS-SVM | CV | .897 (.02) | **+.017** | 97.342 (0.39) | 0.94 (0.01) | 1.44 (0.336) | 9.41 (1.96) |
| PATH | LD | .803 (.10) | -.080 | 95.26 (1.79) | 0.87 (0.11) | 1.91 (0.85) | 20.36 (14.9) |
| NCS-SVM | CV | .870 (.02) | -.013 | 96.69 (0.50) | 0.89 (0.02) | 1.70 (0.48) | 12.31 (2.82) |
| | LD | .812 (.07) | -.071 | 95.59 (1.38) | 0.84 (0.07) | 1.06 (0.65) | 23.06 (11.5) |

and validation sets are used to validate the approach on the test set composed of the remaining samples in the image.

For the NCS-SVM, $M = 7$ breakpoints are preselected equally spaced in the range $\gamma_m = [0.5; 1]$, then the solutions are interpolated by a factor of 10 resulting in $7+6\times9 = 61$ solutions on half of the cost asymmetry path.

The different number $k$ of sample pairs across the boundary for the low density criterion are: $k_{range} = [10, 11, ..., 40]$.

To ensure fair comparison, the standard CS-SVM is trained and evaluated on the same cost asymmetries than the NCS-SVM interpolated breakpoints ($P = 61$). The semi-supervised SVDD ($S^3$VDD), described in Section 4.4.3, uses a 5-NN graph on the samples and finds the optimal kernel distortion parameter $\varepsilon$ via cross-validation. A supervised SVDD is also trained with the same training set than the other methods but using all the labels of the samples considered unlabeled by the other approaches ("changed" samples considered as outliers during training [Li et al., 2011b]). This fully supervised classifier allows comparison with the best classification achievable using standard novelty detection classifier[2].

[2]Experiments were designed in the MATLAB environment based on the standard matlab optimization toolbox, the SVMPath and Nested SVM toolboxes (downloadable at http://www.eecs.umich.edu/∼cscott/code.html)

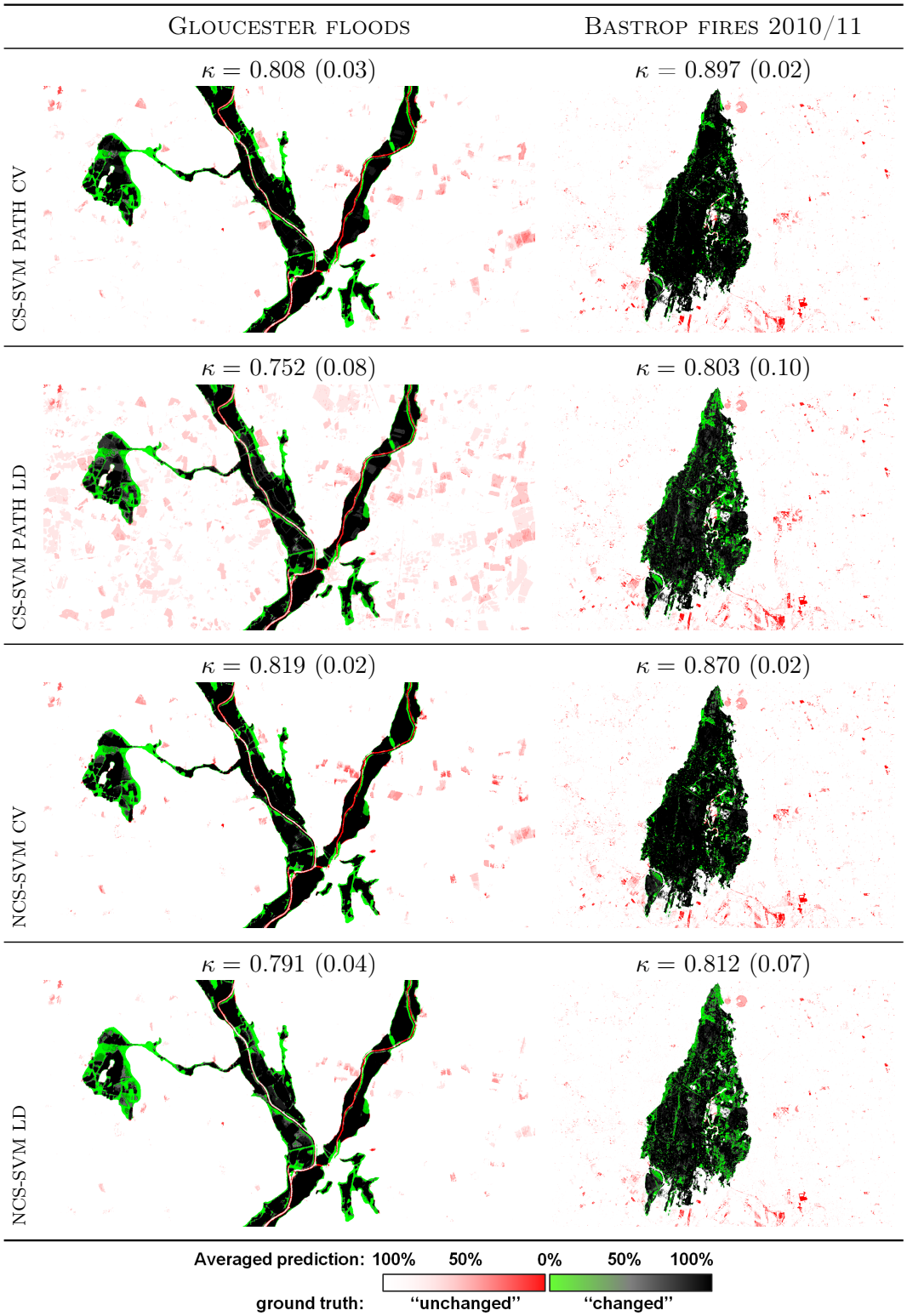



Figure 7.4: Mean (standard deviation) statistics over 10 different realizations using 500 labeled and 500 unlabeled pixels.

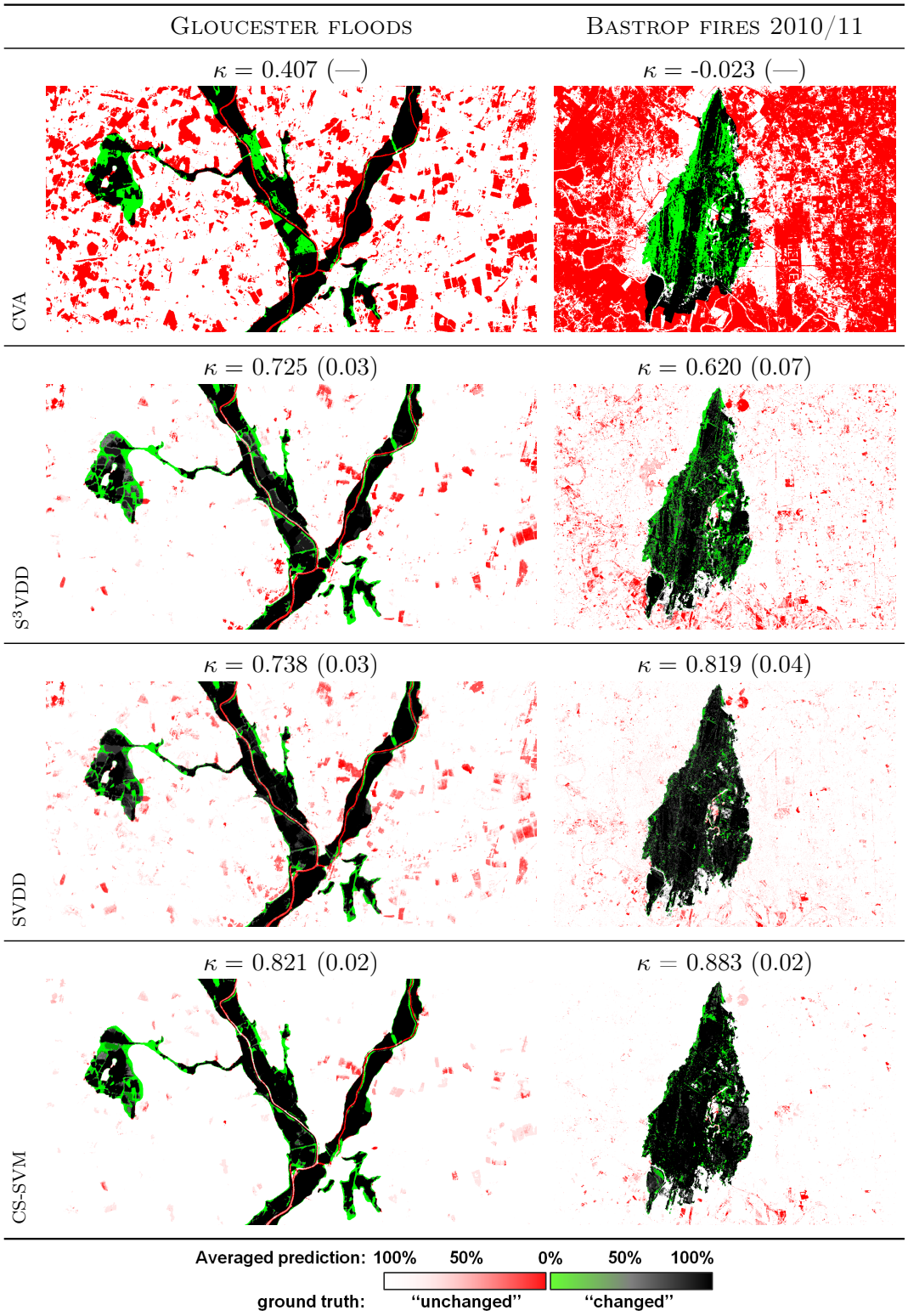
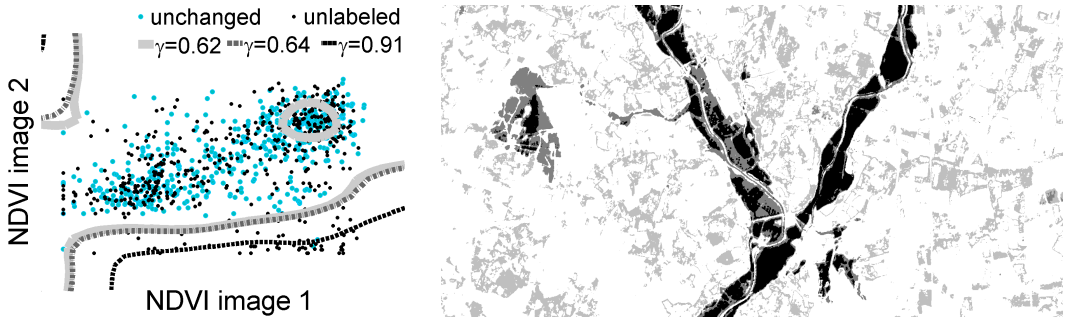



Figure 7.5: Mean (standard deviation) statistics over 10 different realizations using 500 labeled and 500 unlabeled pixels.




(a) (b)

Figure 7.6: (a) 2D plot of 3 nested boundaries along the path and (b) the overlay of the 3 corresponding detection maps for the *Gloucester floods (NDVI)* feature set. In (a) a "hole" can be visualized in the boundary at $\gamma = 0.62$.

### 7.5.2 Numerical results

Table 7.1 summarizes the different statistics for the best parameters averaged over 10 independent experiments (common to all the methods, except the CVA which uses the whole image directly). Figures 7.4 and 7.5 present the averaged change detection maps for the different methods. For CS-SVM PATH and NCS-SVM we report the results obtained by the two strategies for parameters optimization: cross-validation (CV) and the proposed unsupervised low density criterion (LD).

In general, the SSND methods considered provide accurate detection maps in both datasets except for the $S^3$VDD which performs worse than the supervised SVDD and the different CS-SVM. This is not so surprising since it exploits the unlabeled data for retrieving the data structure, while our proposed approaches exploit the unlabeled data in a discriminative way. Therefore, with a relatively small number of unlabeled samples the discriminative approaches take over the graph-based regularization approaches.

The SVDD, which is the only fully supervised classifier, provides results with more false detections compared with the standard CS-SVM (see Table 7.1, False Alarm Rate (FAR)). This can be due to the trade-off for the sphere between fitting precisely the distribution (requiring a small $\sigma$) and covering an important volume of the feature space (requiring a large $\sigma$). Here the sphere is small enough to recover most of the "changed" regions at the price of many other false detection outside the sphere. The SSND approach using the standard CS-SVM is more discriminative and thus performs better than the SVDD in these two change detection scenarios. The best performances of the CS-SVM PATH and NCS-SVM algorithms using CV parameter selection are equivalent to those obtained by the standard CS-SVM, confirming the efficiency of solution path algorithms in SSND situations. Note that the NCS-SVM, with its interpolated solutions can reach the same accuracy than the CS-SVM PATH.

Concerning the proposed low density criterion (LD), we observe that it provides results not far from those obtained by cross-validation (CV). The NCS-SVM LD solutions have slightly less false alarms but more missed alarms, which results in $\kappa$ values $\approx 0.02 - 0.05$ lower than those obtained with CV. This demonstrates that the *cluster assumption* holds in these scenarios and that a low density separation exists between the classes. The LD criterion is less stable than the CV resulting in higher $\kappa$ standard deviation. As it can be observed in Figure 7.3, the criterion along the cost asymmetries $\gamma$ is not smooth, which could result in selecting a suboptimal boundary.

The NCS-SVM LD gives better results than the CS-SVM PATH LD and is more stable. We observed that for cost asymmetries close to 0.5, the nesting of the boundaries could induce boundaries with holes in the middle of the distribution of the "unchanged" class. This phenomenon is illustrated in Figure 7.6 (a), representing the nested boundaries at different cost asymmetries. Figure 7.6 (b) illustrates the corresponding nested detection maps. These boundaries with holes in the middle of the distribution have the advantage of being difficultly selected by the low density criterion (since part of the boundaries are in the high density region inside the "unchanged" cluster). Therefore nested boundaries resulting in false detections are less likely to be selected than un-nested ones.

The sensitivity of the results with respect to the training set is reflected by the standard deviation of the $\kappa$ statistics reported in Table 7.1. The random selection of the training samples can induce a different interpretation of the changes to be detected: if a particular land cover is not present among the labeled "unchanged" samples but it is present in the unlabeled data, it will be interpreted as changed (e.g. the circular field at the top right of the *Bastrop fires*). It is important to remind that the maximum training set size considered ($N_{labeled} = N_{unlabeled} = 500$) corresponds to $\approx 0.1\%$ of the available image pixels. Experiments with a higher number of training samples ($N_{labeled} = N_{unlabeled}$ =1000, 2000, 5000 samples) for the NCS-SVM resulted in $\kappa$ improvement of 0.03 and a false alarm rate slightly lower (under 1%) both for CV and LD. Despite the increase in stability, the missed alarm rate did not change. Visual inspection of the detection maps for 5000 samples in each class (not reported) shows that most of the false detection regions have disappeared where the few remaining could easily be filtered out and that the few regions still miss detected are actually "unchanged" areas mislabeled.

Finally, we compared with the state-of-the-art CVA. For the two case studies, the CVA solution has an important false detection rate, particularly for the Bastrop case study. The method, being completely unsupervised, is unable to properly target the changes of interest in presence of radiometric differences between the images or seasonal changes. In *Gloucester floods*, this corresponds to the presence of wet areas and clouds in the first acquisition, which become challenging "unchanged" regions. For *Bastrop*, the important seasonal changes in the vegetation make the "unchanged" pixels nonzero in the magnitude vector, thus preventing CVA to function correctly. This demonstrates the benefit of SSND, which allows to define "unchanged" areas even if they show radiometric differences.

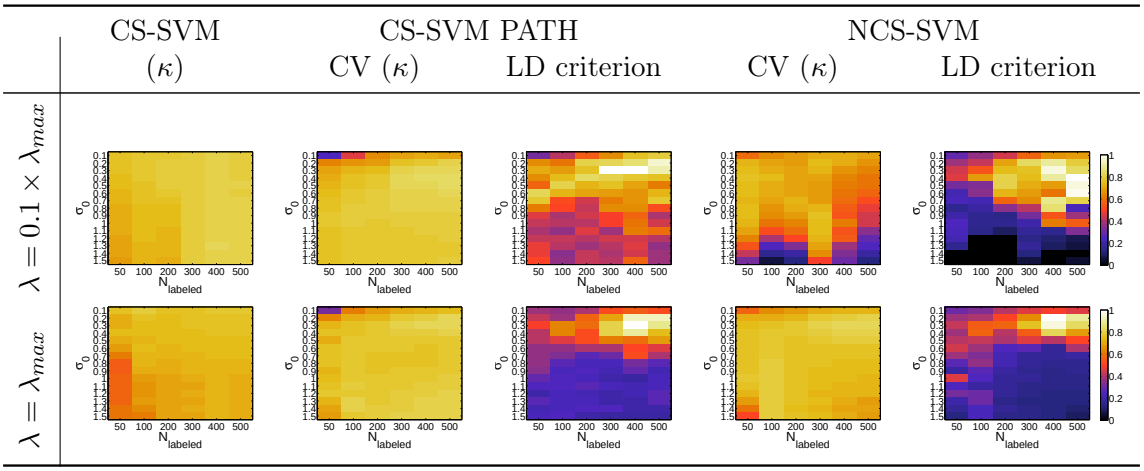


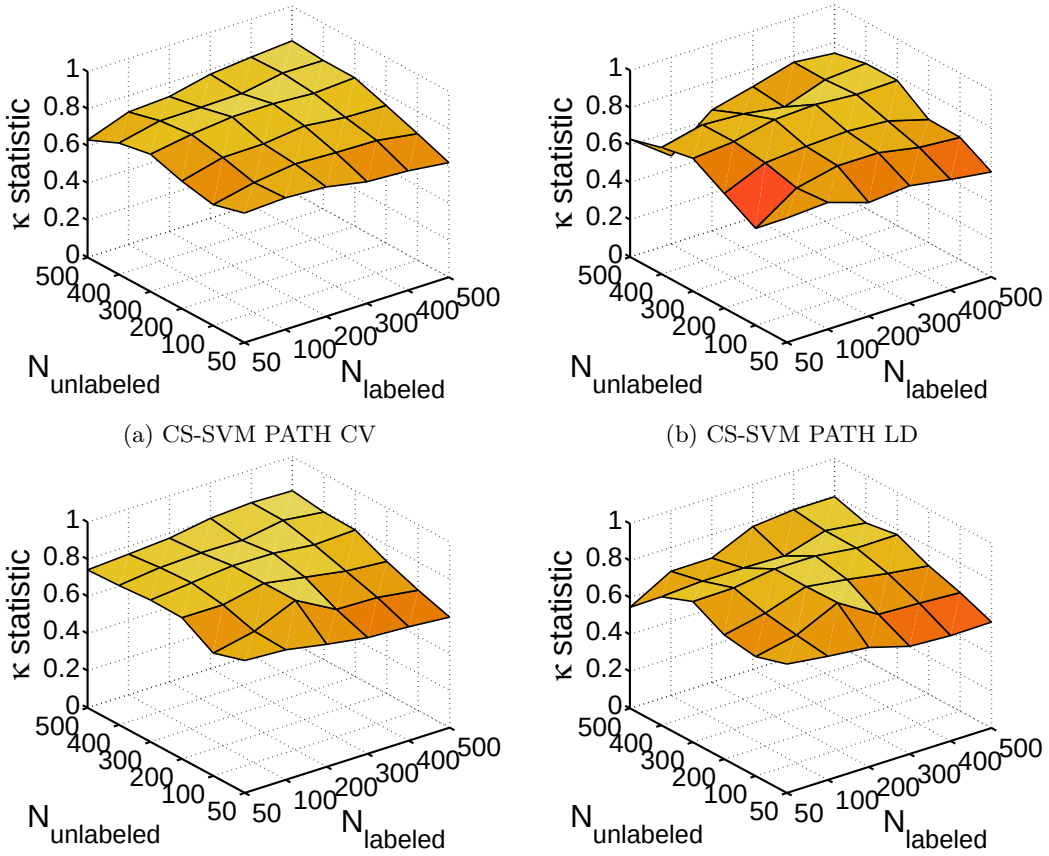

Figure 7.7: Averaged $\kappa$ statistic (CV) and low density criterion normalized (LD) over 3 random runs for the *Gloucester floods* dataset with 500 unlabeled samples and different number of labeled samples $N_{labeled}$, regularization parameters $\lambda$ and kernel parameters $\sigma$ (in terms of $\sigma_0$).

### 7.5.3 Free parameters sensitivity

Figure 7.7 illustrates the optimization surface for the range of $\lambda$ and $\sigma$ parameters, as well as for the number of labeled "unchanged" pixels ($N_{labeled}$). For the two algorithms tracing the entire path, larger regularization gives the best results and reduces the sensitivity to the parameters (to the level of the standard CS-SVM). The low density criterion changes rapidly with $\sigma$ and shows the best performance between 400 and 500 labeled samples. The matching between the maximum of the low density criterion (LD) and the maximum $\kappa$ from cross-validation (CV) can be well observed in Figure 7.7 (bottom row). Therefore, the parameters $\sigma$ and $\lambda$ selected by the unsupervised LD criterion are equivalent than using the supervised CV.

The dependence on the number of labeled and unlabeled samples is further studied in Figure 7.8. NCS-SVM is slightly less sensitive than CS-SVM PATH when few labeled samples are used.

The low density criterion is in comparison more sensitive to the number of unlabeled and labeled samples. A certain amount of unlabeled data is required to observe a cluster of pixels belonging to the "changed" regions. Moreover, the number of labeled samples should be large enough to properly cover the unlabeled samples from the "unchanged" distribution. NCS-SVM being more robust than CS-SVM PATH, it provides more accurate solutions in very unbalanced situations (i.e. $N_{labeled} = 50$ and $N_{unlabeled} = 500$). Experiments with a fixed number of labeled samples ($N_{labeled} = 500$) and a larger number of unlabeled samples ($N_{unlabeled} = 1000, 2000, 5000$) for the NCS-SVM did improve the stability but not significantly the accuracy. The approach is discriminative enough to ensure an accurate detection with a small amount of unlabeled samples. Adding more




(a) CS-SVM PATH CV

(b) CS-SVM PATH LD

(c) NCS-SVM CV

(d) NCS-SVM LD

Figure 7.8: Averaged $\kappa$ over three random runs as a function of the number of labeled and unlabeled data for the *Gloucester floods* using: (a) CS-SVM PATH cross-validation, (b) CS-SVM PATH low-density, (c) NCS-SVM cross-validation and (d) NCS-SVM low-density.

| | | | | CS-SVM Path | | NCS-SVM | |
|---|---|---|---|---|---|---|---|
| | | SVDD | CS-SVM | CV | LD | CV | LD |
| $N = 100$ | Train (s) | 11.2 (2.0) | 11.2 (2.0) | 0.7 (0.3) | | 9.9 (6.9) | |
| | Test (s) | 19.3 (0.3) | 19.2 (0.3) | 14.1 (1.5) | 1.2 (0.5) | 22.5 (1.5) | 0.66 (0.03) |
| | Total (s) | 30.5 (2.3) | 30.4 (2.3) | 14.8 (1.8) | **1.9 (0.8)** | 32.4 (8.4) | 10.6 (6.9) |
| $N = 500$ | Train (s) | 40000 (17125) | 1505 (700) | 36.4 (4.6) | | 95 (51) | |
| | Test (s) | 89.9 (0.6) | 89.5 (0.9) | 82.3 (1.6) | 31.4 (5.5) | 75 (2.9) | 1.3 (0.06) |
| | Total (s) | 40090 (17126) | 1595 (702) | 118 (6) | **67 (10)** | 170 (54) | 96 (51) |

Table 7.2: Runtimes for the different methods with different training set size ($N = N_{labeled} = N_{unlabeled}$) for *Gloucester floods*.

unlabeled samples allow a better discrimination of the boundary passing through low density regions and enforces the unbalanced situation between labeled and unlabeled samples. This is not to be seen as a drawback, since by doing so the proportions of the two classes "unchanged" and "changed" are converging towards their true proportions in the dataset.

### 7.5.4 Algorithms runtime and convergence analysis

In the experiments, the two algorithms tracing the entire solution path converged to accurate solutions with low runtimes. In this section we provide further observations on the training and testing runtimes and convergence of the algorithms.

Table 7.2 reports the training and testing runtimes for different size of the training set. The solution path algorithms are effectively faster for training and grow linearly compared to the standard CS-SVM and SVDD (exploiting both the standard *quadprog* Matlab routine for fair comparisons). The iterative procedure of the NCS-SVM is arbitrarily limited to a maximum of 5000 iterations (corresponding to $5 \times \max(N_{labeled} + N_{unlabeled})$, which is a trade-off between computational cost and accuracy). When setting a too large $\sigma$ or a too small $\lambda$ the breakpoints close to $\gamma = 0.5$ can be difficult to obtain: in this case, the algorithm may not converge in the maximum number of iterations allowed.

The testing runtimes are equivalent for the cross-validation (CV) through the different mehtods but much lower using the low density criterion (LD). The CS-SVM PATH LD is slower than the NCS-SVM LD for testing since it evaluates a larger number of breakpoints $\gamma$ (several times the number of training samples) compared to the 61 interpolated breakpoints of the NCS-SVM. The convergence of the CS-SVM PATH algorithm is guaranteed by the initialization of the algorithm with an arbitrary cost asymmetry in the range $[0.5; 1]$.

## 7.6 Conclusions

In this Chapter, we presented and evaluated two methods for Semi-Supervised Novelty Detection (SSND), that reduce to an unbalanced binary classification between labeled and unlabeled samples.The methods, based on the Cost-Sensitive Support Vector Machine (CS-SVM), assign different error costs for the two classes (cost asymmetry). The errors done on the unlabeled samples containing both classes samples are less penalized than those committed on labeled "unchanged" samples. The novelty of the proposed methods reside in the retrieval of the solutions for different cost asymmetries in a single optimization and the addition of a constraint imposing nested boundaries. This provides us with a

certain hierarchy and stability across the solutions.

We also proposed a low density criterion, which allows to select the optimal cost asymmetry and the other free parameters (kernel and regularization) in an unsupervised way. Such criterion estimates the local density along the boundary, based on pairwise distances across the boundary.

The results on the two multitemporal change detection scenarios (flood and fire detection) showed the efficiency of these SSND approaches that only exploit "unchanged" information and unlabeled data. Thanks to the discriminative use of the unlabeled data, the performances are already excellent with a small number ($< 500$) of labeled and unlabeled samples. The two algorithms deriving the entire solution path performed better than the semi-supervised and supervised SVDD and generally equivalently to the CS-SVM (but at a much lower computational cost). The nested CS-SVM also has the advantage of being less sensitive to the choice of parameters and the size of the training set. Using the low density criterion usually decreases the false alarm rate and slightly the global performance with respect to cross-validation ($\kappa \approx 0,02 - 0.05$ lower). However, we remind that, contrarily to cross-validation, no labeled information about the change is used, and that this is the price to pay to maintain the optimization unsupervised. Nonetheless, the good results obtained confirmed that the two classes are separated by a low density region.

These approaches could benefit from considering active learning strategies [Tuia et al., 2011]. These methods would help in the selection of unlabeled examples for semi-supervised learning allowing to find the most informative pixels for discriminating changed from unchanged areas [Li et al., 2010].

# Part III
# Novelty detection in manifolds

# 8 Background

In this Chapter, we introduce the assumption of high dimensional data lying on multiple manifolds which is exploited in this Part of the Thesis. We review some of the state-of-the-art algorithms in the context of clustering based on similar assumptions. A novel clustering algorithm based on a sparse and low-rank representation of the data in a non-linear feature space is presented in Chapter 9 for clustering and extended to semi-supervised one-class classification in Chapter 10.

## 8.1 High dimensional data on manifolds

High dimensional data can be found in time series, audio, image and video streams and are encountered in various areas ranging from computer vision to financial risk analysis. In most of these applications, the number of data samples available is too low with respect to the high data dimensionality. As presented in Section 3.2, the curse of dimensionality requires an exponential number of samples regarding the number of dimensions which will most of the time not be fulfilled. As a consequence, standard statistical tools tend to fail when applied to these challenging high dimensional situations. Fortunately, in many cases the complexity and effective dimensions of the data inherent structure tends to be much lower than the original high number of dimensions. This introduces the notion of data *manifolds*.

> *A manifold is a topological space where the data lies which is locally diffeomorphic to an Euclidean space. An example of manifold is the surface of a sphere in a 3D space.*

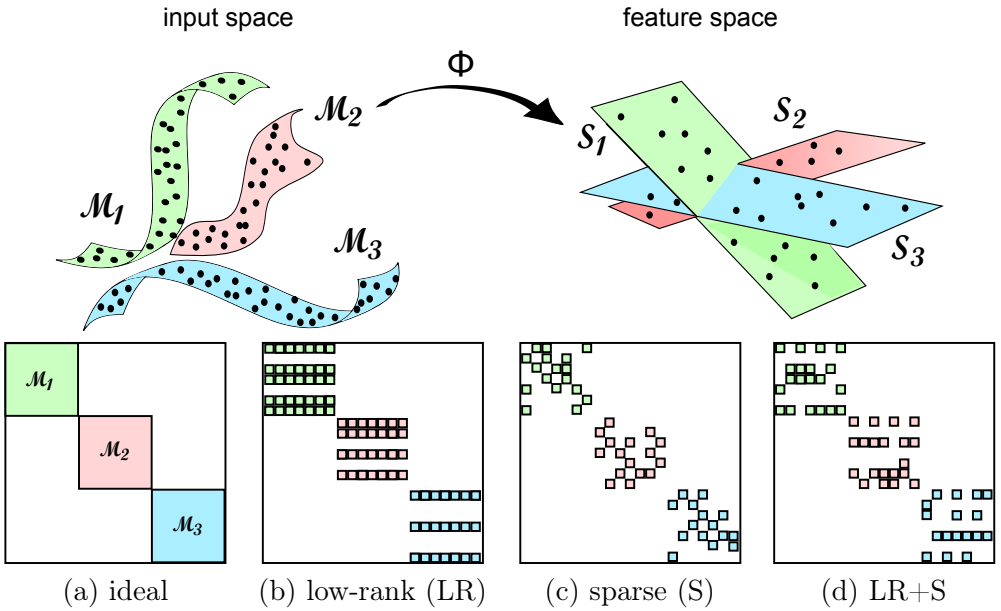
The term *manifold* is more general than *subspace* which defines a linear manifold on which the data lies. In this Thesis, we favour the term *subspace* and indicate about the linearity or non-linearity of the data. Therefore if not stated otherwise, we will assume that a *subspace* is linear and a *manifold* is potentially non-linear.

Many algorithms have been proposed for retrieving the few dimensions that are meaningful to describe the data subspace. The most known example is the Principal Component Analysis (PCA), retaining only the first directions of maximal variance of the data.

In remote sensing, the hyperspectral image data can be assumed as a low dimensional manifold [Bachmann et al., 2005]. This is due to both their high dimensionality and the proximity of the spectral bands in the electromagnetic spectrum (making the data extremely correlated across the dimensions). This allows to search for a low dimensional subspace approximating the data [Healey and Slater, 1999]. However, the non-linear nature of the hyperspectral data requires non-linear dimensionality reduction methods [Lunga et al., 2014]. Non-linear dimensionality reduction techniques have been applied successfully, for example in aligning multi-temporal hyperspectral manifolds [Yang and Crawford, 2011]. However, they are restricted to data lying on a single manifold or, at best, on a collection of well separated subspaces.

Among them, we find the kernel PCA [Schölkopf et al., 1998] and the probabilistic Gaussian process latent variable models (GPLVM) [Lawrence, 2005], ISOMAP [Tenenbaum et al., 2000, Bachmann et al., 2005] which exploits geodesic distances, Laplacian Eigenmaps [Belkin and Niyogi, 2003] and Locally Linear Embedding [Saul and Roweis, 2003] which exploits local neighbourhood distances, Locality Preserving Projections [He and Niyogi, 2004] which forces projections to be in the *span* of the data and more recently, methods aligning the local tangents to the manifold [Zhang et al., 2007] or finding the principal polynomials describing the data [Laparra et al., 2012]. Beyond their own specificities, these methods employ notions of local pairwise similarity, the most often described, either by a graph constructed from an affinity matrix or by a kernel encoding similarities. Both approaches are equivalent, up to some extent [Ham et al., 2004]. Usually, the solution reduces to a generalized eigenvalue problem, which is solved to find a low dimensional structure, where the local similarities are preserved. In this low dimensional space, it is then possible to perform any statistical analysis: classification, regression, or clustering.

The situation observed in real data is often very different: in many circumstances, the data is composed of multiple possibly non-linear subspaces, which may also intersect. In Figure 8.1, the multi-manifold assumption on the data is illustrated. Considering a collection of manifolds would improve their discrimination, as well as the quality of the reduction of dimensionality. More recently, approaches dealing with unions of subspaces have been proposed for the specific case of linear subspaces. They seek a representation of the data which is constrained to either use only few samples for its representation, which will automatically select neighbours from the same subspace (Sparse Subspace Clustering, SSC [Elhamifar and Vidal, 2013]), or to retrieve a low-rank representation, which will impose the use of similar samples for the representation of samples from the same manifold (Low Rank Subspace Clustering, LRSC [Vidal and Favaro, 2013, Liu et al., 2013]). For better visualization of these representations, see Figure 8.1. In (a) the ideal




(a) ideal (b) low-rank (LR) (c) sparse (S) (d) LR+S

Figure 8.1: (top) Schematic of the non-linear mapping of data in a RKHS where linear subspace can be retrieved. (a) The true affinity matrix between the manifolds and the different properties of the representation matrix $\mathbf{C}$: (b) low-rank, (c) sparse and (d) low-rank + sparse. Once a graph is built from these representation coefficients, the matrix properties translate into graph characteristics. The low-rankness induces dense intra-manifold connectivity while the sparsity ensures local affinity and manifold separation. A trade-off between the intra-manifold connectivity and the manifold separation, meaning both low-rank and sparse properties, is ideal to recover the three subspaces via spectral clustering.

representation matrix for clustering has non-zero coefficients only between samples of the same class manifold. In (b) is illustrated a low-rank matrix (here rank= 3), while in (c) is illustrated a sparse representation matrix. The latter may produce matrices where samples from the same manifold are not interconnected by non-zero coefficients. This why the combination of both low-rank and sparse representation can lead to a matrix that has only non-zero coefficients between samples of the same manifold and ensure that they are all inter-connected inside a manifold. These representation are powerful tools which have a very restricted number of user defined parameters and which provided state-of-the-art results in face recognition and motion segmentation problems [Vidal, 2011]. These linear approaches are introduced in Section 8.2, while our proposed non-linear representation combining both low-rank and sparse constraints (KLRSSC) is presented in Chapter 9.

The more general assumption of data composed of *class-wise manifolds*, i.e. each class lies on a dedicated non-linear manifold, has been rarely exploited in the context of clustering. In supervised classification problems, the evaluation of perturbation of each class-wise manifold for a test sample has shown to reach the discrimination performance of the SVM [Courty et al., 2011, Chapel et al., 2012]. Very recently, sparse coding and dictio-

nary learning approaches, which assume that the data lie in linear subspaces, have been extended to non-linear settings [An et al., 2007, Zhang et al., 2012, Nguyen et al., 2012, Xiao et al., 2012, Gao et al., 2013], together with non-linear matrix factorization [Zhang et al., 2006, Zafeiriou and Petrou, 2010] and very recently non-linear compressed sensing [Ramamurthy and Spanias, 2011, Qi and Hughes, 2011, Beck and Eldar, 2012]. Sparse coding and dictionary learning approaches have been exploited for supervised classification of hyperspectral images and produced state-of-the-art results [Chen et al., 2013b, Soltani-Farani et al., 2013].

In the unsupervised setting, only a global dictionary can be learned. This has been exploited for performing vector quantization prior to clustering for unsupervised classification in multi-spectral images Moody et al. [2012].

## 8.2 Linear subspaces

In this section, we review the state-of-the-art approaches for clustering multiple linear subspaces. Some specific applications in computer vision, such as face clustering or linear motion segmentation, meet the assumption of data lying on a union of linear subspaces. The problem of clustering linear subspaces [Vidal, 2011] has been tackled by numerous approaches such as the k-planes [Bradley and Mangasarian, 2000], where hyperplanes are fitted to the data, the generalized PCA [Vidal et al., 2005], that fits and factorizes polynomials algebraically, the Spectral Curvature Clustering (SCC) [Chen and Lerman, 2009] that exploits the local flatness, the low rank subspace clustering (LRSC) [Vidal and Favaro, 2013, Liu et al., 2013] and the SSC [Elhamifar and Vidal, 2013] which exploit the *self-expressiveness property* of the data with different constraints on the matrix of coefficients. Both LRSC and SSC provide the most promising experimental results, with the ability to handle noise and samples with few missing or erroneous dimensions [Liu et al., 2013, Elhamifar and Vidal, 2013]. Moreover, they do not require to set the dimensionality of the subspaces in advance and handle potentially intersecting linear subspaces. A trade-off between the sparse representation and the low-rank representation (LRSSC) has been studied in [Wang et al., 2013a] and used in [Zhang et al., 2013a] for spatially regularized sparse coding.

In remote sensing, the use of sparse representation has been mostly exploited in supervised [Chen et al., 2011a, Wang et al., 2013b, Camps-Valls et al., 2014] or semi-supervised classification [Gu and Feng, 2012] as well as reconstruction [Willett et al., 2014], retrieval of material signatures [Charles et al., 2011] and target detection [Chen et al., 2011c,b, Fowler and Du, 2012, Krishnamurthy et al., 2012, Willett et al., 2014, Nasrabadi, 2014]. More advanced priors such as group-sparse representation, meaning that only elements from the same class are used altogether, have also been investigated in the context of supervised hyperspectral image classification [Sun et al., 2013] and led to improved results. However, the assumption of linear subspaces rarely holds in practice and restricts

drastically the success of these approaches.

Let us enter into some details of the SSC, LRSC and LRSSC algorithms before reviewing non-linear approaches in Section 8.3.

### 8.2.1 Sparse Subspace Clustering

**Problem formulation** Consider a set of samples $\mathbf{x}_i \in \mathbb{R}^d$ with $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_i, .., \mathbf{x}_N]$ lying on an union of $n$ linear subspaces $\cup_{k=1}^{n} \mathcal{S}_k$. Following the self-expressiveness property [Elhamifar and Vidal, 2013], each sample can be expressed as a linear – or affine – combination of the others. If a sparse representation is imposed, the few points selected in the linear combination will tend to belong to the same subspace. This property allows to build a graph $\mathcal{G}$ from the sparse linear coefficients $\mathbf{C} = [\mathbf{c}_1, .., \mathbf{c}_N]$ by setting the graph weights as $W = |\mathbf{C}| + |\mathbf{C}|^T$. The graph will show $n$ disconnected components corresponding to the different subspaces. This graph-cut problem can be solved via spectral clustering, which performs $k$-means clustering on the rows of the eigenvectors of the graph Laplacian [Ng et al., 2002]. As introduced above, SSC represents a sample $\mathbf{x}_i$ by a linear combination of the others samples

$$\mathbf{x}_i = \mathbf{X}\mathbf{c}_i, \quad c_{ii} = 0 \tag{8.1}$$

where $\mathbf{c}_i$ is a sparse vector of coefficients. The trivial solution $c_{ii} = 1$ is avoided by setting it to zero.

Finding the sparse coefficients corresponds to the following sparse optimization problem

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_q \quad s.t. \quad \mathbf{x}_i = \mathbf{X}\mathbf{c}_i, \quad c_{ii} = 0 \tag{8.2}$$

where the $\ell_q$ norm is a sparsifying norm: $q \leq 1$. For $q = 1$, the problem can be solved by convex optimization using Basis Pursuit or ADMM [Elhamifar and Vidal, 2013]. For $q = 0$, the problem is NP-hard but can be solved via Greedy algorithms such as OMP. Even if OMP is suboptimal for sparse signal recovery, it tends to provide affinity matrices with energy spread across more edges, which in turn can lead to better performances with spectral clustering algorithm [Dyer et al., 2013]. An interesting alternative to the formulation of (8.2) is presented in Wang et al. [2011a]. They replace it by $\|C^T C\|_1$ which leads to a convex problem solved via convex quadratic programming. In terms of optimization this quadratic problem with box constraint is less costly than variable splitting schemes for sparse representation.

**Noise** The formulation of (8.2) is not robust to noise. The relaxation of the equality constraint allows a certain noise level $\epsilon$. The noisy SSC formulation is then

$$\min_{\mathbf{c}_i} \|\mathbf{c}_i\|_q \quad s.t. \quad \|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 \leq \epsilon, \quad c_{ii} = 0 \tag{8.3}$$

Notice that for $\ell_q = \ell_1$, 8.3 is the LASSO.

**Outlying entries** In SSC [Elhamifar and Vidal, 2013] and its supervised counterpart SRC [Wright et al., 2009], the problem due to noise present solely in few dimensions, e.g. strong illumination effects or occlusion in face recognition problems, is alleviated by adding a sparse error matrix $\mathbf{E}$ in the representation. The formulation of SSC in matrix notation becomes

$$\min_{\mathbf{C}} \|\mathbf{C}\|_q + \lambda_E \|\mathbf{E}\|_q \quad s.t. \quad \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E}, \quad \operatorname{diag}(\mathbf{C}) = 0 \tag{8.4}$$

The matrix $\mathbf{E}$ represents the outlying entries. The level of sparsity of this specific type of noise is controlled by $\lambda_E$. This can also be seen as the original SSC formulation with the augmentation of the dictionary of signals with the identity matrix: $[\mathbf{X}\,\mathbf{I}]$ and of the coefficients $\mathbf{C}$ with $\mathbf{E}$: $\left[\mathbf{C}^T\mathbf{E}^T\right]^T$.

### 8.2.2 Low-Rank Subspace Clustering

The idea behind LRSC is to exploit the subspace intra-connectivity of data samples. The ideal matrix of coefficients representing the data connectivity is low-rank, where the rank is related to the number of subspaces present in the data.

Our goal is not to find a low-rank approximation of the data as in PCA or robust PCA [Candès et al., 2011] but a low-rank matrix of coefficients representing the data. The motivation for a low-rank coefficient matrix comes from the ideal affinity matrix, illustrated in Figure 8.1 (a). The ideal affinity matrix between the samples, equivalent in classification to the label matrix, is equal to 1 for samples of the same subspace and null otherwise. Therefore this block-diagonal matrix with $n$ blocks for $n$ subspaces has a rank equal to $n$. The number of subspaces $n$ is assumed to be smaller than the number of samples $N$, allowing to seek a low-rank matrix.

The LRSC formulation is

$$\min_{\mathbf{C}} \operatorname{rank}(\mathbf{C}) \quad s.t. \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \mathbf{C} = \mathbf{C}^T \tag{8.5}$$

whose convex relaxation is

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* \quad s.t. \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \mathbf{C} = \mathbf{C}^T \tag{8.6}$$

The desired symmetric property of the coefficients is directly obtained from the closed-form solutions hereafter:

- In the noise-free case, the closed-form solution to (8.6) is the Shape-Interaction Matrix (SIM) [Costeira and Kanade, 1998, Liu et al., 2011]: $\mathbf{C} = \mathbf{V}_1\mathbf{V}_1^T$, where $\mathbf{X} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^T$ is obtained from the skinny SVD (null singular values and the corresponding columns of $\mathbf{U}$ and rows of $\mathbf{V}$ are removed to form $\Sigma_1$, $\mathbf{U}_1$ and $\mathbf{V}_1$).

- If we relax the constraints, the LRSC formulation becomes

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* + \frac{\tau}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 \quad s.t. \quad \mathbf{C} = \mathbf{C}^T \tag{8.7}$$

  and its corresponding closed form solution is $\mathbf{C} = \mathbf{V}\mathcal{P}_\tau(\Sigma)\mathbf{V}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ and $\mathcal{P}_\tau$ is the polynomial-thresholding operator [Vidal and Favaro, 2013] defined as

$$\mathcal{P}_\tau(\sigma) := \begin{cases} 1 - \frac{1}{\tau\sigma} & \sigma > \frac{1}{\sqrt{\tau}} \\ 0 & \sigma \leq \frac{1}{\sqrt{\tau}} \end{cases} \tag{8.8}$$

- For noisy data, the formulation is

$$\min_{\mathbf{A},\mathbf{C}} \|\mathbf{C}\|_* + \frac{\tau}{2}\|\mathbf{A} - \mathbf{A}\mathbf{C}\|_F^2 + \frac{\alpha}{2}\|\mathbf{X} - \mathbf{A}\|_F^2 \quad s.t. \quad \mathbf{C} = \mathbf{C}^T \tag{8.9}$$

  where $\mathbf{A}$ is an unknown clean dictionary. We denote (8.9) as the Low-Rank Subspace Clustering (LRSC). The problem 8.9 is not convex, however it has a unique solution for a certain range of $\tau$ and $\alpha$. Actually, when $3\tau > \alpha$ two critical points exists, with one of the two being the optimal solution. To facilitate the thresholding procedure, it is approximated by two piecewise linear functions. The approximate polynomial-thresholding is

$$\tilde{\mathcal{P}}_{\alpha,\tau}(\sigma) := \begin{cases} \sigma & \sigma > \tilde{\sigma_*} \\ \frac{\alpha}{\alpha+\tau}\sigma & \sigma \leq \tilde{\sigma_*} \end{cases} \tag{8.10}$$

  where $\tilde{\sigma_*} = \sqrt{\frac{\alpha+\tau}{\alpha\tau} + \sqrt{\frac{\alpha+\tau}{\alpha^2\tau}}}$. We see that the singular values above $\tilde{\sigma_*}$ are kept untouched. The behaviour of $\tilde{\sigma_*}$ with respect to $\alpha$ and $\tau$ is first quadratic then linear with respect to $\tau$ and increases for decreasing $\alpha$. The noisier the data, the lower the $\alpha$ and $\tau$ resulting in less singular values preserved.

  The closed-form solution to (8.9) is therefore obtained by the approximate polynomial-thresholding of the SVD of the noisy data: $\mathbf{C} = \mathbf{V}\mathcal{P}_\tau(\tilde{\mathcal{P}}_{\alpha,\tau}(\Sigma))\mathbf{V}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ and the clean dictionary can be recovered from $\mathbf{A} = \mathbf{U}\tilde{\mathcal{P}}_{\alpha,\tau}(\Sigma)\mathbf{V}^T$.

In the case of sparse outlying noise, alternative minimization approaches have been proposed via Augmented Lagrange Multipliers (ALM) [Liu et al., 2013] or exploiting

both shrinkage and polynomial thresholding for recovering both coefficients and outlying noise [Vidal and Favaro, 2013].

### 8.2.3 Sparse and low-rank Subspace Clustering

There is a trade-off between the sparse representation and the low-rank representation [Wang et al., 2013a]. Both SSC and LRSC are based on the self-expressiveness property of the data samples but LRSC shows denser intra-subspace connectivity. In certain situation, they fail to produce a graph allowing clustering of the subspaces. For skewed or high-dimensional data distributions, SSC produces a poorly connected graph [Nasihatkon and Hartley, 2011] which leads to disconnected components, while LRSC appears to experimentally not suffer that same problem. For dependent or intersecting subspaces, SSC can resolve the clustering, while LRSC will fail in these situations.

The ideal graph for spectral clustering should benefit from both inter-class separation and intra-class connectivity and be found as the result of both low-rank and sparse representation. The trade-off between the graph-connectivity and the self-expressiveness of the samples can be formulated in convex form as a weighted sum of the nuclear and $\ell_1$ norm

$$\min_{\mathbf{C}} \frac{1}{1+\lambda}\|\mathbf{C}\|_* + \frac{\lambda}{1+\lambda}\|\mathbf{C}\|_1 \quad s.t. \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \operatorname{diag}(\mathbf{C}) = 0 \tag{8.11}$$

with $\lambda$ balancing the low-rank and the sparsity properties of the coefficient matrix $\mathbf{C}$. Increasing $\lambda$ will add more sparsity in the graph, which helps the recovery of dependent subspaces. Experimental results show a phase transition phenomenon, with the recovery of the subspaces for $\lambda > \lambda_*$. Values of $\lambda$ right after the transition provide the best results.

Table 8.1: Summary of the principal characteristics of state-of-the-art methods for non-linear subspaces/manifolds clustering as well as our proposed methods.
Legend: ⋱ flexible neighbourhood, ◯◯: handles separated subspaces, χ: handles intersecting subspaces, D: dimension free, T: tangent spaces.

| Method | ⋱ | ◯◯ | χ | D | T |
|---|---|---|---|---|---|
| SC [Ng et al., 2002] | × | ✓ | - | ✓ | - |
| STSC [Zelnik-Manor and Perona, 2004] | × | ✓ | - | ✓ | - |
| $K$-Manifolds [Souvenir and Pless, 2005] | × | × | ✓ | × | - |
| Cao *et al.* [Cao and Haralick, 2006] | × | ✓ | (✓) | ✓ | - |
| MiPPS [Koyejo and Ghosh, 2009] | × | ✓ | ✓ | × | - |
| LLMC [Goh and Vidal, 2007] | × | ✓ | × | ✓ | - |
| SMCE [Elhamifar and Vidal, 2011] | (✓) | ✓ | (✓) | ✓ | - |
| SMMC [Wang et al., 2011b] | × | ✓ | ✓ | × | ✓ |
| RMMSL [Gong et al., 2012] | × | ✓ | ✓ | (×) | ✓ |
| SCLPCA [Arias-Castro et al., 2013] | × | ✓ | ✓ | × | ✓ |
| Proposed in this Thesis: | | | | | |
| KSSC → Section 9.1.3 | ✓ | ✓ | ✓ | ✓ | - |
| KLRSC → Section 9.1.2 | ✓ | ✓ | × | ✓ | - |
| KLRSSC → Section 9.1.1 | ✓ | ✓ | ✓ | ✓ | - |

## 8.3 Non-linear subspaces

In this section, we review the state-of-the-art approaches for clustering multiple non-linear manifolds, before introducing our approaches (KLRSSC, KLRSC, KSSC) in Chapter 9.

The problem of clustering multiple non-linear subspaces has been investigated under different assumptions. The major difference between the approaches is their definition of affinity between samples: either based on local neighbourhoods, on locally linear relationships or on tangent spaces similarities. Table 8.1 summarizes their characteristics.

The first family is composed of methods based on local neighbourhoods. Two well known approaches are spectral clustering [Ng et al., 2002] (SC) and its self-tuning extension (STSC) [Zelnik-Manor and Perona, 2004]. Both assume well separated subspaces and use a local neighbourhood of fixed size $k$ to build a graph, which is then clustered via its Laplacian. In STSC, the graph edges are weighted using a local bandwidth, which is the distance to the $k^{th}$ neighbour [Zelnik-Manor and Perona, 2004]. The $k$-Manifolds algorithm [Souvenir and Pless, 2005], which is based on heuristic expectation-maximization, also exploits a local neighbourhood to retrieve geodesic distances, as ISOMAP does. However, it handles only intersecting subspaces and fails, when applied to well separated ones.

The second family of methods assumes a certain smoothness of the subspaces which

can be locally approximated as linear. In general, this assumption requires a large number of samples to be met. In [Cao and Haralick, 2006], a local estimate of the data dimensionality is obtained by local PCA and then used for clustering multiple subspaces. Although successful on non-linear problems, the method cannot deal with intersecting subspaces. In [Koyejo and Ghosh, 2009], a mixture of probabilistic principal surfaces is fit to multiple subspaces in high-dimensional data via the EM algorithm. The model handles intersecting manifolds, but requires a large amount of samples and can fail in presence of important outliers. In [Goh and Vidal, 2007], an affine subspace is fitted to each point and to its $k$ nearest neighbours. The authors of [Goh and Vidal, 2007] propose a closed-form solution. A low-rank embedding is presented in [Liu et al., 2011], which generalizes LLE to multiple manifolds using a low-rank constraint, whereas a sparse optimization program is used in the Sparse Manifold Clustering and Embedding (SMCE) [Elhamifar and Vidal, 2011]. The latter allows to vary the number of neighbors $k$ for each sample. This increases the robustness to outliers and intersections. These methods can perform poorly on under-sampled or intersecting subspaces. In these cases, the local affine subspace assumption might not hold and degrade the clustering results.

The last family of methods uses the estimation of a tangent space at each sample to build the affinity matrix. The Spectral Multi-Manifold Clustering (SMMC) in [Wang et al., 2011b], fits mixtures of probabilistic PCA to the data to estimate the tangent spaces, whereas in [Arias-Castro et al., 2013] local PCA is used instead. In both cases, the affinity matrix is based on similarities between the local tangent spaces instead of single samples, raising the opportunity to handle non-linear and possibly intersecting subspaces of different dimensionality. However, these methods requires the *a priori* knowledge of the number of dimensions and components in the mixture. In the Robust Multi-Manifold Structure Learning (RMMSL) [Gong et al., 2012], the local tangent spaces are estimated by a weighted low-rank factorization which relaxes the need to know the dimensionality of the subspaces. Although it is the first work to raise attention to the problem of outliers, their method is restricted to smooth subspaces with low curvature. Finally, the kernel extension of the Spectral Curvature Clustering (KSCC) [Chen et al., 2009] maps the data to a high dimensional space where flat surfaces are retrieved. The mapping is implicit via the kernel functions, however the high computational complexity requires an iterative computation of the affinity weights and is prone to local minima.

## 8.4 Summary

In this Chapter, we have introduced the problem of retrieving the structure of data lying on a composition of multiple subspaces or manifolds. We have described the state-of-the-art approaches building graphs using sparsity or/and low-rank constraints. In the next Chapters, the similar constraints will be exploited to retrieve the data manifold structure via a RKHS representation and build a graph that will be used for clustering (Chapter 9) or semi-supervised one-class classification (Chapter 10).

# 9 Unsupervised classification via non-linear low-rank and sparse subspace representation

This chapter tackles the problem of unsupervised analysis of high-dimensional data such as hyperspectral images by assuming multiple non-linear subspaces, i.e. manifolds, as illustrated in Figure 8.1. In Section 9.1, we propose a spectral clustering method based on a graph representing the data structure, which is assumed to be an union of multiple manifolds. The method constraints the data samples to be expressed as a low-rank and sparse combination of the others in a RKHS. This captures the global (low-rank) and local (sparse) structures. Spectral clustering is applied on the graph to assign the samples to the different manifolds. In Section 9.2, a scalable approach is proposed in which the optimization is first performed on a subset of the data and then it is applied to the entire image using a non-linear collaborative representation respecting the manifolds structure.

Finally, experiments on standard face and object databases and two hyperspectral images show the adequacy of KLRSSC in dealing with multiple manifolds[1].

## 9.1 Non-linear subspace representation

In this Section we present the KLRSSC algorithm as well as the particular cases where only sparsity (KSSC) or low-rank (KLRSC) constraints are imposed. These approaches extend the methods of Sections 8.2.1, 8.2.2 and 8.2.3 to RKHS and combine the following properties, which are desirable for a non-linear subspace clustering method: i) they can cluster both linear and non-linear data, ii) they utilize a sparse/low-rank representation to provide a flexible neighbourhood and a well connected graph, iii) they can cluster correctly intersecting manifolds and iv) they can provide out-of-sample and scalable clustering.

[1] This Chapter is related to [de Morsier et al., 2014c,d]

### 9.1.1 Kernel Low-Rank and Sparse Subspace Clustering (KLRSSC)

Let us introduce the main problem and the corresponding notations, before going through the main formulation of KLRSSC and the particular cases: KLRSC and KSSC.

We consider a set of $N$ samples $\mathbf{X} = \{\mathbf{x}_i, ..., \mathbf{x}_N\} \in \mathbb{R}^d$, drawn from a union of $n$ non-linear subspaces or manifolds $\{\cup_{k=1}^{n} \mathcal{M}_k\}$. Let us define a non-linear mapping function $\Phi$ between the input space and a higher dimensional feature space $\mathcal{H}$, a RKHS, as

$$\Phi : \quad \mathbf{x} \in \mathbb{R}^d \rightarrow \Phi(\mathbf{x}) \in \mathcal{H} \tag{9.1}$$

We use the following notation when referring to the whole set of mapped samples,

$$\Phi(\mathbf{X}) = [\Phi(\mathbf{x}_1), .., \Phi(\mathbf{x}_N)] \tag{9.2}$$

For the set of samples $\mathbf{X}$, we would like to find the coefficient matrix $\mathbf{C}$ that represents well the data in the mapped feature space

$$\Phi(\mathbf{X}) \cong \Phi(\mathbf{X})\mathbf{C} \tag{9.3}$$

and avoids the trivial solution where $\mathbf{C}$ is the identity matrix $\mathbf{I}$. Finally, based on the sample correspondence described in $\mathbf{C}$, each sample can be assigned to its corresponding manifold. This would lead to the ideal coefficient matrix being block-diagonal, for samples ordered by manifolds, where each block corresponds to a different manifold.

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{C}_2 & \cdots & 0 \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{C}_n \end{bmatrix} \tag{9.4}$$

The generalization of sparse and low-rank subspace clustering for data mapped in a RKHS builds on previous sparse and low-rank subspace clustering approach [Wang et al., 2013a]. As in the linear case, combining sparsity and low-rank criterion allows to exploit the self-expressiveness of data samples with an increased graph-connectivity. In [Elhamifar and Vidal, 2013], the graph-connectivity is improved by adding a $\ell_{row,1}$ norm which minimizes the number of non-zero rows of $\mathbf{C}$. They observed that a low $\ell_{row,1}$ regularization could be beneficial whereas a too large regularization would connect the subspaces altogether. The regularization of the number of non-zero row is very similar to the minimization of the rank of $\mathbf{C}$. We will observe a similar behaviour on the amount of rank regularization needed for recovering correctly the non-linear subspaces.

**Algorithm 3** Kernel Low-Rank and Sparse Subspace Clustering (KLRSSC)

**Input:** A set of $N$ samples $\mathbf{x}_i$ from an union of $n$ subspaces $\mathcal{M}_k$ and a kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ representing the dot-product in an appropriate RKHS.

1: Get $\mathbf{C}$ by solving (9.5) using the optimization scheme of (9.6)
2: Normalize the columns of $\mathbf{C}$ as $\mathbf{C}_i \leftarrow \frac{\mathbf{C}_i}{\|\mathbf{C}_i\|_\infty}$
3: Build the graph representing data manifolds affinity with weights $W = |\mathbf{C}| + |\mathbf{C}|^T$
4: Apply spectral clustering [Ng et al., 2002] to the graph

**Output:** Samples clustered in $n$ groups representing the manifolds

### General formulation

The general formulation for sparse and low-rank subspace clustering of data mapped by an implicit function $\Phi$ is

$$\min_{\mathbf{C}} \frac{1}{1+\lambda}\|\mathbf{C}\|_* + \frac{\lambda}{1+\lambda}\|\mathbf{C}\|_1 \quad s.t. \quad \Phi(\mathbf{X}) = \Phi(\mathbf{X})\mathbf{C}, \quad \text{diag}\,(\mathbf{C}) = 0 \tag{9.5}$$

where $\|\mathbf{X}\|_1$, defined as $\sum_i \|\mathbf{x}_i\|_1$, is the convex relaxation of the pseudo-norm $\ell_0$ and $\|\mathbf{X}\|_*$ is the convex relaxation of the rank $(\mathbf{X})$ function.

We would like to notice that in (9.5), the assumption that the equality $\Phi(\mathbf{X}) = \Phi(\mathbf{X})\mathbf{C}$ holds in the induced feature space is made. However, notice that feature spaces induced by RBF kernels, or any normalized kernel, are hyperspherical with potentially every sample in a different axis [Courty et al., 2012]. This geometry of the space allows $\Phi(\mathbf{X}) \approx \Phi(\mathbf{X})\mathbf{C}$ and not a strict equality. This restriction is not problematic in our case, since this constraint is relaxed in our optimization approaches allowing for an approximation only.

Once the optimal representation coefficients $\mathbf{C}$ are retrieved from (9.5), a graph can be built. Symmetric graph edge weights are obtained as $W = |\mathbf{C}| + |\mathbf{C}|^T$. Finally, spectral clustering cuts the graph into separated components corresponding to the different manifolds. The full KLRSSC algorithm is summarized in Algorithm 3.

### Optimization in the feature space

In order to solve (9.5), we use an augmented Lagrangian approach and the alternating direction method of multipliers (ADMM). Auxiliary variables $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{A}$ and the error term are introduced in order to split the nuclear norm and $\ell_1$ norm and make the objective strictly convex w.r.t. $\mathbf{A}$ and $\mathbf{C}$. Notice that ADMM is designed and converges for an objective composed of two functions and not three. However, here the two terms with $\mathbf{C}_1$

and $\mathbf{C}_2$ are not dependent on each other and count as a single function.

$$\begin{aligned} \min \ & \frac{1}{1+\lambda}\|\mathbf{C}_1\|_* + \frac{\lambda}{1+\lambda}\|\mathbf{C}_2\|_1 + \frac{\lambda_z}{2}\|\Phi(\mathbf{X}) - \Phi(\mathbf{X})\mathbf{A}\|_F^2 \\ s.t. \ & \mathbf{A} = \mathbf{C}_2 - \operatorname{diag}(\mathbf{C}_2), \quad \mathbf{A} = \mathbf{C}_1 \end{aligned} \tag{9.6}$$

The parameter $\lambda_z = \frac{\alpha_z}{\mu_z}$ controls the level of noise, where $\alpha_z$ is a user-defined parameter and $\mu_z$ is the minimal maximum mutual coherence between samples: $\mu_z = \min_i \max_{j\neq i} |\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)| = \min_i \max_{j\neq i} K(\mathbf{x}_i, \mathbf{x}_j)$. This can be interpreted as the potential noise based on the smallest mutual coherence in the data [Dyer et al., 2013]. In [Elhamifar and Vidal, 2013], their Proposition 1 proves that in the SSC $\alpha_z > 1$ otherwise the trivial solution $\mathbf{C} = 0$ is obtained. This is due to the fact that the error will not counter balance enough the $\ell_1$ term for $\alpha_z \leq 0$.

The corresponding augmented Lagrangian objective is

$$\begin{aligned} \mathcal{L}(\mathbf{C}_\cdot, \mathbf{A}, \mathbf{\Delta}_\cdot) = & \frac{1}{1+\lambda}\|\mathbf{C}_1\|_* + \frac{\lambda}{1+\lambda}\|\mathbf{C}_2\|_1 + \frac{\lambda_z}{2}\|\Phi(\mathbf{X}) - \Phi(\mathbf{X})\mathbf{A}\|_F^2 \\ & + \frac{\mu_2}{2}\|\mathbf{A} - \mathbf{C}_2 + \operatorname{diag}(\mathbf{C}_2)\|_F^2 + \frac{\mu_3}{2}\|\mathbf{A} - \mathbf{C}_1\|_F^2 \\ & + tr(\mathbf{\Delta}_2^T(\mathbf{A} - \mathbf{C}_2 + \operatorname{diag}(\mathbf{C}_2))) + tr(\mathbf{\Delta}_3^T(\mathbf{A} - \mathbf{C}_1)) \end{aligned} \tag{9.7}$$

with $tr(\cdot)$ the trace operator on matrices and $\mathbf{\Delta}_2, \mathbf{\Delta}_3$ the Lagrangian multipliers enforcing $\mathbf{A} = \mathbf{C}_2 - \operatorname{diag}(\mathbf{C}_2)$ and $\mathbf{A} = \mathbf{C}_1$, respectively. Setting the derivative of the augmented Lagrangian w.r.t $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{A}$ to zero provides the following update steps

$$\mathbf{A}^{s+1} = (\lambda_z K(\mathbf{X}, \mathbf{X}) + (\mu_2 + \mu_3)\mathbf{I})^{-1}(\lambda_z K(\mathbf{X}, \mathbf{X}) + \mu_2\mathbf{C}_2^s + \mu_3\mathbf{C}_1^s - \mathbf{\Delta}_2^s - \mathbf{\Delta}_3^s) \tag{9.8}$$

$$\mathbf{C}_2^{s+1} = \mathcal{S}_{\frac{\lambda}{(1+\lambda)\mu_2}}(\mathbf{A}^{s+1} + \frac{\mathbf{\Delta}_2}{\mu_2}), \quad \mathbf{C}_2^{s+1} = \mathbf{C}_2^{s+1} - \operatorname{diag}\left(\mathbf{C}_2^{s+1}\right) \tag{9.9}$$

$$\mathbf{C}_1^{s+1} = \mathcal{T}_{\frac{1}{(1+\lambda)\mu_3}}(\mathbf{A}^{s+1} + \frac{\mathbf{\Delta}_3}{\mu_3}) \tag{9.10}$$

$$\mathbf{\Delta}_2^{s+1} = \mathbf{\Delta}_2^k + \mu_2(\mathbf{A}^{s+1} - \mathbf{C}_2^{s+1}), \quad \mathbf{\Delta}_3^{s+1} = \mathbf{\Delta}_3^k + \mu_3(\mathbf{A}^{s+1} - \mathbf{C}_1^{s+1}) \tag{9.11}$$

where $K(\mathbf{X}, \mathbf{X}) = \langle \Phi(\mathbf{X}), \Phi(\mathbf{X}) \rangle$ and $s$ is the iteration. The soft-thresholding operator $\mathcal{S}_\eta(\mathbf{x}) = sign(\mathbf{x})(\|\mathbf{x}\| - \eta)_+$ shrinks the values of $\mathbf{x}$ larger than $\eta$ and sets the others to zero. The singular value soft-thresholding operator is $\mathcal{T}_\eta(\mathbf{X}) = \mathbf{U}\mathcal{S}_\eta(\Sigma)\mathbf{V}^T$ where $\mathbf{U}\Sigma\mathbf{V}^T$ is the skinny SVD of $\mathbf{X}$, see page 121 for details. An adaptive penalty scheme can be adopted which modifies at each iteration the values of $\mu_2$ and $\mu_3$. However, we choose to keep $\mu_2, \mu_3$ fixed to avoid recomputing the matrix inversion at every iteration when updating $\mathbf{A}$.

### 9.1.2 Kernel Low-Rank Subspace Clustering (KLRSC): $\lambda \to 0$

**Formulation for $\lambda \to 0$**

The particular case where $\lambda = 0$ is similar to the low-rank representation (Section 8.2.2) problem but solved in a RKHS induced by the use of kernel functions. We remove the constraint $\text{diag}(\mathbf{C}) = 0$ specific to the sparsity-promoting $\ell_1$ norm and add the symmetry of the coefficients $\mathbf{C} = \mathbf{C}^T$. This last constraint is superficial since always fulfilled by the manner the coefficient matrix is recovered. Notice that the general formulation of (9.5) may produce similarly dense coefficient matrix under sufficient relaxation of the constraints.

We are looking for a low-rank representation $\mathbf{C}$ of the data matrix in an induced feature space $\Phi(\mathbf{X})$. The formulation of the Kernel Low-Rank Subspace Clustering (KLRSC) is

$$\min \|\mathbf{C}\|_* \quad s.t. \quad \Phi(\mathbf{X}) = \Phi(\mathbf{X})\mathbf{C}, \quad \mathbf{C} = \mathbf{C}^T \tag{9.12}$$

With only the nuclear norm in the objective, we can consider the relaxed noisy data formulation

$$\min \|\mathbf{C}\|_* + \frac{\tau}{2}\|\Phi(\mathbf{A}) - \Phi(\mathbf{A})\mathbf{C}\|_F^2 + \frac{\alpha}{2}\|\Phi(\mathbf{X}) - \Phi(\mathbf{A})\|_F^2 \quad s.t. \quad \mathbf{C} = \mathbf{C}^T \tag{9.13}$$

which represents the data in the feature space via a clean dictionary $\Phi(\mathbf{X}) = \Phi(\mathbf{A})\mathbf{C} + \mathbf{E}$.

This low-rank representation in the linear case ($\Phi(\mathbf{X}) = \mathbf{X}$) is derived in closed-form by performing hard, shrinkage or polynomial thresholding of the singular value decomposition (SVD) of the data matrix, as presented in Section 8.2.2. In the non-linear case, these SVD operations cannot be directly performed on the mapped data $\Phi(\mathbf{X})$ since its columns dimensions can be infinite and $\Phi$ is implicit.

Hopefully, the eigen decomposition of the Gram matrix $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ is related to the SVD decomposition of the data $\Phi(\mathbf{X})$. This allows to handle the implicit mapping $\Phi$ through the use of kernel matrices and lead to the following proposition.

**Proposition 1.** *Let $\Phi(\mathbf{X}) \in \mathcal{H}$ be the data samples $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^d$ mapped in a RKHS induced by $\Phi$ equipped with a kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$. The low-rank representation of the mapped data is given by $\mathbf{C} = \mathbf{V}\mathcal{T}(\Sigma)\mathbf{V}^T$, where $\mathbf{V}\Delta\mathbf{V}^T = svd(K(\mathbf{X}, \mathbf{X}))$ and $\Delta = \Sigma^T\Sigma$. The operator $\mathcal{T}$ thresholds the singular values in different ways depending on the noise assumption and constraint relaxation. The hard-thresholding operator for clean data as in* (9.12)*, polynomial-thresholding or approximate polynomial-thresholding for relaxed constraint and noisy data as in* (9.13)*.*

*Proof.* Let the SVD of the mapped data be $\Phi(\mathbf{X}) = \mathbf{U}\Sigma\mathbf{V}^T$. The decomposition cannot be directly computed since the mapping $\Phi$ is implicit. Meanwhile, the eigen decomposition

of its Gram matrix, which is the kernel matrix uniquely defining the RKHS, can be computed as $K(\mathbf{X}, \mathbf{X}) = \mathbf{V}\Delta\mathbf{V}^T$ with $\Delta = \Sigma^T\Sigma$. The closed-form solutions for the low-rank representation of the mapped data only require the singular values $\Sigma$ and the row space $\mathbf{V}$ as for non-mapped data. The proof relates then to the proofs for the linear cases of Vidal and Favaro [2013]. □

#### Optimization in the feature space

The advantage of the formulations (9.12) and (9.13) is the possibility to obtain closed-form solutions. As presented for the linear case in Section 8.2.2, the closed-form solution to the low-rank representation with noise is obtained via a polynomial thresholding $\tilde{\mathcal{P}}_{\alpha,\tau}(\sigma)$ which shrinks the singular values in a softer way than the standard shrinkage operator. The closed-form solution is $\mathbf{C} = \mathbf{V}\mathcal{P}_\tau(\tilde{\mathcal{P}}_{\alpha,\tau}(\Sigma))\mathbf{V}^T$, where $\mathbf{V}$ and $\Sigma$ are obtained from the kernel data matrix eigendecomposition.

### 9.1.3 Kernel Sparse Subspace Clustering (KSSC): $\lambda \to \infty$

#### Formulation for $\lambda \to \infty$

The particular case where $\lambda = \infty$ is similar to the SSC problem (Section 8.2.1), but solved in a RKHS induced by the use of kernel functions. A very similar approach has been developed in parallel by Patel et al. [2013]. They project the data in a latent space where SSC is performed. This latent space can be a RKHS induced by a Mercer kernel.

Similarly to SSC, each sample $\mathbf{x}_i$ is expressed as a sparse linear combination $\mathbf{c}_i$ of the others, exploiting the *self-expressiveness* property but this time in the high dimensional Hilbert space induced by $\Phi$. The formulation of the Kernel Sparse Subspace Clustering (KSSC) is

$$\min \|\mathbf{c}_i\|_q \quad s.t. \quad \Phi(\mathbf{x}_i) = \Phi(\mathbf{X})\mathbf{c}_i, \quad c_{ii} = 0, \tag{9.14}$$

where the norm $q$ is a sparsifying norm: $q \leq 1$. For $q = 1$, the problem can be solved by convex optimization using ADMM or Second-order Cone Programming (SOCP). For $q = 0$, the problem is NP-hard but can be approximated via the Greedy Kernel OMP algorithm [Li et al., 2011a]. The sparse coefficients $\mathbf{c}_i$ are non-null for samples of the same manifold (i.e. self-expressiveness property) and the trivial solution is avoided by setting $c_{ii}$ to zero. The problem of (9.14) is equivalent to finding a non-trivial sparse representation of $\Phi(x_i)$ in the dictionary $\Phi(X_{\setminus i})$.

#### Optimization in the feature space

The problem of (9.14) is optimized differently depending on the sparsity promoting norm used ($\ell_0$ or $\ell_1$). For $\ell_0$, the kernel Orthogonal Matching Pursuit (kOMP) can be used [Li et al., 2011a]. However, we consider the $\ell_1$ formulation being more stable experimentally and consistent with the other proposed optimization approaches.

The problem of (9.14) in its relaxed version with the sparsifying norm $\ell_1$ is a particular case of KLRSSC and can be solved with ADMM. The augmented Lagrangian corresponds to (9.7) without $\mathbf{C}_1$ and $\Delta_3$ terms.

$$\mathcal{L}(\mathbf{C}, \mathbf{A}, \mathbf{\Delta}) = \|\mathbf{C}\|_1 + \frac{\lambda_z}{2}\|\Phi(\mathbf{X}) - \Phi(\mathbf{X})\mathbf{A}\|_F^2$$
$$+\frac{\lambda_c}{2}\|\mathbf{A} - \mathbf{C} + \operatorname{diag}(\mathbf{C})\|_F^2 + tr(\mathbf{\Delta}^T(\mathbf{A} - \mathbf{C} + \operatorname{diag}(\mathbf{C}))). \tag{9.15}$$

with $\Delta$ the Lagrangian multiplier enforcing $\mathbf{A} = \mathbf{C} - \operatorname{diag}(\mathbf{C})$. As for KLRSSC, setting the derivative of the Lagrangian w.r.t. $\mathbf{A}$ and $\mathbf{C}$ to zero provides the same update rules as (9.11), where $\mathbf{C} = \mathbf{C}_2$, $\Delta = \Delta_2$ and $\mathbf{C}_1$, $\Delta_3$ are absent.

### 9.1.4 Handling noise

Most of the time, the data do not lie exactly on a subspace, mainly because of the presence of noise. The representation equality constraint can be relaxed to allow a certain level of noise $\epsilon$ in the feature space or as previously mentioned in Section 9.1.1, for the RBF kernel with its hyperspherical geometry which may not allow a perfect representation.

The geometric analysis of [Soltanolkotabi and Candes, 2012] guarantees the recovery of the subspaces in the noiseless case, even when the subspaces overlap. When considering the noisy case, authors of [Wang and Xu, 2013] provide bounds for the alternative LASSO formulation where a parameter controls the trade-off between sparsity and the reconstruction error. These bounds are a function of the noise level $\epsilon$ and of the distribution of the data in the subspace. The less spread the data are along the subspace, the smaller is the range of the valid parameters balancing sparsity and error. In this case, the recovery becomes more unstable.

For the case of noise acting only on few dimensions, a sparse matrix $\mathbf{E}$ can be added and its $\ell_1$ norm minimized together with $\mathbf{C}$. In our non-linear setting, it is not possible to add the matrix $\mathbf{E}$ in the representation, since the dimensions of the columns of $\mathbf{E}$ would be unknown and possibly infinite in the space induced by $\Phi$. A sparse matrix in the feature space could indicate outliers but not sparse outlying entries. This is further explored in Section 9.1.5 for outlier detection.

### 9.1.5 Handling outliers

The problem of outliers is somehow related to the problem of outlying entries. As seen in the previous section, the outlying entries are sparse when expressed in the coordinate system of the input space. In the the Kernel Empirical Space (KES) [Xiong et al., 2005] spanned by the mapped samples in the RKHS, each sample corresponds to a dimension in this mapped coordinate system. The outliers can be considered as localized on an isolated part of the space. This is equivalent to the problem of finding outlying entries but in the feature space. Actually, this can be done by looking at the $\ell_1$-norm of $\mathbf{c}_i$. For an outlier, its representation $\mathbf{c}_i$ will be less sparse and produces a larger norm, while for an inlier its representation will be of a sparsity related to the dimension of the subspace [Soltanolkotabi and Candes, 2012].

Outliers can then be put aside when performing spectral clustering and get either a random assignment or the one of their most similar sample. For fair comparisons between algorithms, we did not consider the detection of the outliers in our experiments.

### 9.1.6 Handling missing entries

The problem in (9.5) can handle missing entries in $\mathbf{X}$ by simply restricting the evaluation of the objective (and its respective norms $\ell_1$, $\ell_*$ and $\ell_F$) to the observed entries using a diagonal matrix acting as a mask [Nguyen and De la Torre, 2008, Elhamifar and Vidal, 2013]. In the case of missing entries in unknown location, the problem comes back to handling a specific type of noise or outliers if not similar for all samples. There is a forthcoming work of Candès related to the linear SSC with missing data which can be read at www.stanford.edu/~mahdisol/RSC.

## 9.2 Scalable clustering

### 9.2.1 Collaborative representation

The computational complexity of these algorithms is in the order of $\mathcal{O}(N^3)$, which restricts their usage to small to middle sized datasets. The scalability and out-of sample issues are solved by proceeding in three steps: we first find the manifolds on a training set using the KLRSSC described above and then predict manifold membership for test samples. To do so, we use a non-linear collaborative representation, as proposed for the linear case [Peng et al., 2013]: we express the test sample $\mathbf{x}_t$ as a function of all the training samples $\mathbf{X}$: $\Phi(\mathbf{x}_t) = \Phi(\mathbf{X})\mathbf{c}_t$, thus respecting the manifold structure. The $\mathbf{c}_t$ vector representing $\mathbf{x}_t$ is

obtained by the multivariate kernel ridge regression [An et al., 2007] formulation

$$\begin{aligned}\mathbf{c}_t &= (\Phi(\mathbf{X})^T\Phi(\mathbf{X}) + \gamma\mathbf{I})^{-1}\Phi(\mathbf{X})^T\Phi(\mathbf{x}_t) \\ &= (K(\mathbf{X},\mathbf{X}) + \gamma\mathbf{I})^{-1}K(\mathbf{X},\mathbf{x}_t)\end{aligned} \tag{9.16}$$

where the $\gamma$ parameter is a regularization parameter on the coefficients and relates to the inverse noise variance. The $\ell_2$ regularization is less computationally demanding than the $\ell_1$ and $\ell_*$ and has shown to be similarly efficient [Deng et al., 2013, Zhang et al., 2011] when used with training sets of sufficient size in supervised settings or unsupervised but with dependent subspaces only [Lu et al., 2012]. If the superior discrimination power between the $\ell_1$ and the $\ell_2$ norm is still an open debate, the sparse representation allows to be robust to outliers and lower-dimensional input spaces, and last but not least it enables to recover the data structure in unsupervised approaches when the subspaces are dependent or intersecting.

If $\mathbf{x}_t$ belongs to manifold $k$, the reconstruction using only the training samples assigned to $k$ in the training phase will be the most accurate. Therefore, in the final step we compare the residuals obtained when reconstructing $\mathbf{x}_t$ only with the training pixels belonging to each manifold: $\arg\min_k \{\mathbf{r}_{t,k}\}$, where

$$\begin{aligned}\mathbf{r}_{t,k} &= \frac{|\Phi(\mathbf{x}_t) - \Phi(\mathbf{X})\delta_k(\mathbf{c}_t)|_2^2}{|\delta_k(\mathbf{c}_t)|_2^2} \\ &= [K(\mathbf{x}_t,\mathbf{x}_t) - 2K(\mathbf{x}_t,\mathbf{X})\delta_k(\mathbf{c}_t) \\ &+ \delta_k(\mathbf{c}_t)^T K(\mathbf{X},\mathbf{X})\delta_k(\mathbf{c}_t)]/\,|\delta_k(\mathbf{c}_t)|_2^2\end{aligned} \tag{9.17}$$

with $\delta_k(\mathbf{c})$ a function returning 0 to all the elements of $\mathbf{c}$ which are not from manifold $k$.

### 9.2.2 About scalability

The proposed out-of-sample extensions allows to avoid rebuilding the whole graph when a new sample is available. However, the scalability issue remains a problem if the training samples are already too many. In large problems, the kernel matrix can be too large to be manipulated or even stored. In that case, Nyström methods allows to manipulate a smaller matrix which is related to a low-rank approximation of the larger initial matrix (in our case the kernel matrix). This requires to find landmarks representative of the whole initial training set [Chen and Cai, 2011, Kumar et al., 2012]. In [Cawley and Talbot, 2002], the specific problem of finding a low-rank approximation for the kernel ridge regression is investigated. They propose to greedily add samples which minimize the reconstruction error. A convex optimization approach instead of the previous greedy method for reduced rank kernel ridge regression is introduced in [Chen et al., 2013a]. In [Cai and Osher, 2010], a fast Singular Value Thresholding is proposed to obtain the low-rank approximation. Recently Liu and Yan [2012] tackled the subspace clustering problem. The coefficient

matrix $\mathbf{C}$ is retrieved by decomposing it into two smaller matrices $Q$ and $J$, with $Q$ contrained to the Stiefel manifold ($Q^T \cdot Q = I$). The low-rank structure can be retrieved through the smaller matrix $J$, since the nuclear norms of $Q \cdot J$ and $J$ are equivalent. These methods are foundations for extending our proposed KLRSSC formulation to larger scale, meanwhile we kept our experiments at a reduced scale to better study their results and leave this for further research.

## 9.3 Parameters and model selection

**Number of subspaces** The number of subspaces present in the data might be unknown to the user. Nevertheless, the observation of the eigenvalues of the graph Laplacian allows to infer the number of subspaces. In the ideal case, where $n$ disconnected components are present in the graph, there exist $n$ eigenvalues equal to 1. The difference with the $n + 1^{th}$ value called spectral gap is used to infer the number of subspaces [Von Luxburg, 2007].

$$Spectral\,gap = \sigma_{n+1} - \sigma_n \tag{9.18}$$

where $\sigma_n$ is the $n^{th}$ eigenvalue of the graph Laplacian. The spectral gap shows how well the components are intra-connected and separated between each other. The $n$ top eigenvalues will present a small decrease while the largest difference will still be found with the $n + 1^{th}$ component, when the components are not properly disconnected and present some connections between them.

The drawback of the spectral gap is its dependence to the magnitude of the eigenvalues. A more robust approach based on the eigenvalues and less affected by the range of magnitude of the eigenvalues is the spectral gap ratio [Lauer and Schnorr, 2009].

$$Spectral\,gap\,ratio = \frac{\sigma_n - \sigma_{n+1}}{\sigma_{n-1} - \sigma_n} \tag{9.19}$$

The spectral gap ratio evaluates the ratio between the $n^{th}$ and $n + 1^{th}$ difference and the $n - 1^{th}$ and $n^{th}$ difference.

**Kernel parameters** The bottleneck of the use of kernel functions is the selection of the function itself and of its parameters. In the experiments that follow, we consider the linear, polynomial and Gaussian (RFB) kernels, the latter having shown to be generic and flexible in numerous applications [Gao et al., 2013]. Once a valid kernel is chosen, its free parameters have to be tuned.

The selection of the kernel parameters, for instance the polynomial degree $p$ or the RBF bandwidth $\sigma$, can be based on the previously mentioned spectral gap or spectral ratio. Also it can be done using a cluster validity index which describes the result of spectral

clustering. The *Davies-Bouldin (DB)* index is used for this purpose:

$$R_{k,l} = \frac{S_{w,k} + S_{w,l}}{S_b} = \frac{\sum_{i \in \mathcal{M}_k} d(\mathbf{v}_i, \mu_k) + \sum_{i \in \mathcal{M}_l} d(\mathbf{v}_i, \mu_l)}{dist(\mu_k, \mu_l)} \tag{9.20}$$

$$DB = \frac{1}{n} \sum_k \max_{k \neq l} R_{k,l} \tag{9.21}$$

where $dist(\cdot)$ is the Euclidean distance between two samples, $\mu_k$ are cluster centers and $\mathbf{v}_j \in \mathcal{R}^{1 \times n}$ is the $j^{th}$ row of the matrix formed by the $n$ first eigenvectors of the graph Laplacian. $DB$ describes the ratio of the inner cluster variance over the minimal distances between clusters and is to be minimized in order to obtain a mapping towards a space where the $n$ subspaces are characterized by compact and well separated clusters in the spectral embedded space.

## 9.4 Subspace independence measures

### 9.4.1 Linear subspaces

A measure between two disjoint or independent subspaces is their *smallest principal angle* denoted by $\theta_{1,2}$

$$\cos \theta_{1,2} = \max_{\mathbf{v}_1 \in \mathcal{S}_1, \mathbf{v}_2 \in \mathcal{S}_2} \frac{\mathbf{v}_1^T \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2} \tag{9.22}$$

where $\mathbf{v}_i$ are the basis vectors of the corresponding subspace $i$. Two subspaces that are said to be dependent, are actually sharing some dimensions and have a null smallest principal angle.

A richer measure is the normalized affinity between two subspaces, based on all their principal angles

$$\text{aff}(\mathcal{S}_1, \mathcal{S}_2) = \sqrt{\frac{\sum_i \cos^2 \theta_i}{d_1 \wedge d_2}} \tag{9.23}$$

where the angles $\theta_i$ are between shared dimensions $d_1 \wedge d_2$, with $d_1, d_2$ the dimensions spanned by subspace $1, 2$ respectively.

Another notion is the subspace incoherence defined as the separation of a set of samples from a subspace to the others. This has been studied in detail by [Soltanolkotabi and Candes, 2012, Wang et al., 2013a] based on geometric notions such as dual directions and different data distribution models. In the general case, with data drawn at random, the incoherence measure is proportional to the principal angles between the subspaces. The incoherence is close to 1 for small principal angles and close to 0 for orthogonal subspaces, as defined by Soltanolkotabi and Candes [2012].

Both SSC and LRR have shown to recover perfectly independent subspaces [Elhamifar and Vidal, 2013, Liu et al., 2013]. The recovery of disjoint subspaces [Elhamifar and Vidal, 2013], and of subspaces with potential overlap between them [Soltanolkotabi and Candes, 2012] has been studied. SSC can deal with outliers under mild assumptions [Soltanolkotabi and Candes, 2012] and guarantees recovery under noise and corruption [Wang and Xu, 2013]. LRR can deal with large outliers [Liu et al., 2012] and its graph connectivity is typically dense [Liu et al., 2013], but it requires independence. Whereas, the dense connectivity may not be satisfied for SSC when dimensions are greater than 3 [Nasihatkon and Hartley, 2011].

### 9.4.2 Non-linear subspaces

In the non-linear case, the notions and associated measures of subspaces dependence, incoherence, and representation are different.

For instance, the *smallest principal angle* between two subspaces in RKHS can be retrieved by

$$\cos\theta_{1,2} = \max_{\mathbf{x}_1\in\mathcal{M}_1,\mathbf{x}_2\in\mathcal{M}_2} \frac{\Phi(\mathbf{x}_1)^T\Phi(\mathbf{x}_2)}{\|\Phi(\mathbf{x}_1)\|_2\|\Phi(\mathbf{x}_2)\|_2} \overset{RBF}{=} \max_{\mathbf{x}_1\in\mathcal{M}_1,\mathbf{x}_2\in\mathcal{M}_2} K(\mathbf{x}_1,\mathbf{x}_2) \tag{9.24}$$

where the kernel $K$ is a kernel inducing a normalized feature space, e.g. the RBF kernel. In opposition to the linear case, the measure given by (9.24) will not be informative of the independence of two non-linear subspaces in many cases. In the specific case of the RBF kernel, the data samples are mapped to an hypersphere surface of potentially infinite dimension. The RBF bandwidth controls how the samples are distributed at the surface. A small bandwidth will signify that samples are all orthogonal to each other, each one along its own dimension. Whereas a large bandwidth will move the samples close to each other at the surface of the hypersphere and become more dependent between each other. Therefore, the ideal RKHS will set the samples in a way maximizing the independence between subspaces, facilitating their recovery.

Therefore, the independence of two manifolds $\mathcal{M}_1, \mathcal{M}_2$ can be evaluated by considering their sets of samples $\mathbf{X}^1$ and $\mathbf{X}^2$ as two random variables drawn from two different distributions in $\mathcal{H}$. The Maximum Mean Discrepancy (MMD) [Gretton et al., 2012] evaluates the difference between the mean map of the two sets. The empirical unbiased

MMD estimate is given by

$$\begin{aligned}\mathrm{MMD}^2(\mathbf{X}^1, \mathbf{X}^2) &= \|\mu_{\mathcal{M}_1} - \mu_{\mathcal{M}_2}\|^2_{\mathcal{H}} \\ &= \frac{1}{N_1(N_1-1)} \sum_{i\neq j, i\in\mathbf{X}^1} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{N_2(N_2-1)} \sum_{i\neq j, i\in\mathbf{X}^2} K(\mathbf{x}_i, \mathbf{x}_j) \\ &- \frac{2}{N_1 N_2} \sum_{i\in\mathbf{X}^1, j\in\mathbf{X}^2} K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{9.25}$$

with $N_1, N_2$ the number of samples in each set. MMD is large if the subspaces are independent and close to 0 if dependent, meaning potentially issued from the same probability distribution function.

## 9.5 Experiments

In this Section, we first perform experiments on standard real-world datasets generally used to evaluate subspace and manifold clustering methods. These experiments aim at grouping images of objects or faces by subjects with varying illuminations, poses or expressions. Then, we perform experiments on different hyperspectral datasets for unsupervised classification and discuss their results.

### 9.5.1 Experiment 1: Face and object clustering

Given a set of images of objects or faces from different subjects, we aim at clustering the images of the same subject together. We discuss the results obtained on different real-world datasets showing either linear or non-linear characteristics: clustering of faces under certain illuminations (linear), expressions or poses (non-linear), clustering of objects under various poses (non-linear) and recognition of hand-written digits (non-linear). We compare the performance of our non-linear approach with state-of-the-art algorithms, namely: SSC [Elhamifar and Vidal, 2013], LRSC [Vidal and Favaro, 2013], LRSSC [Wang et al., 2013a], SMCE [Elhamifar and Vidal, 2011] and STSC [Zelnik-Manor and Perona, 2004].

#### Experimental setup

In these experiments, we have investigated the following three non-linear approaches

1. KLRSSC which seeks a low-rank and sparse representation in a RKHS with the low-rank-sparsity trade-off parameter $\lambda = 10^{[-4,-3,-2,-1,0,1,2]}$.

2. KLRSC which seeks a low-rank representation in a RKHS.

Table 9.1: Selected parameters for the different databases and methods.

| Dataset | method | SSC | LRSC | LRSSC | KSSC | | KLRSC | | KLRSSC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Poly | RBF | Poly | RBF | Poly | RBF |
| Yale B. E. | $\gamma/\sigma/d$ | 20 | 10000 | - | 2 | 2 | 2 | 2 | 2 | 2 |
| | $\tau/\lambda$ | - | 0.075 | 0.1 | - | - | 3 | 0.5 | 0.1 | 0.1 |
| | $\alpha_z$ | 20 | 0.0075 | 20 | 20 | 20 | 20 | 20 | 50 | 50 |
| AR | $\gamma/\sigma/d$ | 20 | 10000 | - | 2 | 9 | 2 | 2 | 3 | 5 |
| | $\tau/\lambda$ | - | 0.075 | 0.01 | - | - | 1.5 | 3 | 0.01 | 0.01 |
| | $\alpha_z$ | 20 | 0.075 | 20 | 20 | 20 | 100 | 10 | 20 | 20 |
| AT&T | $\gamma/\sigma/d$ | - | 0 | - | 4 | 1.1 | 2 | 1.1 | 4 | 1.2 |
| | $\tau/\lambda$ | - | 1 | 10 | - | - | 1.5 | 1.5 | 1 | 1 |
| | $\alpha_z$ | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Sheffield | $\gamma/\sigma/d$ | - | 0 | - | 5 | 0.5 | 4 | 0.3 | 3 | 0.3 |
| | $\tau/\lambda$ | - | 1 | 0.1 | - | - | 3 | 1 | 0.001 | 0.001 |
| | $\alpha_z$ | 20 | 100 | 5 | 20 | 20 | 20 | 50 | 0.5 | 0.5 |
| COIL20 | $\gamma/\sigma/d$ | - | 0 | - | 2 | 0.5 | 3 | 0.2 | 4 | 0.3 |
| | $\tau/\lambda$ | - | 1.5 | 0.1 | - | - | 1.5 | 1 | 0.001 | 0.001 |
| | $\alpha_z$ | 20 | 1 | 2 | 20 | 20 | 1 | 50 | 0.5 | 0.5 |
| USPS 1&2 | $\gamma/\sigma/d$ | - | 0 | - | 3 | 0.5 | 2 | 0.4 | 3 | 0.5 |
| | $\tau/\lambda$ | - | 3 | 1 | - | - | 3 | 0.5 | .1 | 1 |
| | $\alpha_z$ | 20 | 20 | 1 | 20 | 20 | 100 | 100 | 1 | 20 |

3. KSSC which seeks a sparse representation in a RKHS.

For all three, the noise penalty is searched over the range $\alpha_z = [1, 5, 10, 20, 50, 100]$, and the Lagrangian constraints are set either fix or equal to the noise penalty: $\mu_2 = \mu_3 = [20, \alpha_z]$. Two different kernels are used: polynomial and RBF. The kernel free parameters are selected in the ranges $p = [2, 3, .., 10]$ and $\sigma = [0.1, 0.2, .., 1.6] \cdot \sigma_0$, where $\sigma_0$ is the median distance among the data samples. These parameters are tuned with 100 random projections by finding the optimal missrate and kept fixed for the other number of dimensions. We observed no significant differences in the optimal parameters chosen across the different number of dimensions (random projections). We present in Figure 9.3 the correlation between different measures (*DB index, spectral gap and spectral gap ratio*) and the clustering results. For the comparisons with state-of-the-art algorithms we used the settings reported by the respective authors, and their code if provided. For SSC, the noise penalty is set to $\alpha_z = 20$ [Elhamifar and Vidal, 2013]. The neighbourhood for STSC [Zelnik-Manor and Perona, 2004] and SMCE is set to $k = 7$ and its noise parameter $= 10$ [Elhamifar and Vidal, 2011]. For Yale B Extended and AR database, both SSC and LRSC exploit the noise term for sparse outlying entries.

The different datasets used in the experiments are presented in Appendix B.3.

Table 9.2: Clustering missrate results averaged over all pairs of subjects.

| | LINEAR | | | NON-LINEAR | | PROPOSED | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | SSC | LRSC | LRSSC | SMCE | STSC | KSSC | | KLRSC | | KLRSSC | |
| kernel | | | | | | Poly | RBF | Poly | RBF | Poly | RBF |
| Yale B. E. | | | | | | | | | | | |
| *mean* | **.024** | .041 | .039 | .173 | .329 | .093 | .103 | .176 | .490 | .173 | .199 |
| *median* | **.000** | .039 | .039 | .047 | .445 | .039 | .031 | .109 | .492 | .117 | .109 |
| AR | | | | | | | | | | | |
| *mean* | .205 | .139 | .110 | .325 | .256 | .257 | .239 | .126 | .099 | .142 | **.092** |
| *median* | .214 | .036 | **.000** | .286 | .214 | .214 | .161 | .036 | **.000** | .036 | **.000** |
| AT&T | | | | | | | | | | | |
| *mean* | .041 | .056 | .066 | .077 | .058 | .068 | .052 | .055 | .042 | .055 | **.038** |
| *median* | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** |
| Sheffield | | | | | | | | | | | |
| *mean* | .277 | .243 | .249 | .064 | .062 | .259 | .095 | .192 | **.057** | .282 | .080 |
| *median* | .333 | .252 | .277 | **.000** | **.000** | .280 | **.000** | .149 | **.000** | .357 | **.000** |
| COIL20 | | | | | | | | | | | |
| *mean* | .048 | .057 | .053 | **.001** | .010 | .041 | .011 | .040 | .006 | .065 | .009 |
| *median* | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | **.000** | .021 | **.000** |
| USPS 1&2 | | | | | | | | | | | |
| *mean* | .121 | .160 | .091 | .098 | .091 | .108 | .081 | .499 | **.068** | .135 | .088 |

### Results

Experiments aim at performing face or object clustering using the different clustering methods. In each dataset, the clustering is performed for all pairs of subjects or objects and the results averaged. Additionally, experiments with reduced input dimensions based on random projections are averaged over three random runs. The average and median missrate over all the pairs are evaluated for each dataset. The selected parameters, such as the kernel parameters (polynomial degree and RBF bandwidth), the weight $\gamma$ for the sparse noise term in SSC and LRSC, the low-rank representation parameter $\tau$ for (K)LRSC, the low-rank sparsity trade-off parameter $\lambda$ for (K)LRSSC and the noise penalty $\alpha_z$, are reported in Table 9.1. The results without random projections are summarized in Table 9.2 and the results for different number of random projections, ranging between 10 and 200, are presented in Figure 9.2 for the RBF kernel. The results with the polynomial kernel are shown in Appendix C Figure C.5.

**Yale B Extended** The Yale B Extended face dataset contains frontal photos of 38 subjects under 64 different illuminations [Georghiades et al., 2001]. The cropped face images, resized to $42 \times 48$ pixels, of the first 10 persons are used. For this dataset, it has been proven that the images of a single person lie close to a 9-D linear subspace.

In this frontal face dataset with illuminations variations, both SSC and LRSC exploit the term for sparse outlying entries. The selected RBF kernel bandwidth is relatively high, between $2 \cdot \sigma_0$ and $9 \cdot \sigma_0$ for all three non-linear methods. These large bandwidth values confirm the close-to-linear type of data. The optimal trade-off parameter found for LRSSC and KLRSSC is 0.1, which is a balanced state exploiting both the sparse and low-rank properties. The best results are obtained with SSC followed by the LRSSC and then the LRSC. It is important to emphasize that LRSSC does not have an additional sparse noise term specific to outlying entries and still achieves a result comparable to LRSC. The combination of the sparse and low-rank properties provides a certain robustness to the noise present in this dataset. The advantage of the low-rank characteristic is observed at lower dimensions with a stable recovery until 30 random projections for both LRSC and LRSSC. SSC has a missrate which increases for reduced input dimensions. Under 100 random projections it performs worse than kernel approaches. Both RBF and polynomial kernels produces similar results. This was expected from the linear nature of this dataset. The polynomial kernel of degree 2 and the RBF with a large bandwidth are both closer to a linear mapping than higher degree polynomials for instance. The linear kernel is atually the most appropriate with this data (e.g. LRSSC result). The subspaces formed by each subject are extremely close to each other with very small principal angles [Elhamifar and Vidal, 2013], but they are still independent enough to be able to recover them via the low-rank property (LRSC) until very reduced input dimensions ($\sim 20$).

**AR Faces** The *AR* face database contains 4000 frontal images of 126 subjects. Each subject is photographed in 26 images taken in two sessions [Martinez and Benavente, 1998]. In comparison to the Yale B Extended database, the AR database contains variations of facial expressions, as well as of illumination. We consider the 20 first subjects and use grayscale images cropped and resized to $55 \times 40$ pixels. The images with occlusions (for instance scarfs and sun glasses) have been left aside, resulting in 14 images per subject.

In this frontal face dataset with varying illuminations and expressions, again both SSC and LRSC exploit the sparse noise term for outlying entries. The selected RBF kernel bandwidth is between $2 \cdot \sigma_0$ and $9 \cdot \sigma_0$ for all three non-linear methods, confirming here also the close-to-linear type of data as for the Yale B Extended dataset. The best results are obtained with the KLRSSC and RBF kernel, followed by the KLRSC. These two approaches accommodate well this dataset which has under-sampled subspaces (only 10 image per subject). Here as well, the low-rankness property allows to more robustly segment the subspaces than with the sparse property alone. The SMCE and STSC do not perform well on this dataset, which does not fulfil their assumptions on the nearest neighbours. A decrease in performance is observed under 100 random projections, due to the small number of samples per manifold.

**AT&T Faces** The *AT&T* (ex-ORL) database of faces contains 40 subjects in frontal position with tolerance to certain movements [Samaria and Harter, 1994]. The images contain 10 facial expressions and sometime show occlusions (for instance glasses). Images of the 20 first subjects are used after being cropped and resized to $32 \times 32$ pixels.

The best results are obtained also by KLRSSC and the RBF kernel. The low-rank sparsity trade-off parameter $\lambda$ is well balancing the two with $\lambda = 1$. The linear SSC performs better than the other linear approaches. The sparse representation and the sparse error term allow to handle the facial details and the small pose variations. The LRSC does not behave similarly, since worse results are obtained when the sparse error term is added. The STSC results are among the best while SMCE does not perform well, since the local affine relationship requirement is not fulfilled. This is expected with under-sampled dataset containing some non-linear variations.

**Sheffield (ex-UMIST)** The *Sheffield* (ex-UMIST) Face database contains 20 subjects under poses from frontal to lateral (0 to 90°) position [Graham and Allinson, 1998]. Each subject has between 19 and 48 images. The images are cropped and rescaled to $28 \times 23$ pixels. The database contains no occlusions nor variations in expressions or illumination. This dataset presents non-linear characteristics with faces under varying poses, which is an ideal case for manifold clustering evaluation. The independence between the pairs of subjects is illustrated in Figure 9.1. Sample images of the most and least dissimilar subjects are also presented. Among the MMD of the different pairs of subjects, the clustering failures for the KLRSSC seem to occur with MMD under $\approx 0.25$.

The kernel RBF bandwidths selected are around $\sigma = [0.3, 0.5]$, since the data is highly non-linear. The trade-off parameter is in favour of the low-rank property ($\lambda = 0.001$) in this dataset, the manifolds being independent enough. This is confirmed by the best results obtained using the KLRSC and the RBF kernel. The KLRSSC results follow close-by together with the SMCE and STSC. The two latter have good performances since their neighbourhood assumptions hold with these better sampled manifolds. All kernel approaches are stable until 30 random projections.

**COIL Database** The COIL-20 database [S. A. Nene and Murase, 1996] contains 1440 images of 20 objects rotated over 360°. Each object is represented in 72 images. The processed images are cropped and rescaled to $32 \times 32$ pixels. This object database is an interesting non-linear dataset. The best results are obtained with SMCE, since the assumption of locally affine relationship between samples holds, considering the high sampling of the manifold. Similar results are obtained by the KLRSC, STSC and KLRSSC. The low-rankness property is taking over the sparsity with $\lambda = 0.001$, the manifold being independent. The linear approaches cannot achieve comparable results with this dataset, even if locally linear.

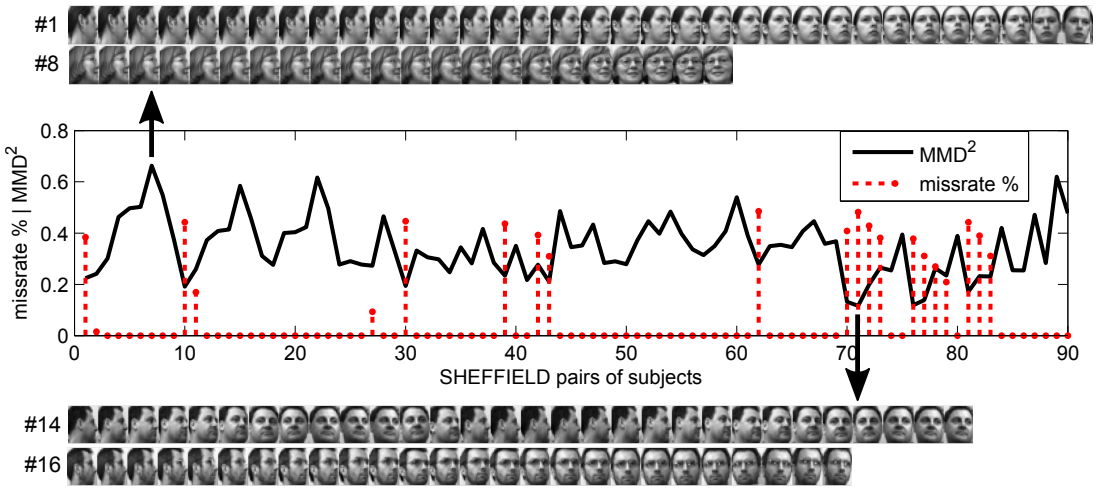



Figure 9.1: Maximum Mean Discrepancy (MMD) for each pair of subjects in the Sheffield face database with a RBF kernel ($\sigma = 0.5 \cdot \sigma_0$). The relation between the non-linear subspace independence (MMD) and the non-linear subspace clustering missrate can be well observed. Two pairs of subjects are presented: the two most dissimilar subjects with highest MMD and the two most similar subjects with lowest MMD.

**USPS Digits** The *USPS hand-written digits recognition* dataset are images of digits from the US Postal Service [Hull, 1994]. We consider a subset composed of the digits 1 and 2, each one having 1100 samples. The best results are obtained by the KLRSC followed by the KSSC and KLRSSC. The sparsity required is more important with $\lambda = 1$ for KLRSSC RBF and LRSSC. The results are stable above 40 random projections. The linear methods are performing worse on this non-linear dataset.

**Discussion** Across the different experiments, the following observations can be made: the KLRSSC and its two particular extension, namely the KSSC and KLRSC, generalize well to both linear and non-linear datasets and are more robust to reduced input spaces and insufficient subspace sampling (AR or AT&T), see Figure 9.2. The flexibility of having a trade-off between low-rankness and sparsity is advantageous to adapt to datasets with different properties. Moreover, this trade-off appears to be important with datasets containing strong variations like AR, AT&T and Yale B Extended. The self-expressiveness property is efficient alone but produces better representation graphs when combined with the low-rank property. The gain in subspace intra-connectivity due to the low-rank property is not counter-acting the clustering, recalling that spectral clustering can handle additional weak connections between different subspaces. Regarding the two competing non-linear approaches, the SMCE shows more flexibility than STSC mainly because the latter uses a fixed neighbourhood. Nonetheless, in AR Faces SMCE does not perform well. This effect is due to the subspaces lying close to each other and having non-linear variations due to expressions variations, which make them even closer or intersecting. We

believe that in such situation the imposed affine relationship in SMCE captures samples from other subspaces.

**Reduced input space** The kernel approaches are more robust to reduced input space. Stable results can be obtained with input dimensions as low as 40. This is equivalent to only about 2% of the original data dimensions. This is advantageous for speeding up computations and suggests also that data corruption like low-resolution may be better handled using non-linear kernel approaches. With Yale B Extended, the linear approaches exploiting the low-rankness property (LRSC, LRSSC) are robust to reduced input dimensions, whereas the SSC results are getting worse more quickly (see Figure C.4 top row). In opposition, with USPS1&2 the LRSC is getting worse quickly while SSC and LRSSC remains stable.

**DB index and spectral gap criteria** The different measures such as the *Davies-Bouldin* index, the spectral gap and the spectral gap ratio, have been proposed to qualify the resulting spectral embedding helping in the parameter selection.

The *DB* index is useful for KSSC where it correlates with the missrate. When used with the low-rank property, it has a certain bias toward low $\sigma$ values as it can be observed in Figure 9.3 second row and therefore cannot be used for parameter selection. This can be understood by the fact that KSSC is over-segmenting at too low $\sigma$ which produces a certain dispersion of the clusters in the spectral embedding. A similar observation can be made for the spectral gap, third row in Figure 9.3, where it correlates with the missrate only for KSSC and biased towards low $\sigma$ otherwise. Finally, in Figure 9.3 bottom row we can observe that the spectral gap ratio is not affected by this bias since it involves only the ratio of the eigenvalues. The spectral gap ratio is largest around $0.4 \cdot \sigma_0$ and starting from $\lambda = 10$ and above. It better correlates with the missrate when a certain amount of sparsity is imposed. Despite its slight instability, it indicates combination of $\sigma$ and $\lambda$ parameters leading to low missrate and thus appears to be a reliable indicator of the clustering quality for the selection of the parameters.

Some additional details on these experiments are reported in Appendix C.1.

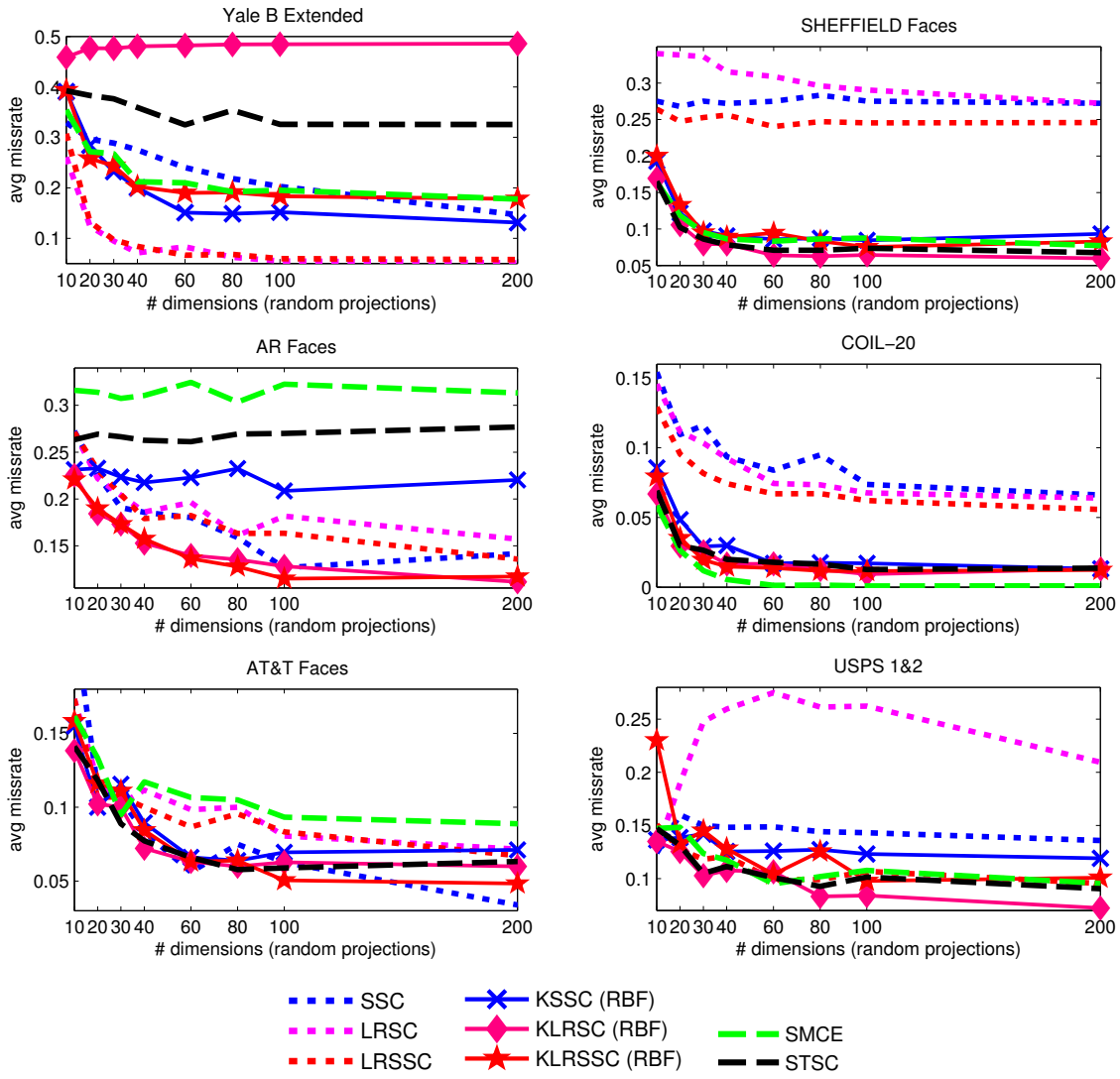



Figure 9.2: Average missrate (the lower the better) over all pairs of subjects for various number of random projections for 3 different datasets: Yale B Extended Faces, AR Faces, AT&T Faces, Sheffield Faces, COIL-20 and USPS-1&2. The left column shows the results for the non-linear methods with the RBF kernel.

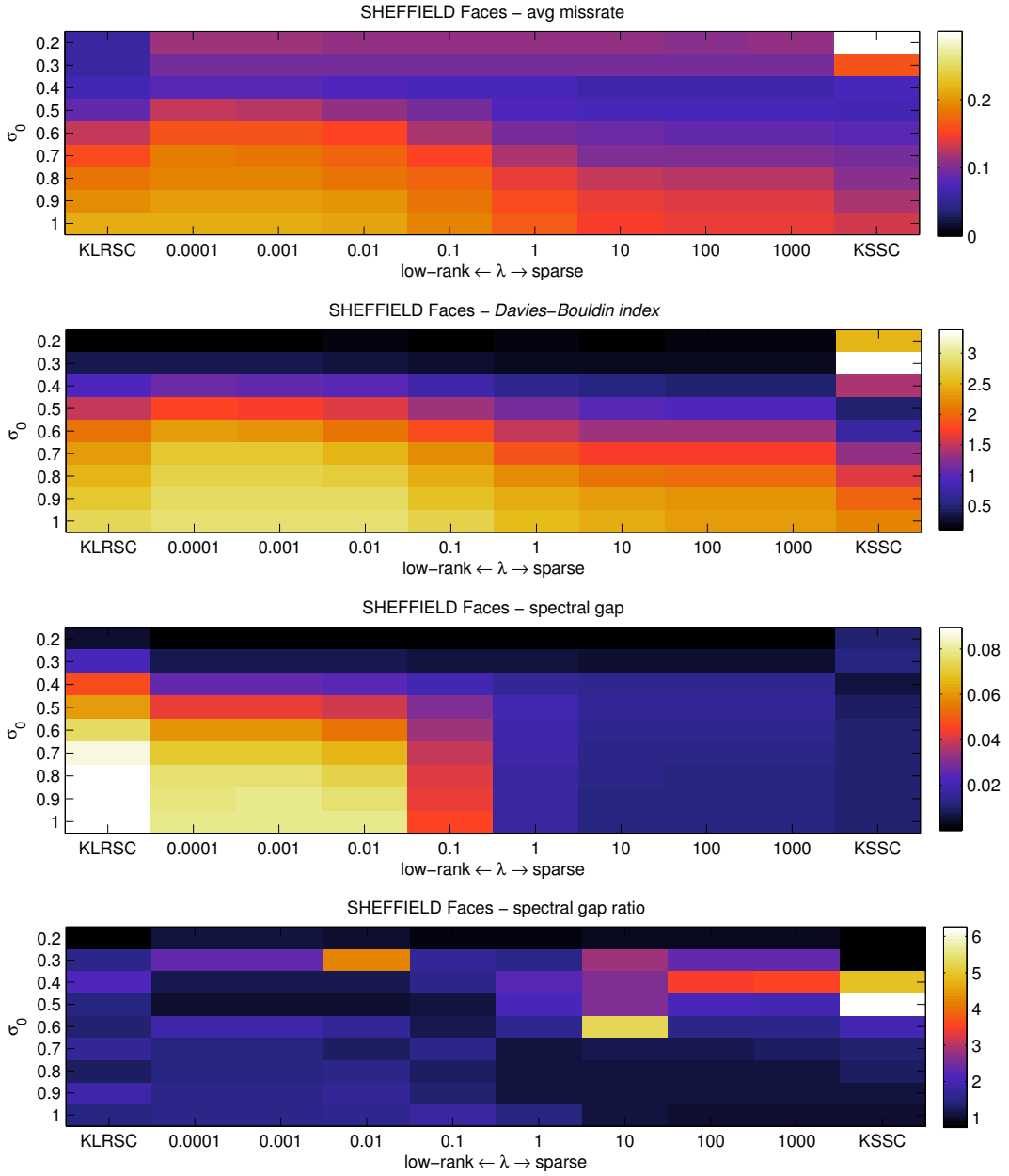



Figure 9.3: Different values of $\lambda$ and $\sigma$ in Sheffield Faces with (a) missrate, (b) *DB* index, (c) spectral gap and (d) spectral gap ratio, averaged over all pairs of subjects. The other parameters are $\alpha_z = \mu_2 = \mu_3 = 20$.

### 9.5.2 Experiment 2: Unsupervised hyperspectral image classification

We perform experiments aiming at retrieving the manifolds present in different hyperspectral scenes from the ROSIS optical sensor. Details on the two scenes can be found in Appendix B.2.

#### Experimental setup

In order to properly evaluate the recovery of different manifolds in the images, we use the available groundtruth to provide a subset of pixels to the different methods. This allows to control the experiment to be able to assess the recovered structures. The subset $\mathbf{X}$ is formed by $n_s$ pixels randomly sampled, with $n_s = 1000$ for *Pavia DFC* and $n_s = 990$ for *Pavia center*. The number of clusters is given to all algorithms for fair comparisons. This avoids a bias against weakly represented classes with spectral clustering methods and facilitates the comparison of the algorithms.

We use the proposed low-rank and sparse representation with a linear kernel – LRSSC, thus assuming linear subspaces –, and with a RBF kernel – KLRSSC, thus assuming smooth manifolds. The low-rank or sparse representation alone are also used for comparisons, being the KLRSC and KSSC with the RBF kernel. The linear LRSC and SSC are also evaluated for comparisons. They are run on the subset $\mathbf{X}$ and the remaining pixels of the image are assigned using the collaborative representation of (9.16).

Moreover the following methods are used for comparisons: i) $k$-means (KM), ii) kernel $k$-means (KKM), which both assign the remaining pixels to their closest center in the input space or feature space respectively [Volpi et al., 2012], iii) the sparse manifold clustering and embedding (SMCE) [Elhamifar and Vidal, 2011], iv) the spectral clustering (SC) with a $k$-NN graph and the weights from the heat kernel and its v) self-tuning extension [Zelnik-Manor and Perona, 2004] which are followed by the assignment of the remaining pixels with 1-NN classification.

The quality and stability of the results are assessed with the adjusted Rand Index (aRI), Normalized mutual information (NMI) and the Cohen's $\kappa$ statistic on the labeled groundtruth after finding the best match between clusters and classes using the Hungarian matching method (see Appendix A.1).

The parameters are tuned in the following ranges: $\lambda = [10^{-6}, ..., 10^{6}]$, $\sigma = [0.5, .., 10] \cdot \sigma_0$, where $\sigma_0$ is the median distance among pixels, $\mu = 1000$, $\gamma = 10^{-6}$ and $k = 7$ for SC and STSC and the neighbourhood of SMCE equal 10.

| | *Pavia DFC* | | | *Pavia center* | | |
|---|---|---|---|---|---|---|
| Method | $\kappa$ | aRI | NMI | $\kappa$ | aRI | NMI |
| KM | $.50 \pm .02$ | $.51 \pm .01$ | $.61 \pm .02$ | $.50 \pm .05$ | $.50 \pm .02$ | $.66 \pm .02$ |
| KKM | $.71 \pm .17$ | $.65 \pm .12$ | $.72 \pm .08$ | $.58 \pm .10$ | $.55 \pm .06$ | $.70 \pm .04$ |
| SMCE | $.14 \pm .04$ | $.05 \pm .03$ | $.07 \pm .04$ | $.37 \pm .03$ | $.27 \pm .04$ | $.47 \pm .04$ |
| SC | $.85 \pm .01$ | $.77 \pm .01$ | $.81 \pm .01$ | $.70 \pm .03$ | $.63 \pm .01$ | $.76 \pm .01$ |
| STSC | $.85 \pm .02$ | $.77 \pm .04$ | $.81 \pm .03$ | $.69 \pm .03$ | $.64 \pm .03$ | $.77 \pm .02$ |
| SSC | $.62 \pm .01$ | $.64 \pm .03$ | $.71 \pm .02$ | $.46 \pm .06$ | $.33 \pm .06$ | $.42 \pm .07$ |
| LRSC | $.86 \pm .02$ | $.76 \pm .03$ | $.78 \pm .02$ | $.59 \pm .07$ | $.46 \pm .06$ | $.57 \pm .05$ |
| LRSSC | $.62 \pm .01$ | $.63 \pm .02$ | $.71 \pm .01$ | $.51 \pm .05$ | $.37 \pm .06$ | $.46 \pm .06$ |
| KSSC | $.91 \pm .01$ | $.84 \pm .01$ | $.85 \pm .01$ | $.73 \pm .03$ | $.68 \pm .02$ | $.78 \pm .01$ |
| KLRSC | $.92 \pm .01$ | $.85 \pm .02$ | $.86 \pm .01$ | $.82 \pm .07$ | $.74 \pm .04$ | $.82 \pm .02$ |
| KLRSSC | $\mathbf{.93 \pm .01}$ | $\mathbf{.87 \pm .02}$ | $\mathbf{.87 \pm .02}$ | $\mathbf{.86 \pm .01}$ | $\mathbf{.77 \pm .01}$ | $\mathbf{.84 \pm .01}$ |

Figure 9.4: The $\kappa$ statistic, adjusted Rand Index (aRI) and Normalized Mutual Information (NMI) over 10 random runs for the *Pavia DFC* image with subset size $n_s = 1000$.

**Results**

The Figure 9.4 presents the numerical performances for the unsupervised classification of *Pavia DFC* and *Pavia center* averaged over 10 independent experiments for the different methods. Figure 9.7 presents the resulting land-cover classification maps for both datasets and some of the methods. In both datasets, the KLRSSC outperforms the other methods in terms of all the statistics reported ($\kappa$, aRI and NMI) and is followed closely by the KLRSC, indicating a benefit in adding some sparsity to the low-rank constraint. When using the sparsity only, the KSSC, the performances decrease ($-.13$ $\kappa$ in *Pavia center* w.r.t KLRSSC) due to a lack of intra-class connections over-partitioning each class and leading to miss-classifications. The performances vary slightly with the trade-off parameter $\lambda$ with the maximum clustering accuracy is obtained for small $\lambda$, indicating a very small contribution of the sparsity w.r.t. the low-rank representation.

On the side of the linear methods, the LRSC achieves the best results compared to SSC and LRSSC. Seeking a low-rank representation matrix is a more flexible constraint than the sparse representation and provides spectral clustering with a matrix where the connected components can be well distinguished. The limitation of LRSSC and the other linear methods can be appreciated in Figure 9.7 for *Pavia Center*: the result of the unsupervised classification is poor (.51 $\kappa$) with the water class assigned to too many pixels. This is due to the low magnitude of water pixels, which can span linear subspaces in every direction using large coefficients and will produce large edge weights in the graph. The SC is performing close to our proposed approach. This is not surprising since SC relies on a nearest-neighbour graph weighted by the heat kernel whereas our KLRSSC finds representative samples which are local neighbours with the RBF kernel. Even if both rely on local neighbourhood, the KLRSSC has the advantage of letting the

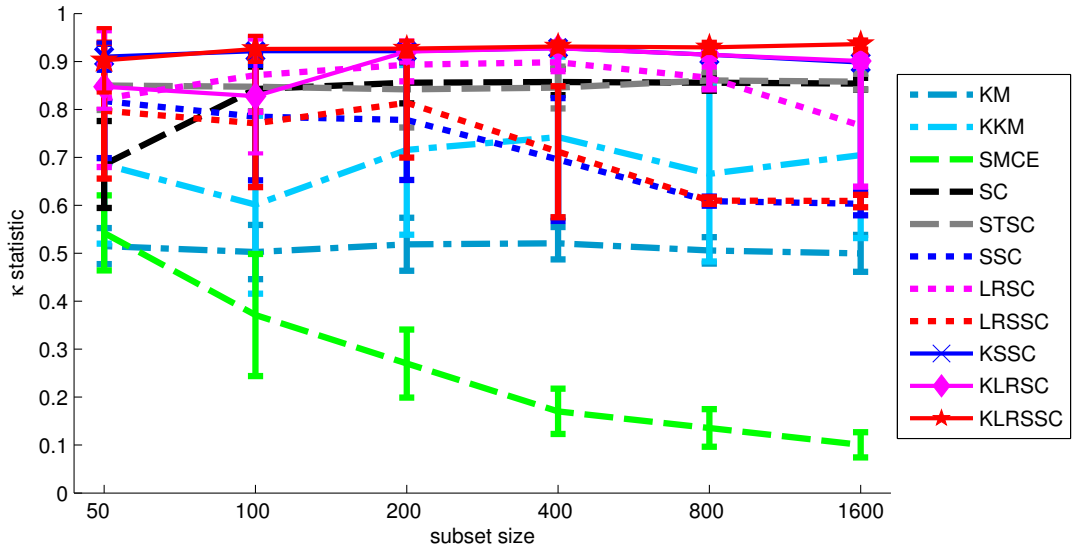



Figure 9.5: The $\kappa$ statistic over 10 random runs for *Pavia DFC* with different subset size ($n_s = [50, .., 1600]$).

neighbourhood adapt to the data, while in SC it is fixed. For manifolds that are close to each other, wrong neighbours can be picked by SC, as for some pixels in the river in *Pavia center* which are attributed to the *asphalt* manifold lying very close to the *water* manifold. The KM does not provide satisfying solutions and performs more than 30% worse (in $\kappa$) compared to KLRSSC. The KKM is also performing worse than SC and KLRSSC. It does not recover properly the multiple manifolds, since it assumes clusters in the feature space and not subspaces. It therefore produces unstable solutions, as shown by the large standard deviations in Figure 9.5.

Figure 9.5 illustrates the $\kappa$ results for different subset sizes $n_s$ over 10 runs. KLRSSC is very accurate at recovering the manifolds already for subset of size $n_s = 50$, confirming the well-posed assumption of multiple manifolds. The other non-linear KLRSC and KSSC stabilizes above $n_s = 200$. The linear LRSC follows closely while the SSC and LRSSC are decreasing above $n_s = 200$. With very small subset, the data structure is less clear and can satisfy more methods while with larger subset the linear assumption does not hold any more. The KKM accuracy has an important variability across the different subset sizes showing difficulties to retrieve the multiple manifolds. The decreasing performances of the SMCE is due to the neighbourhood limit which should probably be increased together with the increasing subset size. As a matter of fact, the SMCE representation of the data by affine subspaces is too flexible for retrieving accurately the different class manifolds in hyperspectral images.

**Low-rank and sparsity trade-off** The main advantage of the proposed formulation for non-linear subspace clustering reside in the fact that it exploits the self-expressiveness of the samples from the same subspace and injects subspace intra-connectivity by the additional low-rank constraint. The low-rank representation is very effective for clustering since it is equivalent to the ideal affinity matrix: fully connected within the manifolds and disconnected between them. However, the low-rank constraint is sensitive to manifolds dependence. Therefore, the sparsity constraint helps when independence fails between manifolds. In Figure 9.6, the positive symmetric and normalized representation coefficients corresponding to the graph edge weights ($W = |C| + |C|^T$) are presented for *Pavia DFC* and the KLRSC, KLRSSC and KSSC methods as well as their linear counterpart, the LRSC, LRSSC and SSC. Also graphical confusion tables which represents the corresponding samples and their label predicted by clustering allow to better see how spectral clustering handles the different coefficient matrices. In Figure 9.6 (a) the linear LRSC groups the classes *roads* and *water* together which produces out-of-diagonal non-null blocks in $\mathbf{C}$. This due to the low-rank constraint which seeks dependent columns of the matrix $\mathbf{C}$. However spectral clustering can handle these smaller coefficients resulting only in few miss-classifications with other classes. LRSC provides the best result among the three linear method ($\kappa = .86$) demonstrating the benefit of representation matrix with dense connections within classes. The addition of the sparse constraint in LRSSC remove the off-block-diagonal coefficients. It finds a compromise between blocks of coefficients and sparse representation which leaves only a few non-null out-of-diagonal coefficients. However, spectral clustering produces more miss-classifications in the class *roads* which is too sparsely connected. When only the sparsity constraint is used (SSC), too few non-zero coefficients remains which produces miss-classifications in the clustering results as observed in (c). The non-linear approaches better retrieve the manifolds structure with representation matrix better filled inside each class block. The KLRSC presents a near-ideal block-diagonal structure. However, miss-classifications remain between the class *roads* and *building*. The KLRSSC eliminates certain of the undesired off-block-diagonal coefficients which reduces the number of miss-classifications between classes *roads* and *building*. The KSSC produces a neat block-diagonal matrix structure, but reduces the intra-manifold connections and induces certain miss-classifications.

**High dimensional input space** In practice the manifolds are all embedded in a very correlated space meaning that many dimensions are not accounting for the manifolds variations. Therefore, we have investigated to first reduce the dimensionality of the data by keeping the 15 principal components, which account for 99.8% of the data total variance. The result obtained where similar than without any prior reduction of the dimensions for our proposed methods. This demonstrates their ability to handle high-dimensional data as well as to analyze correctly the remaining non-linear variations in reduced dimensions.

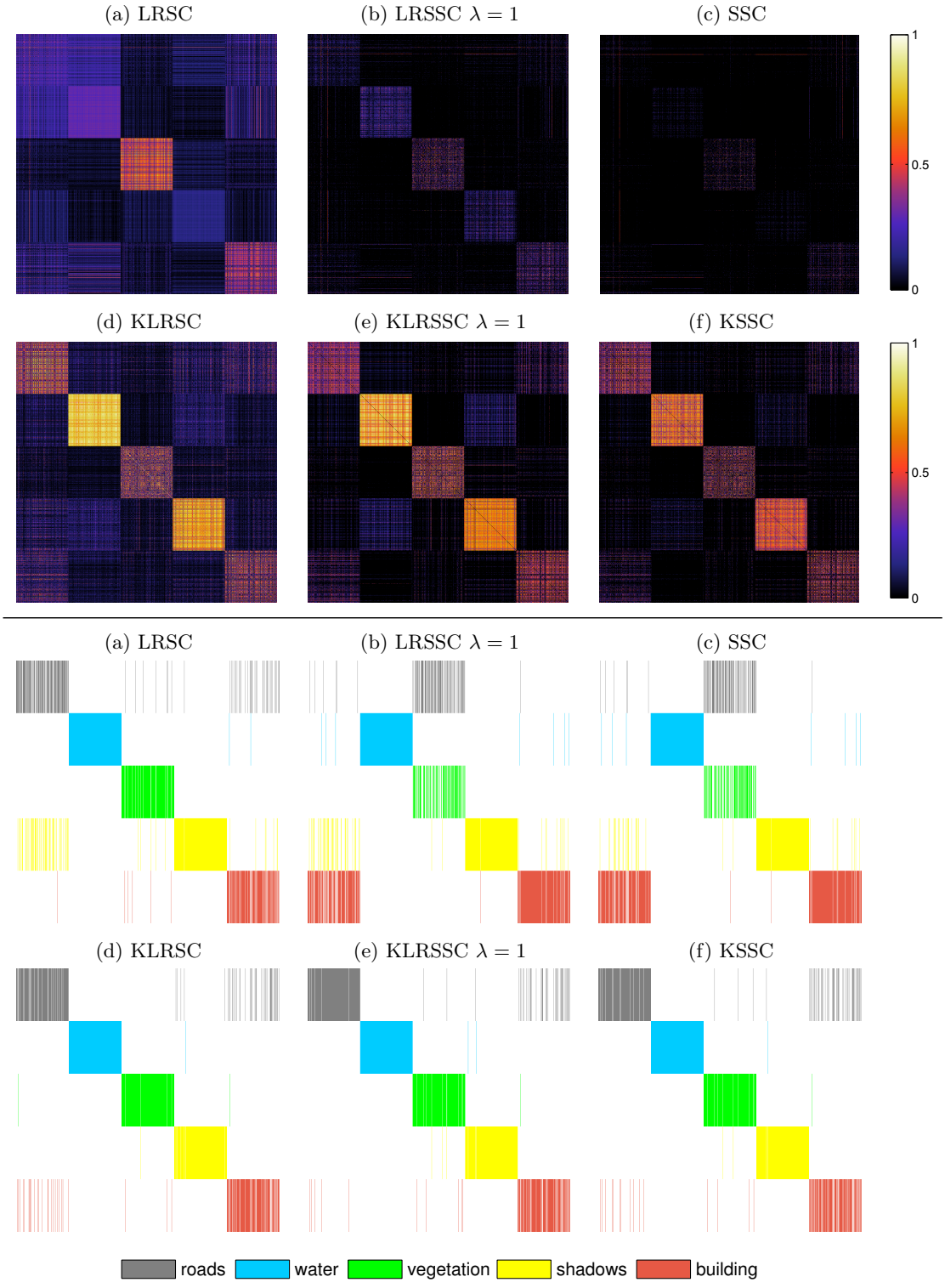



Figure 9.6: **Top part:** graph weights $W = |C| + |C|^T$. The samples from *Pavia DFC* are in class order. **Bottom part:** confusion matrix presenting for each sample (columns) its class labels obtained by clustering (rows). Samples that are miss-classified appear as a vertical segment out of the main block diagonal.

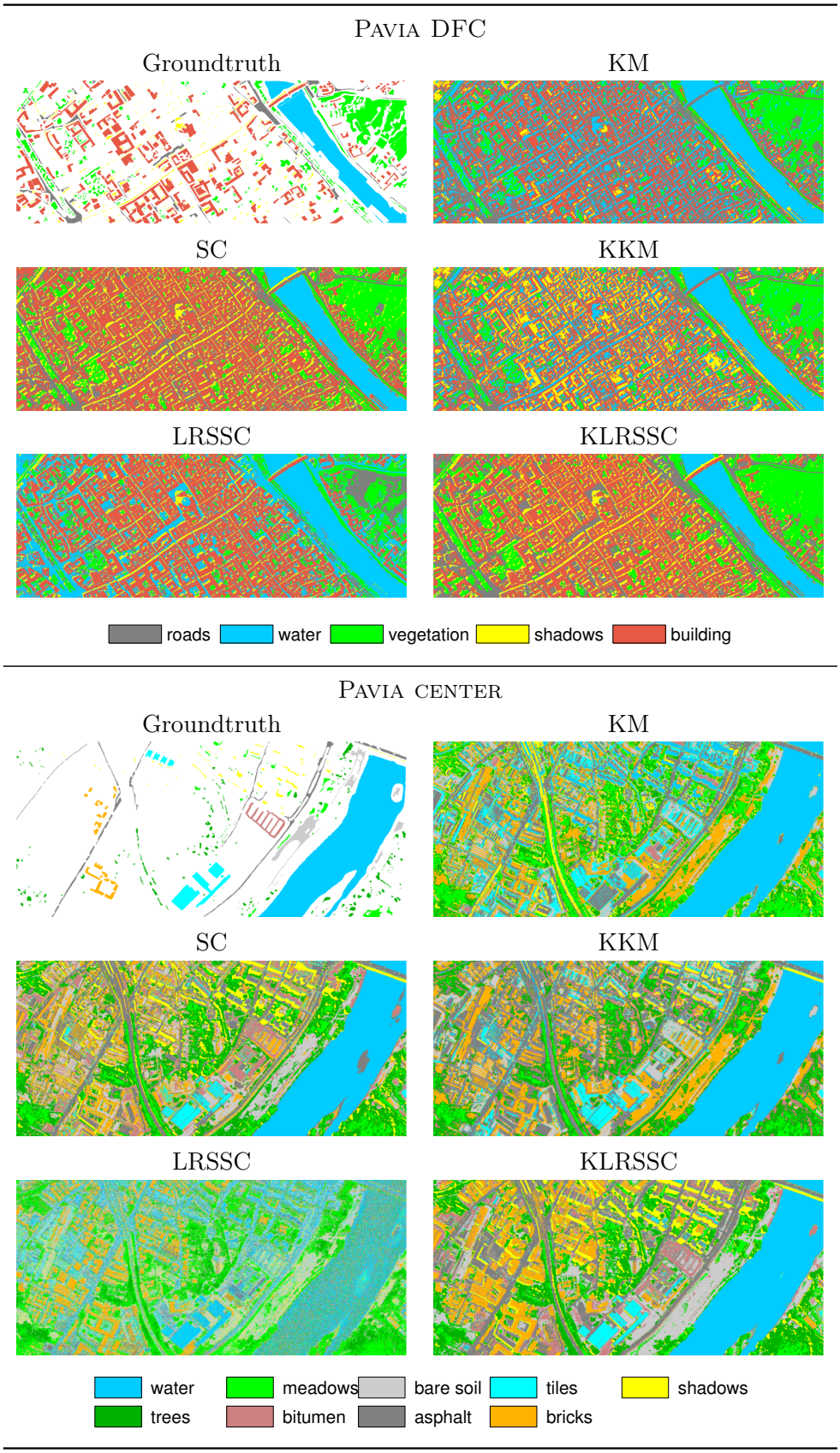



Figure 9.7: Classification maps for *Pavia DFC* and *Pavia center*.

## 9.6 Conclusions

This chapter introduced a non-linear low-rank and sparse representation method for clustering, the Kernel Low-rank Sparse Subspace Clustering (KLRSSC). KLRSSC assumes a union of multiple low-dimensional manifolds which are recovered as subspaces in Reproducing Kernel Hilbert Spaces. It benefits from the self-expressiveness property due to the sparsity constraint and from the improved intra-manifold connectivity due to the low-rank constraint. This allows to handle situation where manifolds are overlapping. We have also investigated the particular cases where only the sparse or low-rank constraint is active. The experimental behaviour of KLRSSC is evaluated over a set of databases of faces and objects and applied to the unsupervised classification of hyperspectral images.

This new representation of the data can be exploited in other settings such as active learning or semi-supervised learning. The latter is investigated in Chapter 10.

# 10 Semi-supervised one-class classification via graphs

In this Chapter, we introduce several graph-based methods for one-class classification and evaluate their respective performances with different types of graphs, among which the KLRSSC introduced in Chapter 9. We compare label propagation, constrained clustering and graph-regularized classification and observe the benefit of exploiting jointly the data, the graph and the labels to obtain the best detection performances in one-class classification problems.

## 10.1 Introduction

In Semi-Supervised Learning (SSL) situations, the data available is only partially labeled. The unlabeled samples carry information about the global geometry of the data and the labeled samples help recovering semantic classes meaningful for the user. A flexible approach for incorporating the data structure takes the form of a graph where the vertices represent the samples and the edges represent the similarities among them. As we have seen in the previous Chapter 9, many approaches exist to build a graph on the data, from the naive $k$-NN graph to our proposed KLRSSC graph representation, that tries to unveil the inherent data multi-manifold structure. In this Chapter, we investigate the SSL situation where additionally to the data a graph representing the data structure is available together with a few labels from one class, as illustrated in Figure 10.1.

The main families of methods compatible with this setting are the following:

- Label propagation: the labels are spread along the graph edges [Zhu et al., 2003, Zhou et al., 2004].
- Constrained spectral clustering: a spectral graph embedding is optimized to satisfy the constraints due to the labels [Wang et al., 2012a,b].
- Graph-based regularization: a classifier is adapted to the geometry of the data by

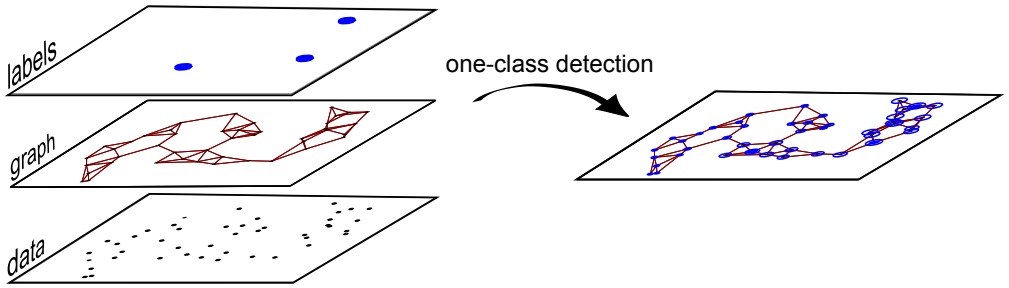



Figure 10.1: The different information to be exploited for the detection of a class of interest: the data features, a graph built on the data and some class labels.

distorting the training samples similarities accordingly to the graph structure [Belkin et al., 2006].

Before diving into the different approaches, let us recall some notations: the given graph $\mathcal{G} = (\nu, \epsilon, W)$ is composed of a set of vertices $\nu$, a set of edges $\epsilon$ and an edge weight matrix $W$ representing the weights between each pair of vertices. The normalized graph Laplacian is $L = \mathbf{I} - D^{-1/2}WD^{-1/2}$ and the transition probability matrix is $P = D^{-1}W = D^{-1/2}WD^{-1/2} = \mathbf{I} - L$, where $D_{j,j} = \sum_i W_{i,j}$ is the diagonal degree matrix and $\mathbf{I}$ the corresponding identity matrix.

## 10.2 One-class label propagation

The propagation of labeled information on the graph is an intuitive approach for combining the labels and the data structure. The labels are spread on the graph based on transition probabilities which are related to the graph Laplacian $L$ and the matrix of random walk transition probabilities $P$ [Zhu and Goldberg, 2009]. Zhu et al. [2003] introduced the Gaussian fields and harmonic functions (GFHF) which propagates the labels across the unlabeled samples, while the learning with local and global consistency (LLGC) proposed by Zhou et al. [2004] propagates the labels over the whole graph allowing a trade-off between the initial label and the propagated information. This trade-off allows the original labeled samples to be modified by the propagated information, which is interesting in case of erroneous labels.

The original formulations of GFHF and LLGC are established for two classes with the labels set to $\{-1, +1\}$. The labels are retrieved based on the positive or negative sign of the propagated information. The multi-way formulations, presented in Algorithm 4 and 5, spread the labeled information $\{+1\}$ separately for each class. The class likelihood after propagation can be directly used as a one-class classification score or to retrieve the labels by looking at the maximum class information in multi-class classification problem.

**Algorithm 4** Label propagation (GFHF) [Zhu et al., 2003]

**Input:** A graph $\mathcal{G}$, its weights $W$ and a label matrix $Y \in \mathbb{R}^{N \times k}$

1: Compute the transition probability matrix $P = D^{-1}W = D^{-1/2}WD^{-1/2} = \mathbf{I} - L$, where $D = \sum_i W_{i,j}$ is the degree matrix.

2: $\hat{Y}^{\infty}_{unlabeled} = (\mathbf{I} - P_{uu})^{-1}P_{ul}Y_l$

**Output:** Label $\{\mathbf{x}_i\}_{i=N_l+1}^{N}$ from $\arg\max \hat{Y}^{\infty}_{unlabeled}$

**Algorithm 5** Label spreading (LLGC) [Zhou et al., 2004]

**Input:** A graph $\mathcal{G}$, its weights $W$ and a label matrix $Y \in \mathbb{R}^{N \times k}$

1: Compute the normalized symmetric graph Laplacian $L = \mathbf{I} - D^{-1/2}WD^{-1/2}$, where $D = \sum_i W_{i,j}$ is the degree matrix and $\mathbf{I}$ the corresponding identity matrix.

2: Choose a parameter $\alpha \in [0, 1)$ controlling how much the labels can spread

3: $\hat{Y}^{\infty} = (1 - \alpha)(\mathbf{I} - \alpha L)^{-1}Y$

**Output:** Label $\mathbf{x}_i$ from $\arg\max \hat{Y}^{\infty}$

The one-class GFHF label propagation after convergence [Zhu et al., 2003] is obtained by

$$\hat{Y}^{\infty}_{unlabeled} = (\mathbf{I} - P_{uu})^{-1}P_{ul}Y_l \tag{10.1}$$

where $Y_l = \{y_i\}_{i=1}^{N_l}$, $\hat{Y}^{\infty}_{unlabeled}$ are the class probability obtained for each unlabeled sample $\{\mathbf{x}_i\}_{i=N_l+1}^{N}$ and $P = \begin{bmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{bmatrix}$. A sample issued from the class of interest will have $\hat{Y}^{\infty} \sim 1$.

The one-class label spreading (LLGC) after convergence [Zhou et al., 2004] is obtained by

$$\hat{Y}^{\infty} = (1 - \alpha)(\mathbf{I} - \alpha L)^{-1}Y \tag{10.2}$$

where $\alpha$ is the parameter controlling the trade-off between initial labelling and propagated information. It is interesting to notice that $(\mathbf{I} - \alpha L)^{-1}$ is a diffusion kernel [Schölkopf and Smola, 2002] and that the parameter $\alpha$ controls the diffusion. For $\alpha = 0$ , the initial labels are obtained without any diffusion. The values of $\hat{Y}^{\infty}$ can be interpreted as class probabilities after the diffusion process.

## 10.3 One-class constrained spectral clustering

The graph representing the data ideally shows disconnected components corresponding to the classes of interest, i.e. here the labeled class and the rest. The spectral embedding from the graph Laplacian has for principal dimensions the indicators of disconnected components in the graph. The spectral clustering relies on this graph property to infer class labels, see Algorithm 1 in Section 3.3.2. In the spectral embedding of Ng et al. [2002], the $k$ first eigenvectors $V \in \mathbb{R}^{N \times k}$ of the Laplacian $L \approx V\Sigma V^T$ are normalized along

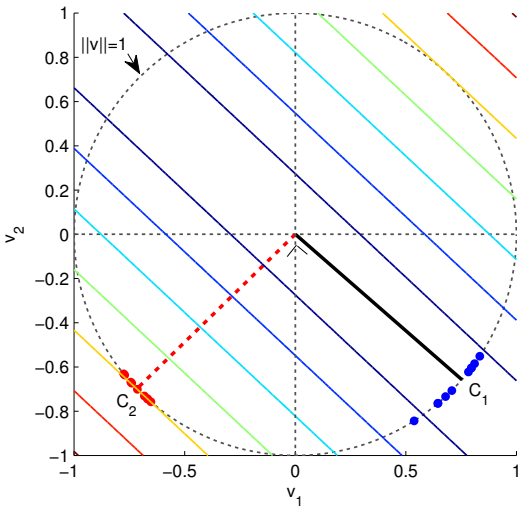



Figure 10.2: Schematic of the two dimensional spectral embedding based on the normalized Laplacian rescaled to the unit circle. $C_1$ represents the class of interest, where labels are available, and $C_2$ the novelties, composed of other classes. The spectral orthogonal subspace projection (SOSP) distance is represented by the colored isolines.

the rows to unit norm: $\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}$. Thus, the embedded samples $V^T$ are concentrated around points on the unit-hypersphere representing their cluster. Figure 10.2 illustrates this spectral embedding for $k = 2$ dimensions. The inner product between pair of samples is equivalent to the cosine of their angle $\theta$, which represent their correlation. Ideally, for a graph having $k$ disconnected components, the embedded samples are exactly lying on $k$ distinct points in $\mathbb{R}^k$. These $k$ vector points representing the different clusters are orthogonal between each other. Therefore, two samples from different clusters are orthogonal ($\theta = 90°$) and have zero correlation since their cosine is null. The $k$ cluster points are not aligned with the basis formed by the eigenvectors, but correspond to a rotation of them. This rotation ambiguity justifies the use of $k$-means [Ng et al., 2002] or the need to find the appropriate rotation matrix [Yu and Shi, 2003] to assign the cluster labels.

### 10.3.1 Spectral orthogonal subspace projection

In the spectral embedding represented in Figure 10.2, the orthogonal distance to the subspace representing the class $C_1$ provides a measure on how dissimilar a sample is from $C_1$. With this distance measure, the most dissimilar point on the hypersphere from $C_1$ is the one in the direction representing $C_2$ orthogonal to the direction of $C_1$. The point on the unit hypersphere representing the subspace of class $C_1$ can be identified based on the labeled samples of the class of interest, denoted $\mathbf{v}_l$. Let us define $\mathbf{p}$ the principal eigenvector of $\mathbf{v}_l$ representing the subspace of $C_1$. The *span* of the subspace $C_1$ is $\mathbf{p}\mathbf{p}^\dagger$,

therefore the orthogonal projection operator is $\mathbf{p}^{\perp} = \mathbf{I} - \mathbf{p}\mathbf{p}^{\dagger}$, with $\mathbf{I}$ the identity matrix. The spectral orthogonal subspace projection distance is

$$d_{sosp}(i) = \|\mathbf{v}_i \mathbf{p}^{\perp}\|_2 \tag{10.3}$$

where samples belonging to the class of interest will have a lower distance than the others.

### 10.3.2 Spectral learning

The spectral embedding may not be optimal for the class of interest, since not exploiting the labeled information. A straightforward approach enforcing class constraints is to modify the affinity matrix $W$ [Kamvar et al., 2003]. The affinities between labeled samples of the same class are set to 1, while affinities of labeled samples from different classes are set to 0. In the context of one-class classification only affinities between the same class can be modified in the absence of negative class information. Therefore, we modify the affinity matrix as follows

$$W_{i,j} = \begin{cases} +1 & \text{if i,j are both labeled} \\ W_{i,j} & \text{otherwise left unchanged} \end{cases} \tag{10.4}$$

Notice that the weights matrix $W$ is normalized in order to have $W_{i,j} \in [0, 1]$ prior to modifying the weights. The spectral embedding is obtained from the eigendecomposition of the normalized graph Laplacian of the modified weights matrix $W$. Finally, the spectral orthogonal subspace projection distance can be used as described in Section 10.3.1.

### 10.3.3 Constrained spectral clustering

The class labels available can be used to establish contraints and allow to find an embedding dedicated to the class of interest which respects the graph structure. In the constrained clustering with multiple classes, the two main ingredients are the Laplacian matrix $L$ reflecting the graph structure and the constraint matrix $Q$ [Wang et al., 2012a]. The $N \times N$ matrix $Q$ contains the class constraints, i.e. a positive value for samples $i, j$ from the same class and a negative value for samples $i, j$ from different classes. The values $Q_{i,j}$ represent the strength of the link belief. A large absolute value will be a strong constraint, while a value close to zero is a weak constraint. The unconstrained spectral clustering is recovered for $Q = \mathbf{I}$.

**Must-link constraints** In our one-class classification situation, the assignment of the matrix $Q$ is ill-posed. As for spectral learning in Section 10.3.2, only *must-link* constraints can be set based on the class of interest. The *cannot-link* constraints cannot be set in the

absence of a negative class. In this situation, the matrix $Q$ is defined as

$$Q_{ij} = \begin{cases} +1 & \text{if i,j are both labeled} \\ 0 & \text{otherwise} \end{cases} \tag{10.5}$$

The matrix $Q$ can be normalized as: $\bar{Q} = D^{-1/2}QD^{-1/2}$. Then the constrained spectral clustering [Wang et al., 2012a] solves the following problem

$$\begin{aligned} \min_{\mathbf{v}\in\mathbb{R}^N} \quad & \mathbf{v}^T L\mathbf{v}, \\ \text{s.t.} \quad & \mathbf{v}^T\bar{Q}\mathbf{v} \geq \beta, \quad \mathbf{v}^T\mathbf{v} = 1, \quad \mathbf{v} \neq D^{1/2}\mathbf{1} \end{aligned} \tag{10.6}$$

where $\mathbf{v}^T L\mathbf{v}$ is the graph cut to be minimized and $\mathbf{v}^T\bar{Q}\mathbf{v} \geq \beta$ is the lower bound on the constraints satisfied. The two other constraints are here to normalize the embedding and avoid trivial solutions.

This problem can be solved by the following generalized eigenvalue problem

$$L\mathbf{v} = \lambda(\bar{Q} - \beta\mathbf{I})\mathbf{v} \tag{10.7}$$

The solutions of (10.7) corresponding to positive eigenvalues $\lambda_i$ are kept and normalized such that $\mathbf{v}^T\mathbf{v} = 1$. These are the ones satisfying the constraints $\mathbf{v}^T\bar{Q}\mathbf{v} \geq \beta$. Hence, among the feasible set of solutions the one minimizing the cost of the cut $\mathbf{v}^T L\mathbf{v}$ is retained. It represents the solution satisfying sufficiently the constraints according to $\beta$ and respecting at best the graph structure.

**Must-link constraints from label propagation** The matrix $Q$ solely based on the few class labels is sparse (most entries are zeros) leading to instabilities [Basu et al., 2008]. Wang et al. [2012b] proposed an alternative to this issue. First the labels are propagated on the graph and then constraints are defined based on the propagated information. The matrix of constraints $Q$ is defined as

$$Q_{ij} = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma^2}\right) \tag{10.8}$$

where $\sigma$ is the bandwidth defining the similarity among the propagated labels $y_i$. The labels can be propagated using either GFHF or LLGC. Notice that matrix $Q$ is a kernel matrix between the vertices in the semi-supervised embedding induced by the propagated labels. In this case, the normalization of $Q$ is $\bar{Q} = D_Q^{-1/2}QD_Q^{-1/2}$, where $D_Q = \text{diag}\left(\sum_i Q_{i,1}, .., \sum_i Q_{i,N}\right)$. The embedding is obtained by solving the generalized eigenvalue problem (10.7) for $\beta = 0$. After removing the eigenvectors corresponding to

null eigenvalues, the embedding that best satisfies the constraints is

$$\mathbf{v}^* = \max_{\mathbf{v}} \mathbf{v}^T \bar{Q} \mathbf{v} \tag{10.9}$$

## 10.4 One-class classifier with graph-based regularization

Within the SSL context, standard classifiers can be regularized by the graph Laplacian [Belkin et al., 2006] thus enforcing to maintain the structure of the data manifold. For one-class classification, the semi-supervised SVDD ($S^3$VDD) introduced in Section 4.4.3 allows to provide a better class description using the data manifold geometry embedded in the graph Laplacian. The SVDD is trained with a kernel between labeled samples distorted by the graph Laplacian which contains information on both labeled and unlabeled samples. The details on the kernel distortion are presented in (4.15) of Section 4.4.3.

## 10.5 Experiments

The experiments aim at studying the detection performance of a class of interest based on different graphs built on the data and few labeled samples.

### 10.5.1 Experimental setup

In the experiments we compare four different types of graphs build on the data: the standard $k-NN$ graph with $k = 7$, the KLRSC, the KLRSSC with $\lambda = 10^{-6}$ and the KSSC, introduced in Chapter 9. As a reminder, $\lambda$ is balancing the trade-off between the low-rank and the sparse representation, meaning a weight matrix $W$ that is between being block-diagonal and being sparse with only few non-zero weights present in each column.

We compare the following different approaches:

- GFHF and LLGC: label propagation where the diffusion parameter for LLGC is $\alpha = 0.5$ [Wang et al., 2012b].
- SOSP: the SOSP distance in the standard spectral embedding [Ng et al., 2002].
- SL: the SOSP distance in the embedding of spectral learning [Kamvar et al., 2003].
- CSC: constrained spectral clustering with *must-link* constraints from labeled samples.
- CSC+GFHF/LLGC: constrained spectral clustering with *must-link* constraints from label propagation and the bandwidth $\sigma$ set to the mean value in the matrix $Q$.

| | GFHF | LLGC | SOSP | CSC | CSC+GFHF | CSC+LLGC | S$^3$VDD |
|---|---|---|---|---|---|---|---|
| $k$-NN | $.95 \pm .03$ | $.81 \pm .07$ | $.94 \pm .02$ | $.80 \pm .00$ | $.86 \pm .07$ | $.97 \pm .01$ | $.97 \pm .01$ |
| KLRSC | $.55 \pm .01$ | $.97 \pm .00$ | $.93 \pm .02$ | $.84 \pm .02$ | $.80 \pm .00$ | $.97 \pm .01$ | $.97 \pm .01$ |
| KLRSSC | $.68 \pm .01$ | $.98 \pm .00$ | $.95 \pm .01$ | $.73 \pm .01$ | $.80 \pm .00$ | $.98 \pm .01$ | $.97 \pm .01$ |
| KSSC | $.74 \pm .02$ | $.97 \pm .01$ | $.86 \pm .01$ | $.69 \pm .06$ | $.77 \pm .00$ | $.96 \pm .00$ | $.97 \pm .01$ |

Table 10.1: The AUC results for *Pavia DFC* averaged over all classes for 10 labeled samples. Each row is a different graph: $k$-NN, KLRSSC and KSSC

- S$^3$VDD: the SVDD distorted by the graph Laplacian with the distortion $\varepsilon$ parameter searched in the range $\varepsilon = [0, ..., 1000]$. The regularization $\lambda$ is set to 0, for fair comparison and avoiding the extensive use of cross-validation.

The hyperspectral datasets *Pavia DFC* and *Pavia center* are used with each of the thematic class from their available groundtruth alternatively considered as the class of interest. The datasets are presented in Appendix B.2. The different graphs are built based on a subset of samples from the images, of size 1000 with 200 samples from the class of interest in *Pavia DFC* and of size 990 with 110 samples from the class of interest in *Pavia center*, in the same way than in Chapter 9. Five random runs, where labels of the class of interest are randomly selected, are performed for each fixed number of labels and for each specific class.

The detection performances are evaluated using the ROC curve and the area under the ROC (AUC), as well as the Cohen's $\kappa$ statistic. The higher the AUC, the better the class detection.

### 10.5.2 Results

In Tables 10.1 and 10.2 are reported the average AUC over all the classes and fives random runs for ten labeled samples and different types of graphs in *Pavia DFC* and *Pavia center*, respectively.

First of all, we observed that the SOSP distance with the standard spectral embedding always outperformed the SOSP distance in the embedding of spectral learning with *must-link* constraints only. The spectral learning with must-link constraints only and in small number cannot provide a very different embedding. Therefore, we only report the

| | GFHF | LLGC | SOSP | CSC | CSC+GFHF | CSC+LLGC | S$^3$VDD |
|---|---|---|---|---|---|---|---|
| $k$-NN | .86 ± .07 | .85 ± .01 | .88 ± .01 | .79 ± .05 | .78 ± .06 | .91 ± .08 | .98 ± .01 |
| KLRSC | .57 ± .01 | .96 ± .01 | .88 ± .02 | .95 ± .01 | .73 ± .00 | .95 ± .03 | .98 ± .01 |
| KLRSSC | .69 ± .02 | .99 ± .01 | .91 ± .02 | .70 ± .13 | .77 ± .02 | .96 ± .03 | .97 ± .01 |
| KSSC | .73 ± .06 | .99 ± .00 | .95 ± .00 | .70 ± .05 | .77 ± .00 | .97 ± .03 | .98 ± .01 |

Table 10.2: The AUC results for *Pavia center* averaged over all classes for 10 labeled samples. Each row is a different graph: $k$-NN, KLRSC, KLRSSC and KSSC

results for the SOSP with the standard spectral embedding in the two Tables.

In both datasets, we observe that the GFHF works only with the $k$-NN graph. It requires a graph with few connections that are mostly reliable, which the $k$-NN provides. In opposition, the KLRSC graph has many non-zero edge weights among which may be erroneous but weak edges. The LLGC provides good performances for our three graphs, that outperform the standard $k$-NN graph.

The SOSP distance is performing well across the different graphs with its best performances using the KLRSSC in *Pavia DFC* and the KSSC in *Pavia center*. The latter having more classes potentially lying closer to each other, the fewer connections present in the KSSC graph could help in discriminating certain classes.

The CSC produces lower performances in general than the other methods. The use of only few labels from the class of interest produces a very sparse matrix of constraints with only *must-link* constraints. This seems insufficient to detect accurately a class of interest. Moreover, we observe that CSC requires densely connected graphs like the KLRSC, or at least not as sparse as the KSSC. The additional graph connections allows to get a certain flexibility in the embedding to more easily satisfy the constraints. A graph like $k$-NN or KSSC with very few connected edges cannot provide a rich enough spectral embedding that would satisfy the constraints. The constraints derived from GFHF improve slightly the performances except with KLRSC where it decreases. This is due to the propagated labels on wrongly connected edges that impose erroneous constraints. On the contrary, CSC with the constraints from LLGC propagation provides among the best performances, more specifically with our proposed graphs. In this case, the diffusion parameter allows to control the amount of propagation and thus the shape of the constraints matrix.

The S$^3$VDD provides performances among the best which are stable across the different

graphs. In comparison to the other methods, it exploits the original data through the kernel matrix, while the CSC and label propagation approaches only rely on the graph and the labels.

In Figure 10.3, the $\kappa$ statistic for the S$^3$VDD for each class to be detected, various number of labeled samples and different graphs is presented. We notice that the accuracy increases faster with the increasing number of labels for KLRSC and KSSC than the standard $k$-NN. As an example, the vegetation class (green line) is only around 0.5 $\kappa$ for ten labels using the $k$-NN graph, while already at 0.7 for the KSSC and 0.8 for the KLRSC.

In Figure 10.4, we can better see the impacts of propagating the labels on the graph and using them as constraints. For LLGC alone, the AUC results are stabilized from 10 labels and above, while CSC needs 20 labeled samples (= 10% of the class samples in the graph). When both are combined in the form of constraints from propagated labels, the performances are already stabilized from 3 labeled samples.

In Figure 10.5, the ROC curves using the KLRSSC graph and some of the different methods are presented. In general we observe some common difficulties in detecting certain classes across all methods. In *Pavia DFC*, the classes *roads* and *building* have the lower AUC since sharing some of their spectral characteristics. The CSC and CSC+GFHF are only performing well for few classes: the vegetation class in *Pavia DFC* and the shadows in *Pavia center*. The LLGC propagation provides very sharp transition, with the best AUC for *Pavia center*. Its usage as constraints together with the CSC are providing among the best performances along with the S$^3$VDD. We can also observe the limitations of the SOSP, which performances are decreased in *Pavia center* due to the larger number of classes in the image.

Finally, detection maps with the S$^3$VDD and the KLRSSC for different classes of interest in the two datasets are presented in Figure 10.6. The maps present the decision function of the S$^3$VDD at every pixel where the pixels belonging to the class have positive values and the pixels detected as novelties, i.e. from different classes, have negative values. We can clearly observe the relationships between the classes: the class *roads* in *Pavia DFC* has values that are close to 0 for certain building regions and the class *building* has also positive values for regions over roads.

We see that this problem of semi-supervised one-class classification or novelty detection can be tackled from different perspectives having some advantages and disadvantages. Among the compared methods, we retain the constrained spectral clustering using propagated labels as constraints (CSC+LLGC) which provided very good performances and relies only on solving general eigenvalue problems based on the data graph and labels. Also the S$^3$VDD provided among the best performances. In terms of computations, distorting the kernel of the S$^3$VDD is approximately similar to solving the general eigenvalue problem if implemented using the efficient Nyström method. Finally, the training of the S$^3$VDD

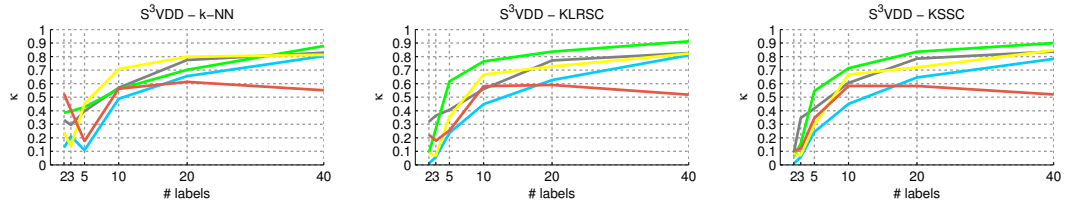



Figure 10.3: $\kappa$ statistic for the different graphs and number of labeled samples in *Pavia DFC*.

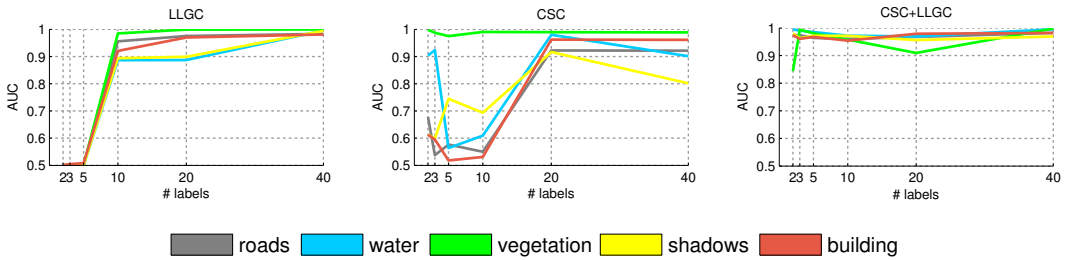



Figure 10.4: AUC for different methods with the graph KLRSSC averaged over 5 runs and for different number of labeled samples in *Pavia DFC*.

would need to compare with finding the embedding maximizing the constraints, which is typically done in a much faster way.

## 10.6 Conclusions

In this Chapter, we have investigated the problem of semi-supervised one-class classification using a graph representing the data structure and few class labels. The comparisons between label propagation, unconstrained and constrained spectral clustering satisfying both the graph structure and the labeled constraints, as well as a one-class classifier regularized by the graph provided insights on the different approaches and the benefit of different graph structures. The constrained spectral clustering using propagated labels from LLGC as constraints, the semi-supervised support vector data description and the simple label propagation provided the best performances and using the different graphs proposed in Chapter 9, which better retrieve the data structure. The use of these particular graph in semi-supervised learning could be probably even enhanced by considering certain transformations of the graph weights as suggested in Zhang et al. [2013b].

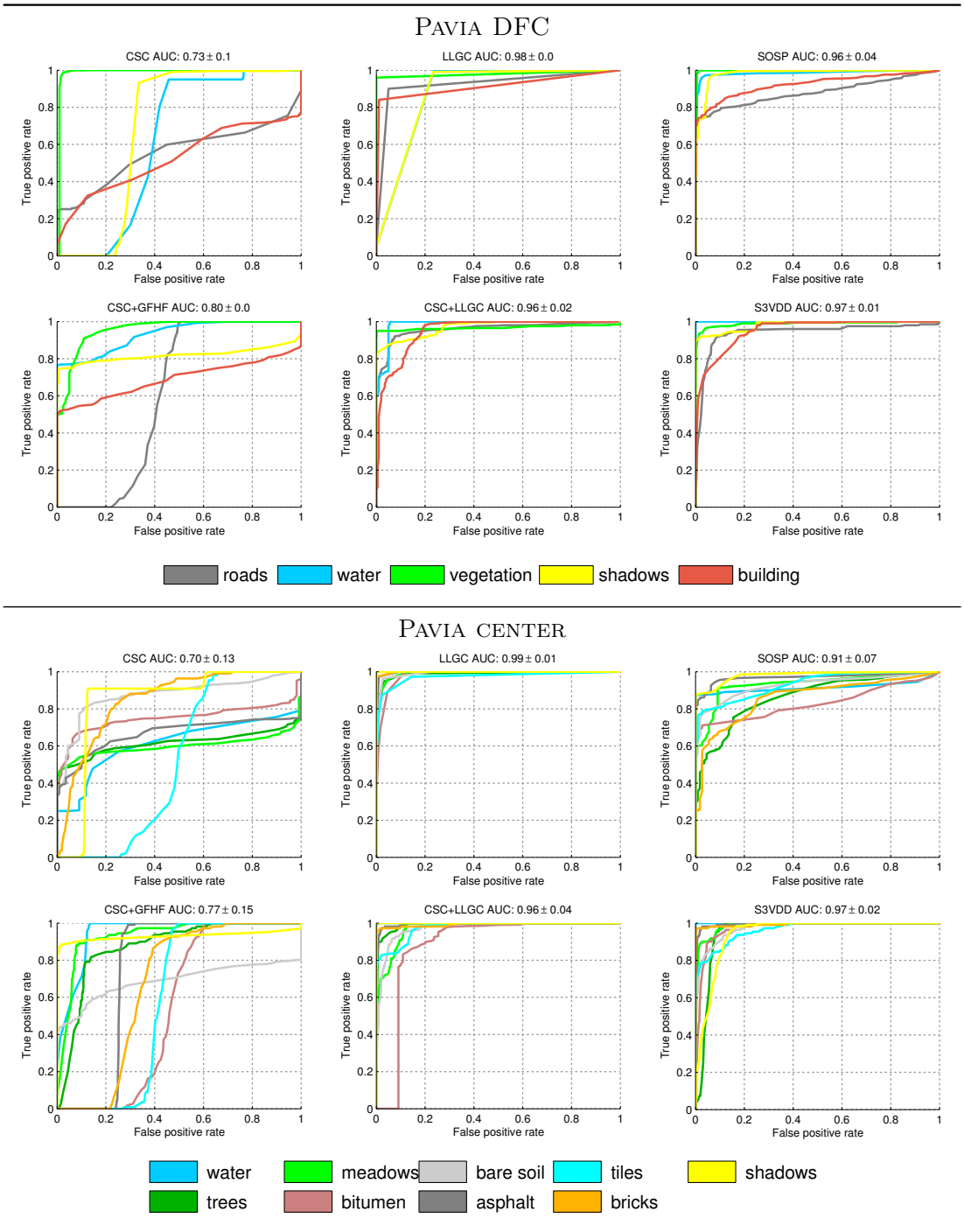



Figure 10.5: ROC curves for *Pavia DFC* and *Pavia center* averaged over 5 different sets for 10 labeled samples and using KLRSSC graph.

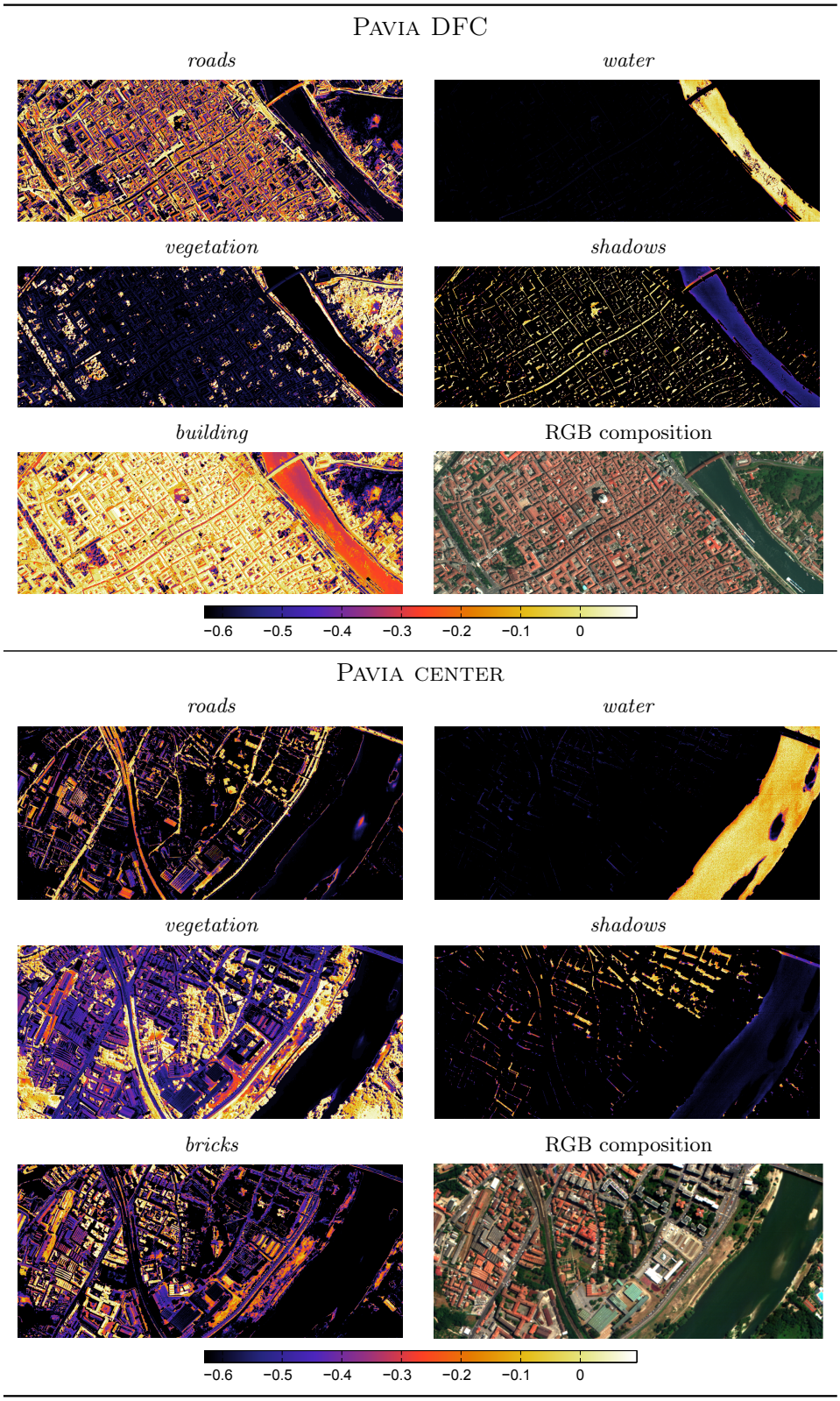



Figure 10.6: Detection maps for classes in *Pavia DFC* and *Pavia center* using the S$^3$VDD with the KLRSSC graph and 10 labeled samples. The color legend corresponds to the decision function which is positive for the class of interest and negative for novelties.

# Conclusions Part IV

# Conclusions

## Towards more realistic detection approaches

The problems of detection and classification with a very low level of supervision is at the forefront of research. It is of prime importance in order to meet the requirements of a broad range of real applications based on the automatic processing of massive data. In these situations, better assumptions on the data are required in order to alleviate the lack of supervision. The success of kernel methods for pattern recognition problems and classification, particularly in remote sensing is partially due to their adaptation to the data structure and their inherent robustness to noise and complex high-dimensional data.

In this Thesis, we have investigated the ability of kernel methods to handle the detection of unknown classes with partial or none supervision, under two different data structure assumptions. We demonstrated the appropriateness of the approaches for novelty detection in remote sensing, taking the form of novelty detection, unsupervised classification or one-class classification. In these challenging settings, the assumptions on the data structure are primordial. We have investigated two particular assumptions on the data structure, motivated by remote sensing applications.

The first assumption, investigated in Part II of the Thesis, is that the data is made of overlapping clusters. These clusters are made of core samples separated by regions of low density populated by outlier samples representing mixed or outlier samples. In remote sensing, this assumption is verified with medium resolution multispectral images where mixed pixels are present. Based on this assumption, we developed a clustering system aiming at retrieving changed and unchanged regions in multi-temporal datasets. We also developed a semi-supervised approach detecting changes based solely on the annotation of some unchanged regions. This is more realistic for post-catastrophe situations where the access to the affected region is often impossible, as for example after flooding or fire events.

The second assumption, investigated in Part III of the Thesis, is that the data is made of multiple low dimensional manifolds. These manifolds represent different thematic classes and can be identified by building a graph of the data structure and retrieving

the disconnected components. In high-dimensional data, such as hyperspectral images, the retrieval of these manifolds is challenging. We developed a new graph representation exploiting this assumption based on recent convex optimization tools, allowing to retrieve low-rank and sparse properties in the representation of the data. Finally, we have studied the semi-supervised situation with different approaches for combining a graph representation of the data with few labels from one class.

The computational considerations are important when dealing with automatic methods that should provide early detection warnings. In Part II of the Thesis, we have provided approaches that accommodate with the data and are faster than the state-of-the-art. The entire solution path algorithms allows to derive solutions much faster than standard optimization procedure. In the semi-supervised setting, the selection of parameters is not more costly when done automatically compared to using cross-validation. However, in the completely unsupervised case the support vector clustering algorithm remains extremely costly to compute and this bottleneck should be further addressed. In Part III of the Thesis, the construction of the graph representing the data structure is computationally intensive and directly relates to the subset of the data used and the number of iterations performed. We proposed an out-of-sample extension, which allows to classify the samples in an online fashion and avoid manipulating a large set of data at once.

In all the experiments, the developed approaches provided more accurate results while respecting the partial or complete lack of supervision. This encourages to continue the investigation in these directions towards the real application of automatic image processing for change detection and classification problems. This methods could provide early detection to the user, which then could refine the result depending on its needs. It is important to keep in mind that user is in the processing loop sooner or later. The proposed approaches are extremely useful as first approximations of changes, and detect changes that are highly trusted, which provide very few false detection.

## Contributions

The principal contributions of the Thesis, more or less independently of their context, can be summarized as follows:

- The introduction of truly hierarchical kernel-based clustering methods dealing with overlapping clusters of arbitrary shapes, namely the Nested Support Vector Clustering (NSVC).
- The development of a cluster validity measure handling hierarchical clusters of arbitrary shape for the unsupervised selection of hyperparameters.
- The development of a cluster merging system which groups clusters that overlaps in order to obtain fewer groups closer to the user needs.

- The introduction of hierarchical entire solution path algorithms for lower computational burden in both clustering and semi-supervised novelty detection.
- The development of a low density criterion in unbalanced classification for the unsupervised selection of hyperparameters.
- The development of a kernel-based subspace clustering method that exploits sparsity and low-rank properties to identify multiple low-dimensional manifolds in the data. It allowed the clustering of objects or faces of different subjects and the unsupervised classification of hyperspectral images.
- The introduction of appropriate measures for the unsupervised selection of hyperparameters in non-linear subspace clustering.
- The study of semi-supervised classification, constrained spectral clustering and label propagation in the context of one-class classification exploiting a graph of the data structure.

## Future perspectives

The further perspectives of this research are mostly on the interaction with the user, the incorporation of more information from the data and the use of more complex sources of data. Here are briefly presented the main points that would need to be further investigated in order to overcome the present issues for novelty detection problems dealing with images and the more specific challenges regarding remote sensing applications.

- **Active learning strategies** The interaction with the user is extremely important in the applications. We have presented automatic and semi-automatic methods, which provide an early detection results to a user. At that point the user might be interested to interact with the machine to provide few but relevant information. The machine selects ambiguous data where the information from the user would impact the most the result. Among our proposed methods for parameter selection, most of them can be exploited in an active learning fashion. In Chapter 6, the samples identified as outliers between clusters and used for merging could be given to the user in order to either confirm the merging or avoid the merging of the clusters and refine their boundaries. In Chapter 7, our low density criterion selects an initial solution, which can be refined by giving the user the samples used in the criterion (samples along the boundary) and asking if they are from the same class or not. This could allow to refine the selection of the appropriate parameter defining the boundary. In Chapter 9, the user could interact during the iterations of the optimization to confirm certain coefficients retrieved. If the coefficients are not enough sparse, the sample is potentially an outlier or at the intersection between classes. The user could help by forcing certain connections between samples to

be removed. The out-of-sample strategy based on the residual of a collaborative representation is a direct indicator of the ambiguity and novelty of the test sample, which could be unveiled by the user.

- **Image data considered as tensor** In high dimensional images the spatial information should be incorporated in the clustering/classification process. The spatial regularization can be either imposed by combining spatial and spectral kernels [Camps-Valls et al., 2008] or by a spatial graph Laplacian regularization [Camps-Valls et al., 2007]. The latter would allow more particular forms of regularization such as non-local similarities [Buades et al., 2010]. More precisely, in the collaborative representation used to extend our graph to new test samples, the coefficients could be spatially smoothed based on the following

$$min_{\mathbf{C}_t} \|\mathbf{C}_t\|_2^2 + \|\Phi(\mathbf{X}_t) - \Phi(\mathbf{X})\mathbf{C}_t\|_2^2 + \mathbf{C}_t^T L \mathbf{C}_t$$

  where $\mathbf{C}_t$ are the retrieved coefficients and $L$ is the spatial laplacian, which enforces a smooth representation.

- **Multi-modal sources** The available sources of data for a given problem are very often multiples. In remote sensing, a variety of sensors is covering the Earth and could provide a valuable source of information if combined. This brings the challenge of fusing sources with potentially different spatial and spectral resolutions, or even different modalities (radar and optical). A multi-modal system would overcome the limitations of each single modality, e.g. the cloud coverage in optical sensors and the speckle noise in radar, allowing a more frequent coverage. The kernel methods can help handling different sources via the kernels addition or product (concatenation in feature space) [Camps-Valls et al., 2008], the projection in a common subspace via kernel CCA [Chaudhuri et al., 2009, Volpi et al., 2013], a consensus among the views [Bickel and Scheffer, 2004, Kumar et al., 2011], the multi-modal joint sparse representation [Nguyen et al., 2011] or tensor analysis [Kolda and Bader, 2009] which finds a joint optimal subspace.

- **Alleviating co-registration** One of the major limit about change detection is the need for very accurate co-registration of the multi-temporal images. This provides a pixel correspondence in time on which relies most of the algorithms. In the case of non-registered images, the match between the different samples has to be performed based on other assumptions. Some recent attempts have considered this issue in remote sensing [Theiler and Wohlberg, 2012] which could benefit from graph matching and domain adaptation methodologies [Tuia et al., 2013] to overcome such issue.

# A Appendix A

## A.1 Accuracy evaluation metrics

In order to evaluate the quality of a classification or detection result and to compare the performances of different approaches, we can use different metrics providing a complementary view. They all rely on measuring the agreement between the predicted and the true labels, but by different means.

**Confusion matrix** The Confusion matrix, also called Contingency table, is a matrix where the correspondences between the predicted classes and the true labels (the ground truth) are reported. It is of size $n_{class} \times n_{class}$ where the ground truth labels are in columns and the predictions in rows. In the case of perfect predictions, the matrix is diagonal. In this Thesis and in general for detection problems, the Confusion matrix is a $2 \times 2$ matrix with a positive and a negative class defined as in Table A.1.

| | | Truth | | |
|---|---|---|---|---|
| | Class | 1 | 2 | User acc. |
| Prediction | 1 | $n_{11}$ | $n_{12}$ | $n_{11}/\sum_j n_{1j}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{22}/\sum_j n_{2j}$ |
| | Producer acc. | $n_{11}/\sum_j n_{j1}$ | $n_{22}/\sum_j n_{j1}$ | |

Table A.1: Confusion table: correspondences between predicted and true labels in the case of two classes.

**False Alarm Rate (FAR)** corresponds to the percentage of samples from class 1 detected as class 2, called error of "'Type I".

**Missed Alarm Rate (MAR)** corresponds to the percentage of samples from class 2 detected as class 1, called error of "Type II". In change detection, the FAR relates to

wrongly detected changes, while MAR relates to the missed changes.

**Overall accuracy (OA)** The Overall Accuracy (OA) is defined as the number of correct predictions divided by the total number of samples:

$$OA = \frac{\sum_j n_{jj}}{n} \tag{A.1}$$

where $n$ is the total number of predicted samples. This is the sum of the matrix diagonal over the total number of samples. The OA is biased towards more frequent classes, which may be problematic when assessing the performances in unbalanced class situations.

**Missrate (MR)** The missrate (MR) is the number of incorrectly classified samples divided by the total number of samples. It is $1 - OA$, the opposite of OA.

**Kappa statistic ($\kappa$)** The Cohen's Kappa statistic [Foody and Mathur, 2004] is used to evaluate the number of samples correctly classified with respect to the number of samples potentially classified correctly by chance.

$$\kappa = \frac{p(k) - p(r)}{1 - p(r)} \tag{A.2}$$

where $p(k)$ is the agreement rate, corresponding to OA, and $p(r)$ is the agreement due to chance, corresponding to the sum of the product between the fractions of correctly detected classes and the product between the fractions of correctly predicted classes. Compared to OA, the $\kappa$ measure allows to better cope with unbalanced classes.

**F-score** The F-score measure considers both the precision $p = \frac{n_{22}}{(n_{22}+n_{21})}$ and the recall $r = \frac{n_{22}}{(n_{22}+n_{12})}$ of the predictions to compute the score. The F-score or balanced F-score (F1 score) is the harmonic mean of precision and recall which can be interpreted as a weighted average.

$$F - score = 2 \cdot \frac{p \cdot r}{p + r} \tag{A.3}$$

F-score ranges between 0 and 1, with its best score at 1.

**Receiver Operating Curve (ROC)** The ROC curve is useful to evaluate a model in detection problems. The curve is formed by the True Positive Rate (TPR) and the False Positive Rate (FPR). The curve is connected at (TPR,FPR)=0 and (TPR,FPR)=1. The perfect curve would be a TPR=1 for any FPR> 0. A curve where TPR linearly grows

with FPR is associated to a random prediction. A concave curve would mean a model which performs worse than random (the prediction opposite to the model would be more relevant).

**Identifying clusters** In clustering problems the results are the association of a cluster identity $\Omega = \{\omega_1, .., \omega_N\}$ to each data sample. However, for better interpretation of the result we often want to match these cluster indices to class labels $Y = \{y_1, .., y_N\}$ from a predefined groundtruth map. The *Hungarian method* [Kuhn, 1955] aims at converting the cluster indices into the class labels by finding the best match possible between the two. This assignment method tries to find the best match between nodes of a bipartite graph. For the problem of cluster matching, the cost of the graph edges are inversely related to the entries of a contingency table made between the two assignments.

Once the identification performed, the clustering *purity* is obtained as for the OA.

**Normalized mutual information (NMI)** The NMI [Yao, 2003] is used to evaluate a clustering quality. NMI is defined as the mutual information between the cluster assignments and the labeling of the dataset normalized by the arithmetic mean of the maximum possible entropies of the empirical marginals.

$$NMI = \frac{MI(\Omega, Y)}{(H(\Omega) + H(Y))/2} \tag{A.4}$$

On the contrary to cluster purity, the normalization of the Mutual Information avoids to increase when the number of clusters increases. NMI lies in the range $[0, 1]$ such that the higher the value, the better the clustering quality.

**Adjusted rand index (aRI)** The aRI [Hubert and Arabie, 1985] is a clustering quality measure which is adjusted for the chance of grouping elements. It ranges between $[-1, 1]$ and is maximum for two exactly similar clustering solutions.

$$aRI = \frac{\sum_{i,j} - \left(\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right) / \binom{N}{2}}{\frac{1}{2}\left(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right) - \left(\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right) / \binom{N}{2}} \tag{A.5}$$

where $a_i = \sum_i n_{ij}$ and $b_j = \sum_j n_{ij}$.

# B Appendix B

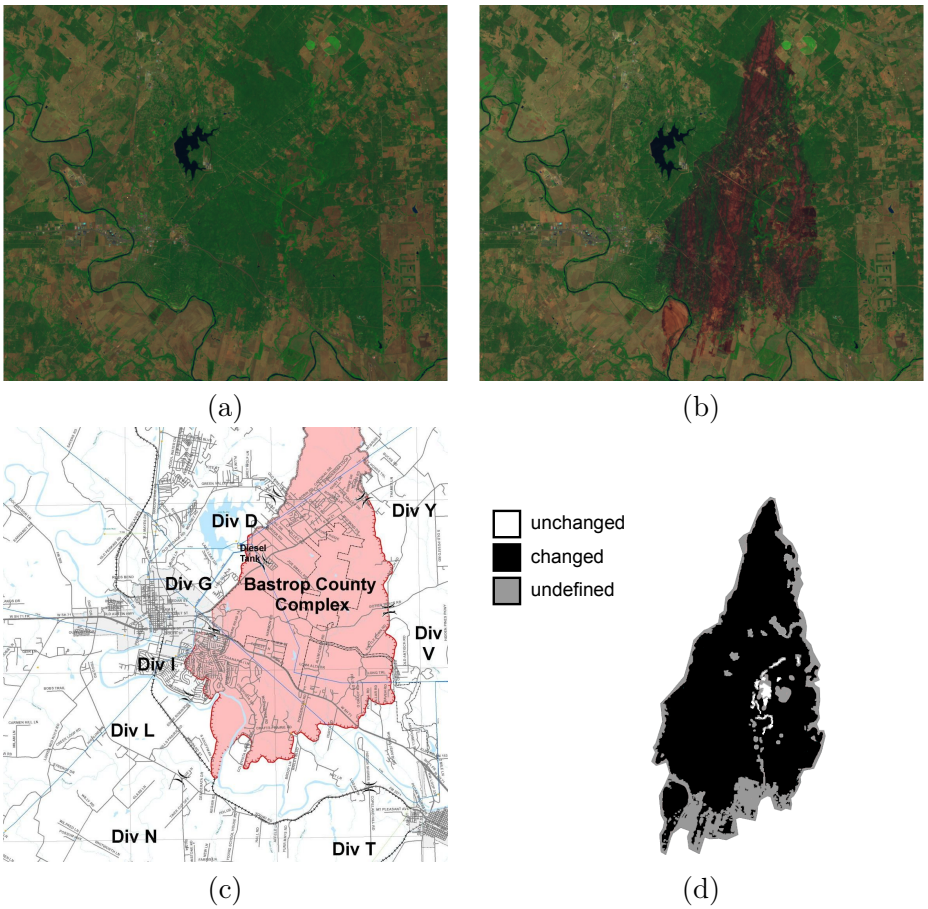
## B.1 Landsat

### B.1.1 Bastrop fires

In september 2011, the most important wildfires of Texas started near Bastrop, Texas, USA. Strong winds and dry conditions caused the most destructive wildfire in Texas history. The fire killed two people, destroyed over 1600 homes, and caused over 325 million in property damage.

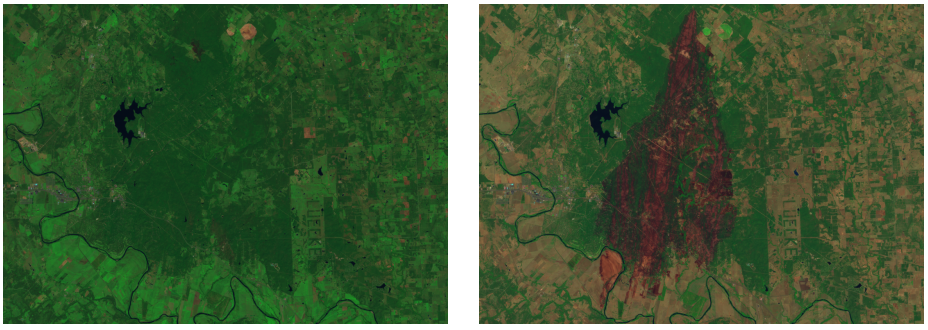
The *Bastrop fires* and *Bastrop fires* 2010/11 image datasets consist in Landsat 5 TM images having 6 spectral bands (from 450 nm to 2350 nm) at a spatial resolution of 30m acquired near Bastrop in Texas, USA. Both datasets use an image taken a week after the fire (11/09/2011) and differ in the image used as reference:

- *Bastrop fires*: the reference image has been acquired few days before (26/08/2011) the fire. The considered subset is of size 785×929 pixels. A (MidInfraRed-Green-Blue) composition representing well the fire is presented in Figure B.1.
- *Bastrop fires* 2010/11: the reference image has been acquired a year before (03/10/2010) the fire. The considered subset is of size 780×1085 pixels. A (R-G-B) composition is presented in Fig. B.2. It presents challenging additional vegetation changes.




(a) (b) (c) (d)

Figure B.1: The *Bastrop fires* multitemporal dataset (a) a week before the fire and (b) after the fire, (c) an official map[1]of the damages that helped setting (d) the groundtruth map.



(a) (b)

Figure B.2: The *Bastrop fires* 2010/11 multitemporal dataset (a) a year before the fire and (b) after the fire.

[0]Credits: Bastrop County Office of Emergency Management. Map can be downloaded at http://kut.org/wp-content/uploads/2011/09/BASTROP-MAP-2.jpg

### B.1.2 Gloucester floods

The *Gloucester floods* dataset consists in two SPOT XS images acquired before and after the floods in 2000 in the Gloucestershire in UK and made available in the context of the 2009 *IEEE* GRSS data fusion contest [Longbotham et al., 2012]. The considered subset (around the city of Tewkesbury at the confluence of the Avon and Severn Rivers) is of size 800×1600 pixels with a spatial resolution of 20m and 3 spectral bands (NearInfraRed-Red-Green). The two images are presented in Figure B.3. The groundtruth has been established over the entire subset. The challenge in detecting the flood extent is due to the noise in the post-flooding image, the limited information (only 3 bands) and the various crop changes occurring on the ground.

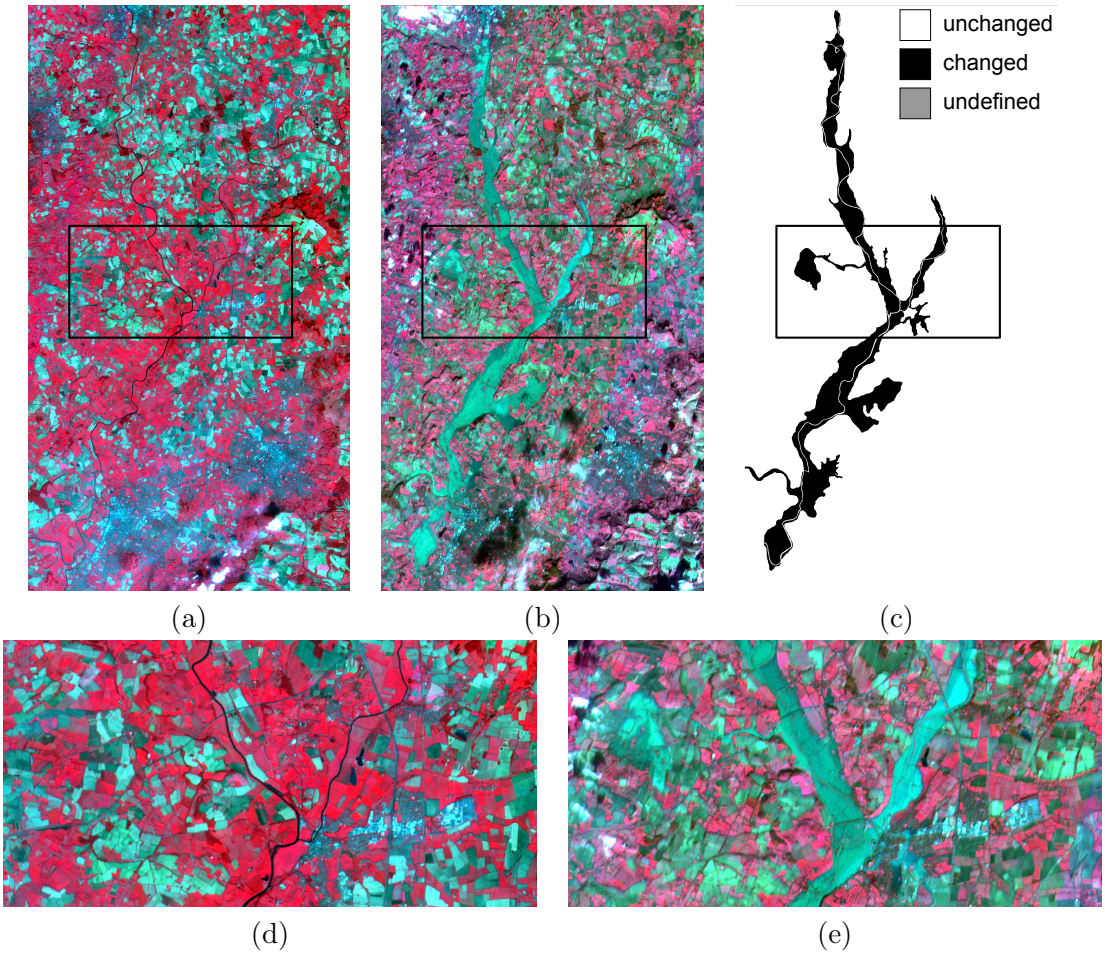



Figure B.3: The *Gloucester floods* multitemporal dataset from the Data Fusion Contest (a) before and (b) after floods delimited in the groundtruth (c). The subset used in the experiments is presented in (d) and (e).

## B.2 ROSIS

The Reflective Optics Spectrometric Imaging System (ROSIS) is an airborne spectrometer in the multispectral region ($430 - 960$ nm) with a spectral resolution of 4 nm per channel. The spatial resolution is generally under the meter pixel-resolution. The following image scene scenes over Pavia were provided by Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, Pavia university (Italy).

### B.2.1 Pavia DFC

The *Pavia DFC* image was part of the GRSS Data Fusion Contest of 2008 [Licciardi et al., 2009]. It is taken over the city of Pavia (Italy) with a spatial resolution of 1.3m. Among the spectral bands, 102 are selected and used. The scene is $1400 \times 512$ pixels and contains 5 broad classes of interest (Buildings, Roads, Water, Vegetation and Shadows) labeled over $206'009$ pixels.

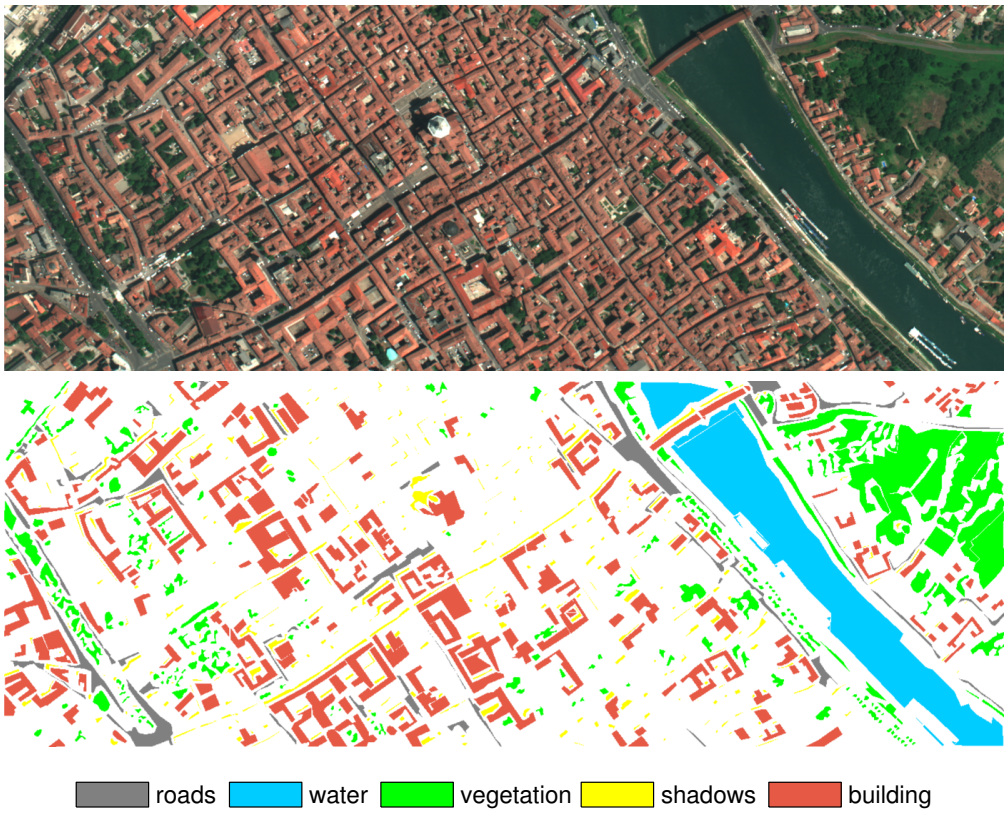



Figure B.4: The *Pavia DFC* hyperspectral dataset.

### B.2.2 Pavia center

The *Pavia Center* image[2] over the city center of Pavia has the same spatial and spectral resolution than *Pavia DFC*. The image subset is 731 × 328 pixels and has 9 classes of interest, namely *water*, *trees*, *meadows*, *bitumen*, *bare soil*, *asphalt*, *tiles*, *bricks* and *shadows*.

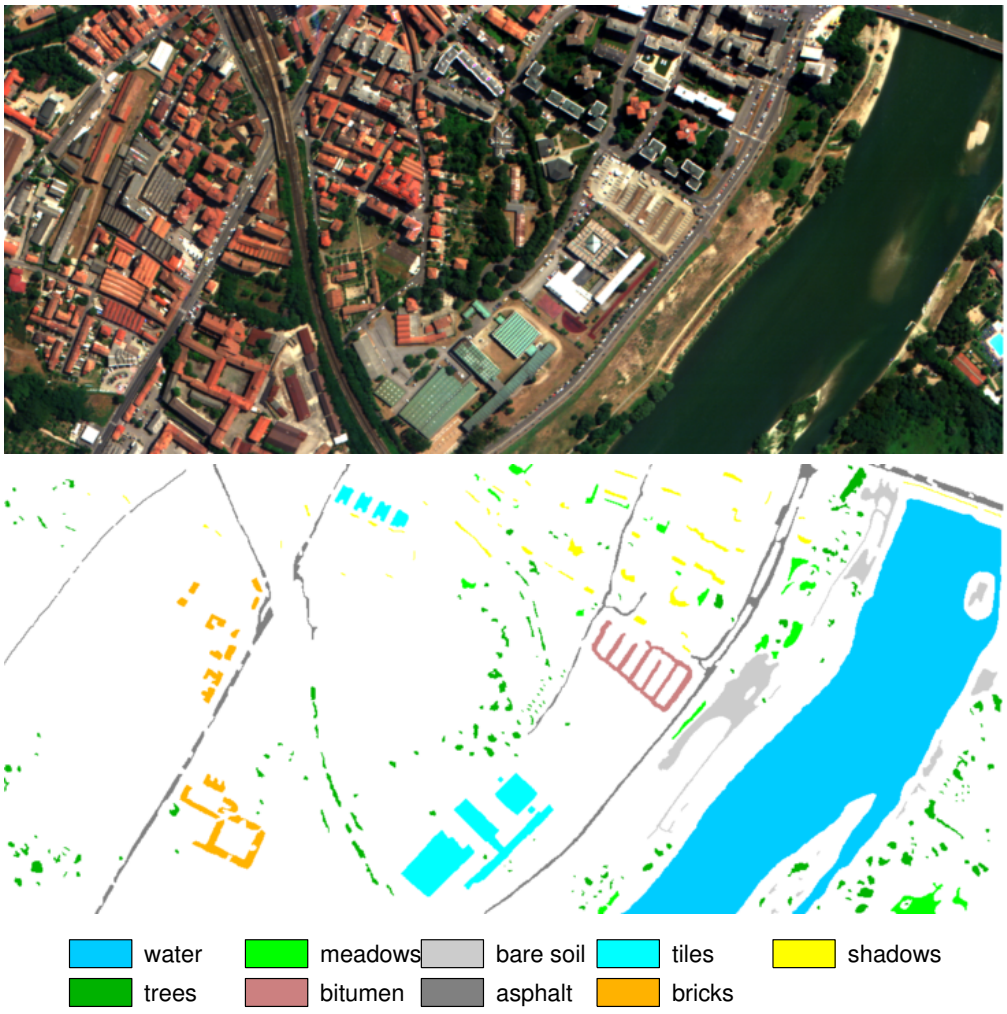



Figure B.5: The *Pavia center* hyperspectral dataset.

[2]The image can be downloaded from http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

## B.3 Faces and objects databases

In this section, we present image samples from the different databases considered in some of the experiments of Chapter 9: Yale B Extended, AR Faces, AT&T Faces, Sheffield Faces, COIL-20 and USPS.

**Yale B Extended** The Yale B Extended face dataset contains frontal photos of 38 subjects under 64 different illuminations [Georghiades et al., 2001]. The cropped face images of the first 10 persons, resized to $42 \times 48$ pixels, are used. For this dataset, it has been proven that the images of a single person lie close to a 9-D linear subspace [Basri and Jacobs, 2003]. Image samples are presented in Figure B.6.

Figure B.6: Samples from the two first subjects in *Yale B Extended* face dataset [Georghiades et al., 2001].

**AR Faces** The *AR* face database contains 4000 frontal images of 126 subjects. Each subject is photographed in 26 images taken in two sessions [Martinez and Benavente, 1998]. In comparison to the Yale B Extended database, the AR database contains variations of facial expressions, as well as of illumination. We consider the 20 first subjects and use grayscale images cropped and resized to $55 \times 40$ pixels. The images with occlusions (for instance scarfs and sun glasses) have been left aside, resulting in 14 images per subject. Image samples are presented in Figure B.7.

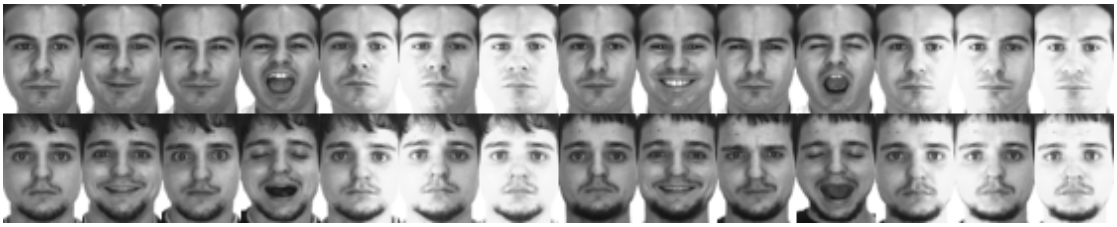


Figure B.7: Samples from the two first subjects in the AR Face database [Martinez and Benavente, 1998]

**AT&T Faces** The *AT&T* (ex-ORL) database of faces contains 40 subjects in frontal position with tolerance to certain movements [Samaria and Harter, 1994]. The images contain 10 facial expressions and sometime show occlusions (for instance glasses). Images of the 20 first subjects are used after being cropped and resized to $32 \times 32$ pixels. Sample of the images are presented in Figure B.8.

Figure B.8: Samples from the two first subjects in the *AT&T* (ex-ORL) database of faces [Samaria and Harter, 1994]

**Sheffield (ex-UMIST)** The *Sheffield* (ex-UMIST) Face database contains 20 subjects under poses from frontal to lateral (0 to 90°) position [Graham and Allinson, 1998]. Each subject has between 19 and 48 images. The images are cropped and rescaled to $28 \times 23$ pixels. The database contains no occlusions nor variations in expressions or illumination. This dataset presents non-linear characteristics with faces under varying poses, as presented in Figure B.9 for the subject 1 and 5 of the database.

Figure B.9: Samples from subjects 1 and 5 in the *Sheffield* (ex-UMIST) Face Database [Graham and Allinson, 1998].

**COIL Database** The COIL-20 database [S. A. Nene and Murase, 1996] contains 1440 images of 20 objects rotated over 360°. Each object is represented in 72 images. The processed images are cropped and rescaled to $32 \times 32$ pixels. The 20 objects and samples of images for two objects are presented in Figure B.10.

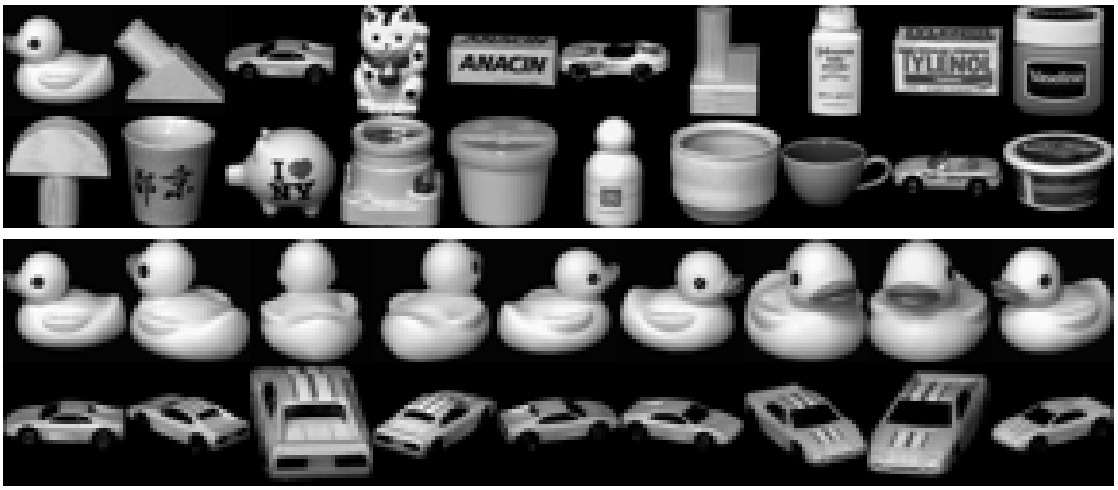


Figure B.10: Top: the 20 objects in the *COIL-20* database [S. A. Nene and Murase, 1996]. Bottom: image samples from objects 1 and 3.

**USPS Digits** The *USPS hand-written digits recognition* dataset are images of digits from the US Postal Service [Hull, 1994]. We consider a subset composed of the digits 1 and 2, each one having 1100 samples, partially presented in Figure B.11.

Figure B.11: The first four rows illustrate samples of digit 1 and the following rows show samples from digit 2 in the *USPS hand-written digits* dataset [Hull, 1994].

# C Appendix C

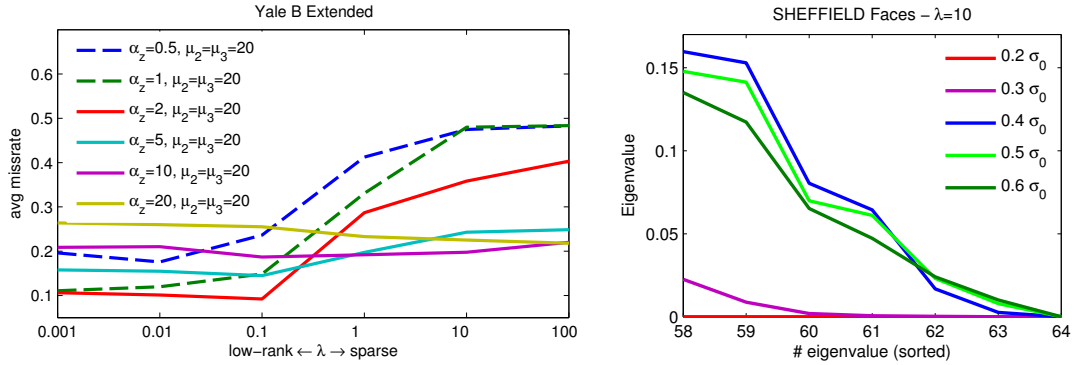
In this Appendix are presented some additional details on the results of the experiment 1 of Chapter 9.

## C.1 Additional face & object clustering results

**Trade-off between sparsity and low-rank properties** The flexibility of our proposed formulation has to adapt to the different types of subspaces. For dependent subspaces, we observe a phase-transition phenomenon with subspace recovery starting from a certain $\lambda$ and above, as observed in [Wang et al., 2013a] and depicted in Figure C.1 (a). Indeed from $\lambda = 1$ and above, a low missrate (good clustering) is obtained. As postulated, the best clustering is obtained for a certain trade-off between sparsity and low-rankness, here at $\lambda = 1$. Interestingly, the phase transition phenomenon is reversed when the representation error is more penalized (high $\alpha_z$), as observed in Figure C.1 (a). A high penalty of the representation error would probably use too many samples with some that are on a different subspace. In that situation, the low-rank constraint should be preferred.

**DB index and spectral gap sensitivity** In Figure C.2 the spectral embeddings of the subject 1 and 5 of the Sheffield Faces database (image samples in B.9) for different $\sigma$ values are presented. We clearly see the shrinkage of the clusters with $\sigma$ decreasing. The spread of the clusters is quantified by the $DB$ index. Following the indication of $DB$ would lead to choose very small $\sigma$. Therefore, samples from different subspaces will shrink onto similar clusters and it will be impossible to assign them correctly. The ideal bandwidth should be around $0.4\sigma_0$ as observed in Figure C.3 where the missrate for all the pairs of subjects in the Sheffield Faces dataset are represented for different RBF kernel bandwidth.

In Figure C.1 (b), the 6 last eigenvalues can be observed for different $\sigma$. The eigenvalues




(a) (b)

Figure C.1: (a) Average missrate over all pairs of subjects for different values of $\lambda$ in Yale B Extended Faces. An inversed phase-transition phenomenon is observed, with the best result obtained for $\alpha_z = 2$ and $\lambda = 0.1$. A large error should be tolerated to handle the strong illuminations differences. (b) The 6 last eigenvalues for subject 1 and 5 in the SHEFFIELD Faces database for different $\sigma$ and $\lambda = 10$ and $\alpha_z = \mu_2 = \mu_3 = 20$.

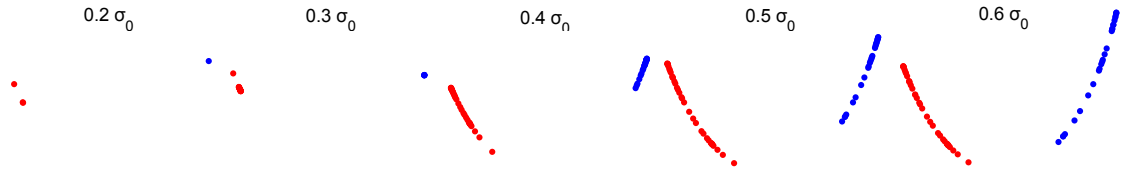



Figure C.2: The spectral embeddings of KLRSSC corresponding to different $\sigma$ for $\lambda = 10$ and $\alpha_z = \mu_2 = \mu_3 = 20$. The clusters shrink with decreasing $\sigma$

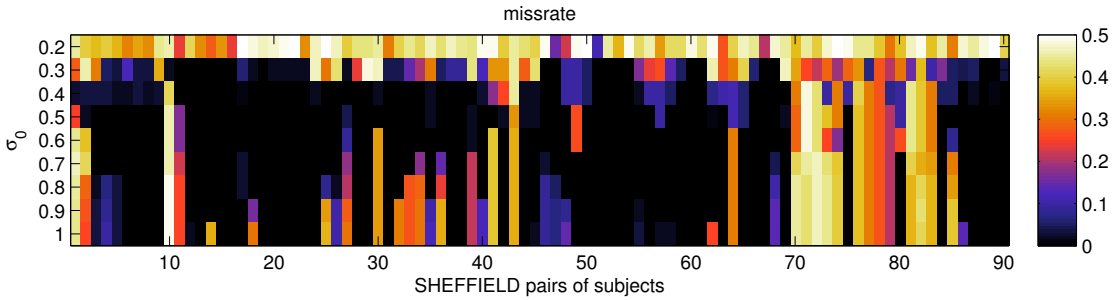



Figure C.3: Missrate of all the pairs of subjects in the SHEFFIELD Faces database for different values of $\sigma$. The optimal $\sigma$ value is around $0.4 \cdot \sigma_0$.

are the largest with $\sigma = 0.4 \cdot \sigma_0$ and decrease together with $\sigma$. This gives to the spectral gap a bias towards large $\sigma$ and prevents us from using it for parameter selection.

**RBF and Polynomial kernels** All the results with the RBF and the polynomial kernel for the different datasets are presented in Figures C.4 and C.5. The RBF kernel provides better results for all datasets and methods except the USPS digits dataset which gets a slightly better missrate for the KLRSC and KLRSSC with the polynomial kernel. The KSSC is on the contrary worse with the polynomial kernel.

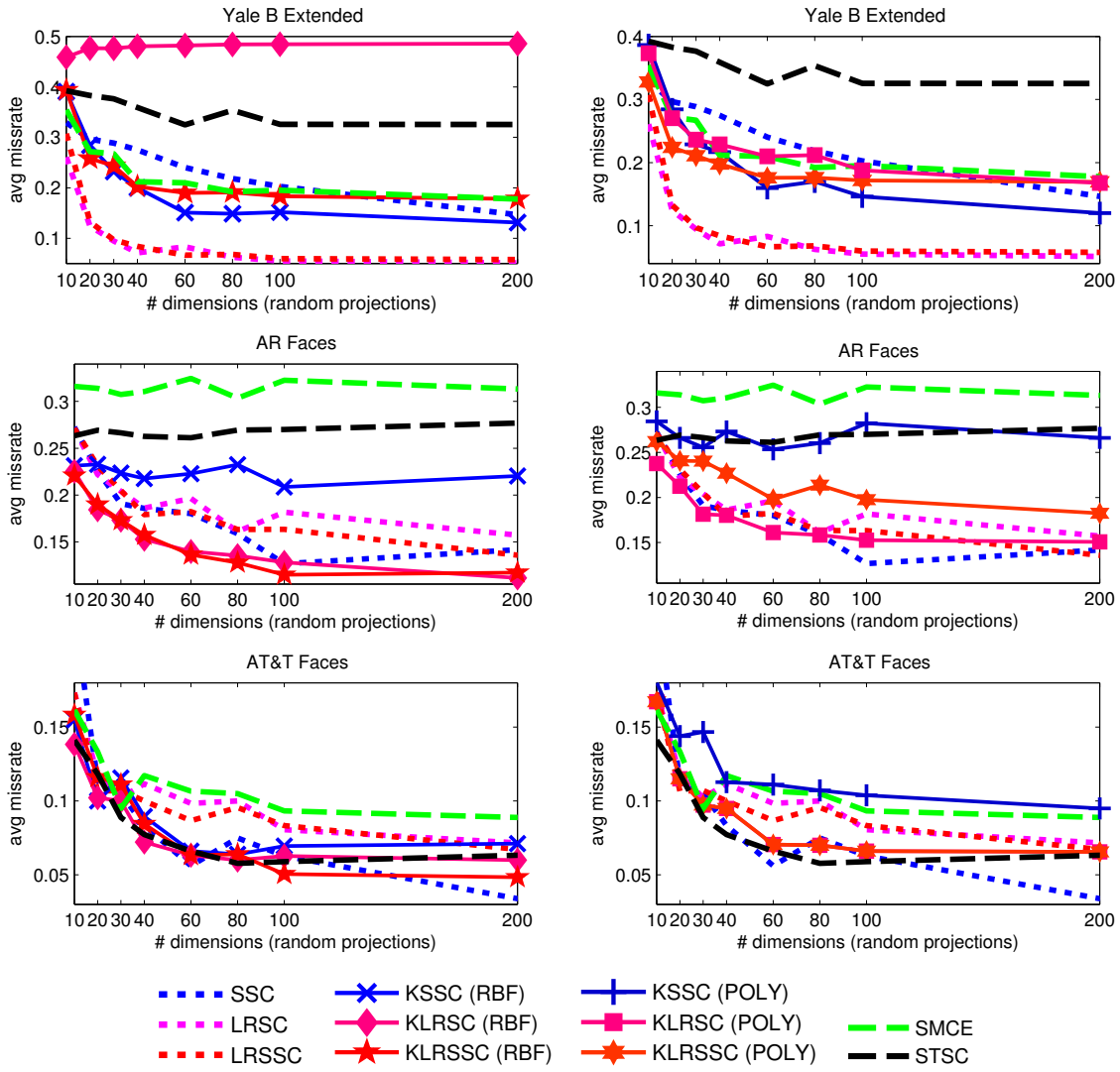



Figure C.4: Average missrate (the lower the better) over all pairs of subjects for various number of random projections for 3 different datasets: Yale B Extended Faces, AR Faces and AT&T Faces. The left column shows the results for the non-linear methods with the RBF kernel. The right column shows the results for the non-linear methods with the polynomial kernel.

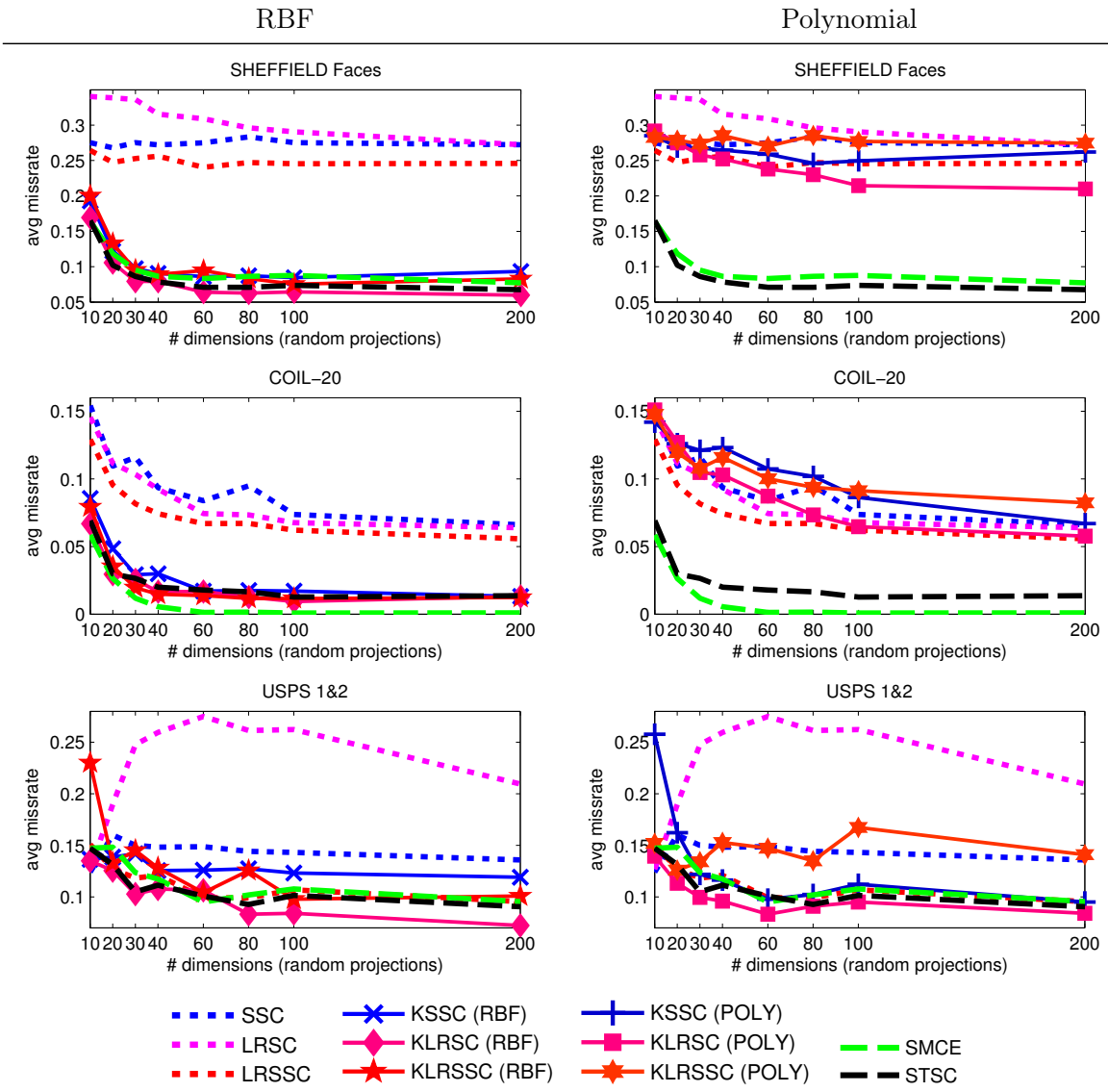



Figure C.5: Average missrate (the lower the better) over all pairs of subjects for various number of random projections for 3 different datasets: Sheffield Faces, COIL-20 and USPS-1&2. The left column shows the results for the non-linear methods with the RBF kernel. The right column shows the results for the non-linear methods with the polynomial kernel.

# Bibliography


C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *ACM SIGMOD Record*, pages 37–46, New York, NY, USA, 2001. ACM. ISBN 1-58113-332-4. doi: 10.1145/375663.375668.

O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3297–3304. IEEE, 2011.

S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2007. doi: 10.1109/CVPR.2007.383105.

F. Angiulli. Prototype-based domain description for one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1131–1144, 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.204.

J. Arenas-Garcia, K. Petersen, G. Camps-Valls, and L. Hansen. Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, 30(4):16–29, July 2013. ISSN 1053-5888. doi: 10.1109/MSP.2013.2250591.

E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *arXiv preprint arXiv:1301.2007*, 2013.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3): 337–404, 1950.

F. Bach, D. Heckerman, and E. Horvitz. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741, 2006.

F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *International Conference Machine Learning (ICML)*, page 6. ACM, 2004.

C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Exploiting manifold geometry in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):441–454, 2005.

G. H. Ball and D. J. Hall. Isodata, a novel method of data analysis and pattern classification. Technical report, DTIC Document, 1965.

A. Banerjee, P. Burlina, and C. Diehl. One-class svm for hyperspectral anomaly detection. In G. Camps-Valls and L. Bruzzone, editors, *Kernel methods for remote sensing data analysis*, pages 169–192. J. Wiley & Sons, NJ, USA, 2009.

R. G. Baraniuk. More is less: Signal processing and the data deluge. *Science*, 331(6018):717–719, 2011. doi: 10.1126/science.1197448.

V. Barnett and T. Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

S. Basu, I. Davidson, and K. Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *arXiv preprint*, abs/1203.4580, 2012.

L. A. Belanche. Developments in kernel design. In *European symposium on artifcial neural networks, computational intelligence and machine learning (ESANN), Bruges*, pages 369–378, 2013.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2002.

K. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, page 368. The MIT Press, 1999.

S. Bickel and T. Scheffer. Multi-view clustering. In *IEEE International Conference on Data Mining (ICDM)*, volume 4, pages 19–26, 2004.

C. M. Bishop. Novelty detection and neural network validation. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 141, pages 217–222. IET, 1994.

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 9999:2973–3009, 2010.

U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *Location and Context Awareness*, pages 192–206. Springer, 2009.

M. B. Blaschko, C. H. Lampert, and A. Gretton. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 133–145. Springer, 2008.

A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *International Conference Machine Learning (ICML)*, 2001.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational learning theory (COLT)*, pages 92–100. ACM, 1998.

J. Bolton and P. Gader. Multiple instance learning for hyperspectral image analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4232–4235. IEEE, 2010.

R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, pages 235–249, 2002.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

R. Borisyuk, M. Denham, F. Hoppensteadt, Y. Kazanovich, and O. Vinogradova. An oscillatory neural network model of sparse distributed memory and novelty detection. *Biosystems*, 58(1–3):265 – 272, 2000. ISSN 0303-2647. doi: http://dx.doi.org/10.1016/S0303-2647(00)00131-3.

R. M. Borisyuk and Y. B. Kazanovich. Oscillatory model of attention-guided object selection and novelty detection. *Neural Networks*, 17(7):899 – 915, 2004. ISSN 0893-6080. doi: http://dx.doi.org/10.1016/j.neunet.2004.03.005.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

F. Bovolo. A multilevel parcel-based approach to change detection in very high resolution multitemporal images. *IEEE Geoscience and Remote Sensing Letters*, 6(1):33–37, 2009.

F. Bovolo and L. Bruzzone. A split-based approach to unsupervised change detection in large size multitemporal images: application to Tsunami-damage assessment. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6):1658–1671, 2007a.

F. Bovolo and L. Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236, 2007b.

F. Bovolo, L. Bruzzone, and M. Marconcini. A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2070–2082, 2008.

F. Bovolo, G. Camps-Valls, and L. Bruzzone. A support vector domain method for change detection in multitemporal images. *Pattern Recognition Letters*, 31(10):1148–1154, 2010.

P. S. Bradley and O. L. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *ACM SIGMOD Record*, pages 93–104. ACM, 2000.

L. Bruzzone and M. Marconcini. Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1108–1122, 2009.

L. Bruzzone and C. Persello. A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2142 –2154, july 2009. ISSN 0196-2892. doi: 10.1109/TGRS.2008.2011983.

L. Bruzzone and D. Prieto. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. 2001.

L. Bruzzone and S. Serpico. An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):858–867, 1997. ISSN 0196-2892. doi: 10.1109/36.602528.

L. Bruzzone, M. Chi, and M. Marconcini. A novel transductive SVM for semisupervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363–3373, 2006.

A. Buades, B. Coll, and J.-M. Morel. Image denoising methods. a new nonlocal principle. *SIAM Review*, 52(1):113–147, 2010.

J.-F. Cai and S. Osher. Fast singular value thresholding without singular value decomposition. *UCLA CAM Report*, 5, 2010.

F. Camastra and A. Verri. A novel kernel method for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):801–805, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.88.

C. Campbell and K. P. Bennett. A linear programming approach to novelty detection. In *Advances in Neural Information Processing Systems*, volume 13, page 395. The MIT Press, 2001.

G. Camps-Valls and L. Bruzzone. *Kernel methods for remote sensing data analysis*. Wiley Online Library, 2009.

G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Mari, L. Alonso, J. Calpe-Maravilla, and J. Moreno. Multitemporal image classification and change detection with kernels. In *SPIE International Symposium Remote Sensing XII*, volume 6365, 2006.

G. Camps-Valls, T. Bandos Marsheva, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3044–3054, 2007.

G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Rojo-Álvarez, and M. Martínez-Ramón. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008.

G. Camps-Valls, D. Tuia, L. Gómez-Chova, S. Jiménez, , and J. Malo. *Remote Sensing Image Processing*, volume 5 of *Synthesis Lectures on Image, Video, and Multimedia Processing*. Morgan & Claypool Publishers, 2011. doi: 10.2200/S00392ED1V01Y201107IVM012.

G. Camps-Valls, D. Tuia, L. Bruzzone, and J. Atli Benediktsson. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31(1):45–54, 2014. ISSN 1053-5888. doi: 10.1109/MSP.2013.2279179.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58 (3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395.

W. Cao and R. Haralick. Nonlinear manifold clustering by dimensionality. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 920–924, 2006. doi: 10.1109/ICPR.2006.865.

L. Capobianco, A. Garzelli, and G. Camps-Valls. Target detection with semisupervised kernel orthogonal subspace projection. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3822–3833, Jul 2009.

G. A. Carpenter, S. Grossberg, and D. B. Rosen. Art 2-a: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4(4):493–504, 1991.

J. C. Castilla-Rubio and S. Willis. Planetary skin: A global platform for a new era of collaboration, 2009.

G. C. Cawley and N. L. Talbot. Reduced rank kernel ridge regression. *Neural Processing Letters*, 16(3): 293–302, 2002. ISSN 1370-4621. doi: 10.1023/A:1021798002258.

T. Celik. Multiscale change detection in multitemporal satellite images. *IEEE Geoscience and Remote Sensing Letters*, 6(4):820 –824, oct. 2009.

J. Chan, J. Bailey, and C. Leckie. Discovering correlated spatio-temporal changes in evolving graphs. *Knowledge and Information Systems*, 16(1):53–96, 2008.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computer Survey*, 41(3): 15:1–15:58, July 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882.

C.-I. Chang. Relationship among orthogonal subspace projection, constrained energy minimization and rx-algorithm. In *Proceedings of SPIE (AeroSense)*, volume 4725, pages 490–500, 2002. doi: 10.1117/12.478782.

C.-I. Chang, C.-C. Wu, and C.-T. Tsai. Random n-finder (n-findr) endmember extraction algorithms for hyperspectral imagery. *IEEE Transactions on Image Processing*, 20(3):641–656, 2011. ISSN 1057-7149. doi: 10.1109/TIP.2010.2071310.

L. Chapel, T. Burger, N. Courty, and S. Lefevre. Classwise hyperspectral image classification with perturbo method. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6883–6886, 2012. doi: 10.1109/IGARSS.2012.6352581.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. 2004.

O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 585–592, 2002.

O. Chapelle, B. Schölkopf, A. Zien, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, MA:, 2006.

A. Charles, B. Olshausen, and C. Rozell. Learning sparse codes for hyperspectral imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):963–978, 2011. ISSN 1932-4553. doi: 10.1109/JSTSP.2011.2149497.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference Machine Learning (ICML)*, pages 129–136. ACM, 2009.

G. Chen and G. Lerman. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3):317–330, 2009. ISSN 0920-5691. doi: 10.1007/s11263-008-0178-9.

G. Chen, S. Atev, and G. Lerman. Kernel spectral curvature clustering (kscc). In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 765–772. IEEE, 2009.

G. H. Chen, C. Wachinger, and P. Golland. Sparse projections of medical images onto manifolds. In J. C. Gee, S. Joshi, K. M. Pohl, W. M. Wells, and L. Zallei, editors, *Information Processing in Medical Imaging*, volume 7917 of *Lecture Notes in Computer Science*, pages 292–303. Springer Berlin Heidelberg, 2013a.

J. Chen, X. Chen, X. Cui, and J. Chen. Change vector analysis in posterior probability space: A new method for land cover change detection. *IEEE Geoscience and Remote Sensing Letters*, PP(99):317 –321, 2010.

X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, 2011.

Y. Chen, X. Dang, H. Peng, and H. Bart Jr. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 288–305, 2008.

Y. Chen, N. M. Nasrabadi, and T. D. Tran. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3973–3985, 2011a.

Y. Chen, N. M. Nasrabadi, and T. D. Tran. Simultaneous joint sparsity model for target detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 8(4):676–680, 2011b.

Y. Chen, N. M. Nasrabadi, and T. D. Tran. Sparse representation for target detection in hyperspectral imagery. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):629–640, 2011c.

Y. Chen, N. Nasrabadi, and T. Tran. Hyperspectral image classification via kernel sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):217–231, 2013b. ISSN 0196-2892. doi: 10.1109/TGRS.2012.2201730.

V. Cherkassky and F. M. Mulier. *Learning from data: concepts, theory, and methods*. Wiley. com, 2 edition, 2007.

J. P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

N. Courty, T. Burger, and J. Laurent. Perturbo: A new classification algorithm based on the spectrum perturbations of the laplace-beltrami operator. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6911 of *Lecture Notes in Computer Science*, pages 359–374. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-23780-5_33.

N. Courty, T. Burger, and P.-F. Marteau. Geodesic analysis on the gaussian rkhs hypersphere. In P. Flach, T. Bie, and N. Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7523 of *Lecture Notes in Computer Science*, pages 299–313. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-33460-3_25.

F. G. Cozman, I. Cohen, and M. Cirelo. Unlabeled data can degrade classification performance of generative classifiers. In *FLAIRS Conference*, pages 327–331, 2002.

M. Dalla Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone. An unsupervised technique based on morphological filters for change detection in very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 5(3):433 – 437, 2008.

D. Dasgupta and S. Forrest. Novelty detection in time series data using ideas from immunology. In *International Conference on Intelligent Systems*, pages 82–87, 1996.

M. Davy and S. Godsill. Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–1313. IEEE, 2002.

F. De Comité, F. Denis, R. Gilleron, and F. Letouzey. Positive and unlabeled examples help learning. In *Algorithmic Learning Theory*, pages 219–230. Springer, 1999.

F. de Morsier, M. Borgeaud, C. Kuchler, V. Gass, and J.-P. Thiran. Semi-supervised and unsupervised novelty detection using nested support vector machines. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 7337–7340, July 2012a. doi: 10.1109/IGARSS.2012.6351935.

F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Unsupervised change detection via hierarchical support vector clustering. In *IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS-ICPR)*, pages 1–4. IAPR, 2012b. doi: 10.1109/PPRS.2012.6398309.

F. de Morsier, M. Borgeaud, C. Küchler, A. Vogel, V. Gass, and J.-P. Thiran. Robust phase-correlation based registration of airborne videos using motion estimation. In *Earth Observation of Global Changes (EOGC)*, pages 37–48. Springer, 2013a.

F. de Morsier, M. Rasamimalala, D. Tuia, M. Borgeaud, S. Rakotoniaina, S. Rakotondraompiana, and J.-P. Thiran. Detection of floods in sar images with non-linear kernel clustering and topographic prior. In *European Signal Processing Conference (EUSIPCO)*, September 9-13 2013b.

F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Semi-supervised novelty detection using svm entire solution path. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):1939–1950, 2013c. ISSN 0196-2892. doi: 10.1109/TGRS.2012.2236683.

F. de Morsier, C. Nathalie, D. DeMaris, M. Gabrani, A. Gotovos, and A. Krause. Fast detection of novel problematic patterns based on dictionary learning and prediction of their lithographic difficulty. In *Conference on Optical Microlithography XXVII, part of SPIE Advanced Lithography*, San Jose, California, USA, February 23-27 2014a.

F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Cluster validity measure and merging system for hierarchical clustering considering outlier. *Pattern Recognition, submitted to*, 2014b.

F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Non-linear low-rank and sparse representation for hyperspectral image analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2014c.

F. de Morsier, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran. Non-linear low-rank and sparse subspace clustering. *IEEE International Journal of Computer Vision, submitted to*, 2014d.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

F. Denis, R. Gilleron, A. Laurent, and M. Tommasi. Text classification and co-training from positive and unlabeled examples. In *International Conference Machine Learning Workshop (ICML Workshop: the continuum from labeled to unlabeled data)*, pages 80–87, 2003.

S. Deris, H. Alashwal, and M. Othman. One-class support vector machines for protein-protein interactions prediction. *International Journal of Biological and Medical Sciences*, 1(2):120–127, 2006.

N. Dobigeon, J.-Y. Tourneret, C. Richard, J. Bermudez, S. McLaughlin, and A. Hero. Nonlinear unmixing of hyperspectral images: Models and algorithms. *IEEE Signal Processing Magazine*, 31(1):82–94, 2014. ISSN 1053-5888. doi: 10.1109/MSP.2013.2279274.

A. R. Doherty and A. F. Smeaton. Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors*, 10(3):1423–1446, 2010.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

R. P. Duin and E. Pekalska. Open issues in pattern recognition. In *Computer Recognition Systems*, pages 27–42. Springer, 2005.

J. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.

E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk. Greedy feature selection for subspace clustering. *arXiv preprint*, 2013.

E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems*, pages 55–63, 2011.

E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.57.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220. ACM, 2008.

E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *International Conference Machine Learning (ICML)*, 2000.

E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

T. Evgeniou, T. Poggio, M. Pontil, and A. Verri. Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38(4):421–432, 2002.

L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

B. Fernandes, G. D. C. Cavalcanti, and T. Ren. Autoassociative pyramidal neural network for face verification. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1612–1617, 2011. doi: 10.1109/IJCNN.2011.6033417.

M. Filippone and G. Sanguinetti. A perturbative approach to novelty detection in autoregressive models. *IEEE Transactions on Signal Processing*, 59(3):1027–1036, 2011. ISSN 1053-587X. doi: 10.1109/TSP.2010.2094609.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

G. Foody and A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6):1335 – 1343, june 2004. ISSN 0196-2892. doi: 10.1109/TGRS.2004.827257.

J. E. Fowler and Q. Du. Anomaly detection and reconstruction from random projections. *IEEE Transactions on Image Processing*, 21(1):184–195, 2012.

G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text classification without negative examples revisit. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):6–20, 2006.

S. Gao, I. Tsang, and L.-T. Chia. Sparse representation with kernels. *IEEE Transactions on Image Processing*, 22(2):423–434, 2013. ISSN 1057-7149. doi: 10.1109/TIP.2012.2215620.

A. B. Gardner, A. M. Krieger, G. Vachtsevanos, and B. Litt. One-class novelty detection for seizure analysis from intracranial eeg. *Journal of Machine Learning Research*, 7:1025–1044, 2006.

A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

N. Goernitz, M. Kloft, K. Rieck, and U. Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research (JAIR)*, 46:235–262, 2013. doi: 10.1613/jair.3623.

A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007. doi: 10.1109/CVPR.2007.383235.

L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe. Semi-supervised image classification with Laplacian support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 5(4):336–340, 2008.

L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla. Mean map kernel methods for semisupervised cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1): 207–220, 2010.

D. Gong, X. Zhao, and G. Medioni. Robust multiple manifolds structure learning. In *International Conference Machine Learning (ICML)*, 2012.

D. Graham and N. Allinson. Characterizing virtual eigensignatures for general purpose face recognition. *Face Recognition: From Theory to Applications; NATO ASI Series F, Computer and Systems Sciences*, 163:446–456, 1998.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, Mar. 2012. ISSN 1532-4435.

Y. Gu and K. Feng. L1-graph semisupervised learning for hyperspectral image classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1401–1404, 2012. doi: 10.1109/IGARSS.2012.6351274.

Y. Gu, Y. Liu, and Y. Zhang. A selective kpca algorithm based on high-order statistics for anomaly detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 5(1):43–47, 2008. ISSN 1545-598X. doi: 10.1109/LGRS.2007.907304.

S.-M. Guo, L.-C. Chen, and J. S. H. Tsai. A boundary method for outlier detection based on support vector domain description. *Pattern Recognition*, 42(1):77–83, 2009.

I. Guyon. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *International Conference on Machine Learning (ICML)*, page 47. ACM, 2004.

M. Hansen, K. Sjöstrand, H. Ólafsdóttir, H. Larsson, M. Stegmann, and R. Larsen. Robust pseudo-hierarchical support vector clustering. *Image Analysis*, pages 808–817, 2007.

S. Harmeling, G. Dornhege, D. Tax, F. Meinecke, and K.-R. Müller. From outliers to prototypes: ordering data. *Neurocomputing*, 69(13):1608–1618, 2006.

J. Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

V. Hautamaki, I. Karkkainen, and P. Franti. Outlier detection using k-nearest neighbour graph. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pages 430–433. IEEE, 2004.

P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In *Advances in Neural Information Processing Systems*, pages 946–952, 2000.

X. He and P. Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, volume 16, page 153, 2004.

G. Healey and D. Slater. Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6):2706–2717, 1999.

S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. In *UbiComp 2006: Ubiquitous Computing*, pages 177–193. Springer, 2006.

H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968. ISSN 0018-9448. doi: 10.1109/TIT.1968.1054102.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, May 1994.

E. B. Hunt. *Concept learning: An information processing problem.* John Wiley & Sons Inc, 1962.

G. K. Ian Dowman, Karsten Jacobsen and R. Sandau. *High resolution optical satellite imagery.* Whittles Publishing, 2012.

J. Im, J. R. Jensen, and M. E. Hodgson. Optimizing the binary discriminant function in change detection applications. *Remote Sensing of Environment*, 112(6):2761–2776, 2008.

T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, pages 487–493, 1998.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computer Survey*, 31(3): 264–323, Sept. 1999. ISSN 0360-0300. doi: 10.1145/331499.331504.

N. Japkowicz. *Concept-learning in the absence of counter-examples: an autoassociation-based approach to classification.* PhD thesis, Rutgers, The State University of New Jersey, 1999.

N. Japkowicz, C. Myers, and M. A. Gluck. A Novelty Detection Approach to Classification. In *International joint Conference on Artificial Intelligence (IJCAI)*, pages 518–523, 1995.

T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference Machine Learning (ICML)*, volume 99, pages 200–209, 1999.

P. Juszczak, D. M. Tax, E. Pe˛kalska, and R. P. Duin. Minimum spanning tree based one-class classifier. *Neurocomputing*, 72(7–9):1859 – 1869, 2009. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2008.05.003.

K. Kamvar, S. Sepandar, K. Klein, D. Dan, M. Manning, and C. Christopher. Spectral learning. In *International joint Conference on Artificial Intelligence (IJCAI)*. Stanford InfoLab, 2003.

S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.

M. Kemmler, E. Rodner, and J. Denzler. One-class classification with gaussian processes. *Computer Vision-ACCV 2010*, pages 489–500, 2011.

N. Keshava and J. F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, 2002.

H. Kim and J. Lee. Clustering based on gaussian processes. *Neural Computation*, 19(11):3088–3107, 2007.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro. Semicca: Efficient semi-supervised learning of canonical correlations. In *International Conference on Pattern Recognition (ICPR)*, pages 2933–2936, 2010. doi: 10.1109/ICPR.2010.719.

R. T. Knight. Contribution of human hippocampal region to novelty detection. *Nature*, 383(6597): 256–259, 1996.

T. Kohonen. *Self-organizing maps*, volume 30. Springer, 2001.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

I. Kononenko and M. Kukar. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited, 2007. ISBN 1904275214, 9781904275213.

O. Koyejo and J. Ghosh. Mipps: A generative model for multi-manifold clustering. In *Fall Symposium on Manifold Learning and Its Applications*, pages 18–25. AAAI Press, 2009.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Loop: local outlier probabilities. In *Proc. of the 18th ACM Conference on Information and Knowledge Management*, pages 1649–1652. ACM, 2009.

K. Krishnamurthy, R. Willett, and M. Raginsky. Target detection performance bounds in compressive imaging. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–19, 2012.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

B. Kulis, S. Basu, I. Dhillon, and R. Mooney. Semi-supervised graph clustering: a kernel approach. *Machine Learning*, 74(1):1–22, 2009.

A. Kumar, P. Rai, and H. D. Iii. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1413–1421, 2011.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(1):981–1006, Apr. 2012. ISSN 1532-4435.

R. Kwitt and U. Hofmann. Unsupervised anomaly detection in network traffic by means of robust pca. In *International Conference on Computing in the Global Information Technology (ICCGI)*, pages 37–37. IEEE, 2007.

H. Kwon and N. M. Nasrabadi. Kernel rx-algorithm: a nonlinear anomaly detector for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(2):388–397, 2005.

V. Laparra, G. Camps-Valls, and J. Malo. Iterative gaussianization: From ica to random rotations. *IEEE Transactions on Neural Networks*, 22(4):537–549, 2011.

V. Laparra, D. Tuia, S. Jímenez, G. Camps-Valls, and J. Malo. Nonlinear data description with principal polynomial analysis. In *IEEE Machine Learning for Signal Processing Workshop (MLSP)*, Santander, Spain, 2012.

F. Lauer and C. Schnorr. Spectral clustering of linear subspaces for motion segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 678–685, 2009.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

G. Lee and C. Scott. The one class support vector machine solution path. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–521. IEEE, 2007.

G. Lee and C. Scott. Nested support vector machines. *IEEE Transactions on Signal Processing*, 58(3): 1648–1660, 2010.

J. Lee and D. Lee. An improved cluster labeling method for support vector clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):461–464, 2005.

K. Lee, D. Kim, D. Lee, and K. Lee. Improving support vector data description using local density degree. *Pattern Recognition*, 38(10):1768–1771, 2005.

K. Lee, D. Kim, K. Lee, and D. Lee. Density-induced support vector data description. *IEEE Transactions on Neural Networks*, 18(1):284–289, 2007.

W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *IEEE Symposium on Security and Privacy*, pages 130–143, 2001. doi: 10.1109/SECPRI.2001.924294.

W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *International Conference Machine Learning (ICML)*, volume 3, pages 448–455, 2003.

Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353. IEEE, 2012.

F. Letouzey, F. Denis, and R. Gilleron. Learning from positive and unlabeled examples. In *Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

H. Li, Y. Gao, and J. Sun. Fast kernel sparse representation. In *2011 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 72–77, 2011a. doi: 10.1109/DICTA.2011.20.

J. Li, J. Bioucas-Dias, and A. Plaza. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (11):4085 –4098, 2010.

W. Li, Q. Guo, and C. Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):717 –725, feb. 2011b. ISSN 0196-2892. doi: 10.1109/TGRS.2010.2058578.

X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In *International joint Conference on Artificial Intelligence (IJCAI)*, volume 18, pages 587–594. Lawrence Erlbaum Associates LTD, 2003.

G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba. Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S data fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3857–3865, 2009.

B. Liu, W. Lee, P. Yu, and X. Li. Partially supervised classification of text documents. In *International Conference Machine Learning (ICML)*, pages 387–394. Citeseer, 2002.

B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *IEEE International Conference on Data Mining (ICDM)*, pages 179–186. IEEE, 2003.

G. Liu and S. Yan. Active subspace: Toward scalable low-rank learning. *Neural Computation*, 24(12): 3371–3394, Sept. 2012. ISSN 0899-7667. doi: 10.1162/NECO_a_00369.

G. Liu, H. Xu, and S. Yan. Exact subspace segmentation and outlier detection by low-rank representation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 703–711, 2012.

G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.88.

R. Liu, R. Hao, and Z. Su. Mixture of manifolds clustering via low rank embedding. *Journal of Information and Computational Science*, 8(5):725–737, 2011.

N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du. Multi-modal change detection, application to the detection of floodedareas: outcome of the 2009-2010 data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):331–342, 2012.

C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *European Conference on Computer Vision (ECCV)*, pages 347–360. Springer, 2012.

D. Lunga, S. Prasad, M. Crawford, and O. Ersoy. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine*, 31(1):55–66, 2014. ISSN 1053-5888. doi: 10.1109/MSP.2013.2279894.

L. Manevitz and M. Yousef. One-class SVMs for document classification. *Journal of Machine Learning Research*, 2:139–154, 2002.

D. Marcos Gonzalez, F. de Morsier, G. Matasci, D. Tuia, and J.-P. Thiran. Hierarchical sparse representation for dictionary-based classification of hyperspectral images. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), submitted to*, Lausanne, Switzerland, June 2014. IEEE.

M. Markou and S. Singh. Novelty detection: a review-part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003a.

M. Markou and S. Singh. Novelty detection: a review-part 2: neural network based approaches. *Signal processing*, 83(12):2499–2521, 2003b.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, pages 570–576, 1998.

A. Martinez and R. Benavente. The ar face database. *CVC Technical Report*, (24), 1998.

A. Marx and S. Goward. Remote sensing in human rights and international humanitarian law monitoring: concepts and methods. *Geographical Review*, 103(1):100–111, 2013. ISSN 1931-0846. doi: 10.1111/j.1931-0846.2013.00188.x.

A. Marx and T. Loboda. Landsat-based early warning system to detect the destruction of villages in darfur, sudan. *Remote Sensing of Environment*, 136(0):126 – 134, 2013. ISSN 0034-4257. doi: http://dx.doi.org/10.1016/j.rse.2013.05.006.

J. Mas. Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing*, 20(1):139–152, 1999.

G. Matasci, F. de Morsier, M. Kanevski, and D. Tuia. Domain adaptation in remote sensing through cross-image synthesis with dictionaries. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS), submitted to*, Quebec city, Canada, July 2014.

L. Meng and J. Kerekes. Object tracking using high resolution satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (99):1–1, 2012.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446, 1909.

G. Mercier and F. Girard-Ardhuin. Partially supervised oil-slick detection by sar imagery using kernel expansion. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10):2839–2846, 2006.

D. Miljković. Novelty detection in machine vibration data based on cluster intraset distance. In *MIPRO 2008*, 2008.

E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 464–471. IEEE, 2000.

J. D. Miller and A. E. Thode. Quantifying burn severity in a heterogeneous landscape with a relative version of the delta normalized burn ratio (dnbr). *Remote Sensing of Environment*, 109(1):66–80, 2007.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. The MIT Press, 2012.

D. Moody, S. Brumby, J. Rowland, and C. Gangodagamage. Unsupervised land cover classification in multispectral imagery with sparse representations on learned dictionaries. In *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–10, 2012. doi: 10.1109/AIPR.2012.6528190.

P. J. Moreno, P. P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, page None, 2003.

G. Mountrakis, J. Im, and C. Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

M. M. Moya and D. R. Hush. Network constraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

S. Muggleton, editor. *Learning from Positive Data*, volume 1314 of *Lecture Notes in Computer Science*, 1997. Springer. ISBN 3-540-63494-0.

J. Muñoz-Marí, L. Bruzzone, and G. Camps-Valls. A support vector domain description approach to supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 45(8):2683–2692, 2007.

J. Muñoz-Marí, F. Bovolo, Gómez-Chova, L. Bruzzone, and G. Camp-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8):3188 –3197, aug. 2010. ISSN 0196-2892. doi: 10.1109/TGRS.2010.2045764.

D. T. Munroe and M. G. Madden. Multi-class and single-class classification approaches to vehicle model recognition from images. *International Conference on Artificial Intelligence and Computer Science (AICS)*, 2005.

B. Nasihatkon and R. Hartley. Graph connectivity in sparse subspace clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2137–2144, 2011. doi: 10.1109/CVPR.2011.5995679.

N. Nasrabadi. Hyperspectral target detection : An overview of current and future challenges. *IEEE Signal Processing Magazine*, 31(1):34–44, 2014. ISSN 1053-5888. doi: 10.1109/MSP.2013.2278992.

A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Kernel dictionary learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2021–2024, 2012.

M. H. Nguyen and F. De la Torre. Robust kernel principal component analysis. Advances in Neural Information Processing Systems, 2008.

N. Nguyen, N. Nasrabadi, and T. Tran. Robust multi-sensor classification via joint sparse representation. In *Proceedings of the 14th International Conference on Information Fusion (FUSION)*, pages 1–8, 2011.

A. A. Nielsen. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–478, 2006.

A. A. Nielsen, K. Conradsen, and J. J. Simpson. Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64:1–19, 1998.

N. J. Nilsson. *Learning machines.* McGraw-Hill, N.Y., 1965.

F. Odone, A. Barla, and A. Verri. Building kernels from binary strings for image matching. *IEEE Transactions on Image Processing*, 14(2):169–180, 2005.

E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. *AI Memo (MIT)*, 1997.

N. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6): 847–857, 1997.

M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, pages 1410–1418, 2009.

J. Paola and R. Schowengerdt. A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. *International Journal of remote sensing*, 16(16):3033–3058, 1995.

S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *IEEE International Conference on Data Engineering*, pages 315–326. IEEE, 2003.

M. Parente and A. Plaza. Survey of geometric and statistical unmixing algorithms for hyperspectral images. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4. IEEE, 2010.

L. Parra, G. Deco, and S. Miesbach. Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2):260–269, 1996.

V. Patel, H. V. Nguyen, and R. Vidal. Latent space sparse subspace clustering. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

D. Pelleg and A. W. Moore. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems*, pages 1073–1080, 2004.

X. Peng, L. Zhang, and Z. Yi. An out-of-sample extension of sparse subspace clustering and low rank representation for clustering large scale data sets. *arXiv preprint arXiv:1309.6487*, pages –, 2013.

J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods: Support Vector Learning*, 208(MSR-TR-98-14):1–21, 1998.

A. Plaza, P. Martinez, R. Perez, and J. Plaza. A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 42(3):650–663, 2004. ISSN 0196-2892. doi: 10.1109/TGRS.2003.820314.

D. Potin, P. Vanheeghe, E. Duflos, and M. Davy. An abrupt change detection algorithm for buried landmines localization. *IEEE Transactions on Geoscience and Remote Sensing*, 44(2):260–272, 2006.

T. Produit, D. Tuia, F. de Morsier, and F. Golay. Picture density in the space of geographic features: mapping landscape attractiveness. In *International Workshop on Environmental Multimedia Retrieval (in conjunction with ACM ICMR)*, Glasgow, U.K., April 2014.

H. Qi and S. Hughes. Using the kernel trick in compressive sensing: Accurate signal recovery from fewer measurements. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3940–3943. IEEE, 2011.

R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.

K. N. Ramamurthy and A. Spanias. Optimized measurements for kernel compressive sensing. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 1443–1446. IEEE, 2011.

T. S. Ranoeliarivao, F. de Morsier, D. Tuia, S. Rakotoniaina, M. Borgeaud, J.-P. Thiran, and S. Rakotondraompiana. Multisource clustering of remote sensing images with entropy-based dempster-shafer fusion. In *European Signal Processing Conference (EUSIPCO)*, September 9-13 2013.

C. Rasmussen and C. Williams. Gaussian processes for machine learning. 2006.

F. Ratle, M. Kanevski, A.-L. Terrettaz-Zufferey, P. Esseiva, and O. Ribaux. A comparison of one-class classifiers for novelty detection in forensic case data. In *Intelligent Data Engineering and Automated Learning (IDEAL)*, pages 67–76. Springer, 2007.

F. Ratle, G. Camps-Valls, and J. Weston. Semisupervised neural networks for efficient hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2271 –2282, may 2010. ISSN 0196-2892. doi: 10.1109/TGRS.2009.2037898.

I. S. Reed and X. Yu. Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(10):1760–1770, 1990.

J. A. Richards. *Remote sensing digital image analysis: an introduction*. Springer, 2013.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

J. Rissanen. *Minimum-Description-Length Principle*. Wiley Online Library, 1985.

J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 223–239, 1987.

S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

S. J. Roberts. Novelty detection using extreme value statistics. *IEE Proceedings-Vision, Image and Signal Processing*, 146(3):124–129, 1999.

L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

B. A. Rowland and A. S. Maida. Spatiotemporal novelty detection using resonance networks. In *Proc. of the 17th Annual Florida AI Research Society Conference (FLAIRS)*, pages 676–681, 2004.

J. Ryan, M.-J. Lin, and R. Miikkulainen. Intrusion detection with neural networks. In *Advances in Neural Information Processing Systems*, pages 943–949. MORGAN KAUFMANN PUBLISHERS, 1998.

S. K. N. S. A. Nene and H. Murase. Columbia object image library (coil-20). *Technical Report CUCS-005-96*, February 1996.

F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Proc. of Workshop on Applications of Computer Vision*, pages 138–142, 1994. doi: 10.1109/ACV.1994.341300.

C. Sanchez-Hernandez, D. Boyd, and G. Foody. One-class classification for mapping a specific land-cover class: Svdd classification of fenland. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4): 1061–1073, 2007.

L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, July 2013.

B. Schölkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002.

B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

B. Scholkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, 12:582–588, 2000.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

A. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.

C. Scott and G. Blanchard. Novelty detection: Unlabeled data definitely help. In *AISTATS*, 2009.

C. Scott and E. Kolaczyk. Annotated minimum volume sets for nonparametric anomaly discovery. In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pages 234–238. IEEE, 2007.

C. Scott and R. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7: 665–704, 2006.

M. Seeger. Learning with labeled and unlabeled data. Technical report, technical report, University of Edinburgh, 2001.

T. J. Sejnowski and G. Tesauro. The hebb rule for synaptic plasticity: algorithms and implementations. *Neural models of plasticity*, pages 94–103, 1989.

J. Seo and H. Ko. Face detection using support vector domain description in color images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages V–729. IEEE, 2004.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge Univ Pr, 2004.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

A. Singh. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003, 1989.

A. Singh, R. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems*, 2009.

A. Skabar. Single-class classifier learning using neural networks: An application to the prediction of mineral deposits. In *International Conference on Machine Learning and Cybernetics*, volume 4, pages 2127–2132. IEEE, 2003.

P. Smagghe, J.-L. Buessler, and J.-P. Urban. Novelty detection in image recognition using irf neural networks properties. In *European symposium on artifcial neural networks, computational intelligence and machine learning (ESANN), Bruges*, 2013.

B. Sofman, J. Bagnell, and A. Stentz. Anytime online novelty detection for vehicle safeguarding. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1247–1254, 2010. doi: 10.1109/ROBOT.2010.5509357.

A. Soltani-Farani, H. R. Rabiee, and S. A. Hosseini. Spatial-aware dictionary learning for hyperspectral image classification. *arXiv preprint*, 2013.

M. Soltanolkotabi and E. J. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

C. Song, C. E. Woodcock, K. C. Seto, M. P. Lenney, and S. A. Macomber. Classification and change detection using landsat tm data: when and how to correct atmospheric effects? *Remote sensing of environment*, 75(2):230–244, 2001.

R. Souvenir and R. Pless. Manifold clustering. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 648–653. IEEE, 2005.

J. Stamenkovic, D. Tuia, F. de Morsier, M. Borgeaud, and J.-P. Thiran. Estimation of soil moisture from airborne hyperspectral imagery with support vector regression. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2013.

I. Steinwart, D. Hush, and C. Scovel. Density level detection is classification. In *Advances in Neural Information Processing Systems*, pages 1337–1344, 2004.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 2013.

X. Sun, Q. Qu, N. Nasrabadi, and T. Tran. Structured priors for sparse-representation-based hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, PP(99):1–5, 2013. ISSN 1545-598X. doi: 10.1109/LGRS.2013.2290531.

I. Sundaram, I. G. Strachan, D. A. Clifton, L. Tarassenko, and S. King. Aircraft engine health monitoring using density modeling and extreme value statistics. In *Proc. 6th Int. Conf. on Condition Monitoring and Machinery Failure Prevention Technologies, Dublin, Ireland*, pages 919–930, 2009.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Artificial Neural Networks, 1995., Fourth International Conference on*, pages 442–447, 1995.

L. Tarassenko, A. Nairac, N. Townsend, and P. Cowley. Novelty detection in jet engines. In *IEE Colloquium on Condition Monitoring: Machinery, External Structures and Health (Ref. No. 1999/034)*, page 4, 1999.

D. Tax. *One-class classification*. PhD thesis, Delft University of Technology, 2001. URL http://ict.ewi.tudelft.nl/~davidt/thesis.pdf.

D. Tax and R. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999.

D. Tax and R. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

J. B. Tenenbaum. Rules and similarity in concept learning. *Advances in Neural Information Processing Systems*, 12:59–65, 2000.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

J. Theiler. Quantitative comparison of quadratic covariance-based anomalous change detectors. *Applied Optics*, 47(28):F12–F26, 2008.

J. Theiler and B. Wohlberg. Local coregistration adjustment for anomalous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3107–3116, 2012.

J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy. Elliptically contoured distributions for anomalous change detection in hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 7(2):271–275, 2010.

D. R. Thompson, W. A. Majid, C. J. Reed, and K. L. Wagstaff. Semi-supervised eigenbasis novelty detection. *Statistical Analysis and Data Mining*, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

H. Tiitinen, P. May, K. Reinikainen, and R. Näätänen. Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature*, 372:90–92, 1994.

A. Tikhonov. *Numerical methods for the solution of ill-posed problems*. 1995.

D. Tuia and G. Camps-Valls. Semi-supervised remote sensing image classification with cluster kernels. *IEEE Geoscience and Remote Sensing Letters*, 6(2):224–228, 2009.

D. Tuia and G. Camps-Valls. Urban image classification with semisupervised multiscale cluster kernels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1):65–74, 2011.

D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.

D. Tuia, J. Munoz-Mari, L. Gomez-Chova, and J. Malo. Graph matching for adaptation in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):329–341, 2013.

M. Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16 (6):22–38, 1999.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

J.-P. Vert, K. Tsuda, and B. Schölkopf. A primer on kernel methods. *Kernel Methods in Computational Biology*, pages 35–70, 2004.

R. Vert and J. Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. ISSN 1053-5888. doi: 10.1109/MSP.2010.939739.

R. Vidal and P. Favaro. Low rank subspace clustering (lrsc). *Pattern Recognition Letters*, 2013. ISSN 0167-8655. doi: 10.1016/j.patrec.2013.08.006.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.244.

A. Villa, J. Chanussot, J. Benediktsson, C. Jutten, and R. Dambreville. Unsupervised methods for the classification of hyperspectral images with low spatial resolution. *Pattern Recognition*, 46(6):1556 – 1568, 2013. ISSN 0031-3203. doi: 10.1016/j.patcog.2012.10.030.

M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection with kernels. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1026–1030, 2012. ISSN 1545-598X. doi: 10.1109/LGRS.2012.2189092.

M. Volpi, F. de Morsier, G. Camps-Valls, M. Kanevski, and D. Tuia. Multi-sensor change detection based on nonlinear canonical correlations. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2013.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

E.-S. Wacker and J. Denzler. Enhanced anomaly detection in wire ropes by combining structure and appearance. *Pattern Recognition Letters*, 2013.

D. Wang, D. Yeung, and E. C. C. Tsang. Structured one-class classification. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 36(6):1283–1295, 2006. ISSN 1083-4419. doi: 10.1109/TSMCB.2006.876189.

G. Wang, D. Yeung, and F. Lochovsky. A kernel path algorithm for support vector machines. In *International Conference Machine Learning (ICML)*, pages 951–958. ACM, 2007.

J. Wang and S. Qin. A new subspace identification approach based on principal component analysis. *Journal of Process Control*, 12(8):841 – 855, 2002. ISSN 0959-1524. doi: 10.1016/S0959-1524(02)00016-1.

J.-S. Wang and J.-C. Chiang. A cluster validity measure with outlier detection for support vector clustering. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 38(1):78–89, 2008. ISSN 1083-4419. doi: 10.1109/TSMCB.2007.908862.

Q. Wang, L. S. Lopes, and D. M. Tax. Visual object recognition through one-class learning. In *Image Analysis and Recognition*, pages 463–470. Springer, 2004.

S. Wang, X. Yuan, T. Yao, S. Yan, and J. Shen. Efficient subspace segmentation via quadratic programming. In *AAAI*, 2011a.

X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *arXiv preprint*, abs/1201.5338, 2012a.

X. Wang, B. Qian, and I. Davidson. Labels vs. pairwise constraints: A unified view of label propagation and constrained spectral clustering. In *IEEE International Conference on Data Mining (ICDM)*, pages 1146–1151. IEEE, 2012b.

Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks*, 22(7):1149–1161, 2011b.

Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In *International Conference Machine Learning (ICML)*, pages 89–97, 2013.

Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When lrr meets ssc. In *Advances in Neural Information Processing Systems*, pages –, 2013a.

Z. Wang, N. Nasrabadi, and T. Huang. Spatial - spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–15, 2013b. ISSN 0196-2892. doi: 10.1109/TGRS.2013.2285049.

R. Willett, M. Duarte, M. Davenport, and R. Baraniuk. Sparsity and structure in hyperspectral imaging : Sensing, reconstruction, and target detection. *IEEE Signal Processing Magazine*, 31(1):116–126, 2014. ISSN 1053-5888. doi: 10.1109/MSP.2013.2279507.

L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

C. Wu, B. Du, and L. Zhang. A subspace-based change detection method for hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2):815–830, 2013. ISSN 1939-1404. doi: 10.1109/JSTARS.2013.2241396.

L. Xiao, B. Dai, Y. Fang, and T. Wu. Kernel l1 graph for image analysis. In C.-L. Liu, C. Zhang, and L. Wang, editors, *Pattern Recognition*, volume 321 of *Communications in Computer and Information Science*, pages 447–454. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33505-1. doi: 10.1007/978-3-642-33506-8_55.

Y. Xiao, H. Wang, W. Xu, and J. Zhou. {L1} norm based {KPCA} for novelty detection. *Pattern Recognition*, 46(1):389 – 396, 2013. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2012.06.017.

E. P. Xing, M. I. Jordan, S. Russell, and A. Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002.

H. Xiong, M. Swamy, and M. O. Ahmad. Optimizing the kernel in the empirical feature space. *IEEE Transactions on Neural Networks*, 16(2):460–474, 2005.

H. Yang and M. Crawford. Manifold alignment for multitemporal hyperspectral image classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4332–4335, 2011. doi: 10.1109/IGARSS.2011.6050190.

L. Yao, J. Tang, and J. Li. Entire solution path for support vector machine for positive and unlabeled classification. *Tsinghua Science & Technology*, 14(2):242–251, 2009.

Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Springer, 2003.

D. Yeung and C. Chow. Parzen-window network intrusion detectors. In *International Conference on Pattern Recognition (ICPR)*, volume 16, pages 385–388, 2002.

F. Yin and G. Huang. Improved density-induced support vector data description. In *International Conference Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 747–750. IEEE.

N. Yousri, M. Kamel, and M. Ismail. A novel validity measure for clusters of arbitrary shapes and densities. In *International Conference on Pattern Recognition (ICPR)*, pages 1–4, dec. 2008. doi: 10.1109/ICPR.2008.4761242.

A. Ypma and R. P. Duin. Support objects for domain approximation. Citeseer, 1998.

H. Yu. Single-class classification with mapping convergence. *Machine Learning*, 61(1-3):49–69, 2005.

S. Yu and J. Shi. Multiclass spectral clustering. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 313–319, 2003. doi: 10.1109/ICCV.2003.1238361.

S. Zafeiriou and M. Petrou. Nonlinear non-negative component analysis algorithms. *IEEE Transactions on Image Processing*, 19(4):1050–1066, 2010.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2004.

D. Zhang, Z.-H. Zhou, and S. Chen. Non-negative matrix factorization on kernels. In Q. Yang and G. Webb, editors, *PRICAI 2006: Trends in Artificial Intelligence*, volume 4099 of *Lecture Notes in Computer Science*, pages 404–412. Springer Berlin Heidelberg, 2006.

D. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision (ICCV)*, pages 471–478, 2011. doi: 10.1109/ICCV.2011.6126277.

L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li. Kernel sparse representation-based classifier. *IEEE Transactions on Signal Processing*, 60(4):1684–1695, 2012. ISSN 1053-587X. doi: 10.1109/TSP.2011.2179539.

T. Zhang, J. Yang, D. Zhao, and X. Ge. Linear local tangent space alignment and application to face recognition. *Neurocomputing*, 70(7):1547–1553, 2007.

T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja. Low-rank sparse coding for image classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2013a.

T. Zhang, R. Ji, W. Liu, D. Tao, and G. Hua. Semi-supervised learning with manifold fitted graphs. In *International joint Conference on Artificial Intelligence (IJCAI)*, pages 1896–1902. AAAI Press, 2013b.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16(16):321–328, 2004.

J. Zhu, X. Fang, Z. Guo, M. H. Niu, F. Cao, S. Yue, and Q. Y. Liu. Ibm cloud computing powering a smarter planet. In *Cloud Computing*, pages 621–625. Springer, 2009.

X. Zhu and A. Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.

X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference Machine Learning (ICML)*, volume 3, pages 912–919, 2003.

# Frank de Morsier

28th of December 1985, Swiss nationality
CH +41 792065281
frank.demorsier@epfl.ch

Av. Général Guisan 76
CH-1009 Pully
Switzerland



**SUMMARY** *Strong practical and theoretical knowledge in **machine learning, data mining** and **pattern recognition** for **Signal/Image Processing** with a Minor degree in **Space Technologies** giving a multidisciplinary approach. Diligent and involved in team projects with broadened perspectives from international studies.10+ papers in international journal and conferences and 2 patents.*


## EDUCATION

10/2009 – 04/2014 **Ph.D student** – Signal Processing Lab (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Supervisor: Prof. Jean-Philippe Thiran (LTS5), Co-supervisor: Dr. Maurice Borgeaud (ESA Earth Observation Science, Applications, and Future Technologies department) Subject: Semi-supervised and unsupervised kernel-based novelty detection with applications to remote sensing imagery

09/2007 – 07/2009 **Master in Information Technology** - (GPA 5.47 / 6) EPFL, Switzerland

**Minor in Space Technologies** - Space Center, EPFL, Switzerland

10/2004 – 07/2007 **Bachelor in Electrical and Electronics Eng.** - (GPA 5.03 / 6) EPFL, Switzerland

09/2006 – 07/2007 **Heriot-Watt University Edinburgh** - Exchange year, United Kingdom

09/2001 – 07/2004 Baccalaureate Certificate - Option physics and maths, Gymnase de Morges, Switzerland

## PROFESSIONAL EXPERIENCES

02/2012 – 08/2012 (6 months) Internship – **IBM** Zurich **Research Laboratory** (ZRL), Switzerland
- Fast novelty detection in VLSI layouts via dictionary learning and compressed sensing
- 1 patent filed and 1 pending.

07/2008 (4 weeks) Internship - **Applied Signal Processing Group (ASPG)** EPFL, Switzerland
- Biomedical Signal Processing on Atrial Fibrillation
- 3D Dipole reconstruction, clustering and dynamic analysis (chaotic characterization)

07/2007 – 08/2007 (6 weeks) Internship - **Flexcell / VHF Technologies** Yverdon, Switzerland
- Specifications of a buffer battery between flexible solar panels and laptops

## PROJECTS

Course Project **Graphical Models**: Variational EM with Markov Random Field for image segmentation

Master Project **Image Processing**: Registration of multi-modal and multi-temporal remote sensing images
- *Robust registration based on phase-correlation and gradient image representation*

Semester Project (Minor) **Remote Sensing**: Automatic download of MERIS images from ESA Rolling Archives
- *PERL algorithm localizing and downloading satellite images in near-real time*

Semester Project (Master) **Biomedical Signal Processing**: Prediction of Atrial Fibrillation, adaptive filtering & classification
- *Co-author for an oral presentation at **« Computers in Cardiology 08 - Bologna »**, Italy*

Bachelor Project **Digital Signal Processing**, Robotics & Electronics: Battery sorter (AAA, AA & C)
- FPGA AHDL coding, C++ interface, USB & PIC control, *Business plan*

## RELATED COURSES

PhD courses Machine Learning (Prof. A. Billard, Dr. F. Fleuret), Probabilistic Graphical Models (Prof. V. Cevher), Advanced Image Processing (Profs. M. Unser & J.-P. Thiran)

Master courses Image Analysis and Pattern Recognition, Advanced Signal Processing, Speech Processing, Biomedical Signal Processing, Aerial mapping, Satellite communications, Principles of finance

**COMPUTER SKILLS** Coding: Matlab/Simulink, C/C++, PERL, Python, Java, PHP/mySQL, AHDL (VHDL)
Engineering: ENVI/IDL, ERDAS Imagine, QGIS, LabVIEW, Solidworks

**SUMMER SCHOOLS**

(2 days) 07/2012 **Remote Sensing Summer School** - Munich, Germany (co-located with IGARSS)
(2 weeks) 04/2012 **Machine Learning Summer School** - La Palma, Canaria islands (co-located with AISTATS)
(2 weeks) 07/2010 **Alpbach Summer School** - Design new space mission for the understanding of climate change

- **Award: Jury's chairman (**Roger Bonnet, ISSI**) price**
- *Limited to 60 European engineering and science students*

**TEACHING EXPERIENCES**

(120hours) 2012-13 Teaching Assistant - Master course « Earth Observation »
(20 hours) 2011 Teaching Assistant - Master course « Image Processing »
(60 hours) 2009 Teaching Assistant - Master course « Remote Sensing of the Earth by Satellites »
(60 hours) 2008 Teaching Assistant - Minor course « Spacecraft data processing & Interfaces »

**ADDITIONAL EXPERIENCES**

09/2012 – 04/2014 Comity member of the EDEE students association (doctoral school association)

09/2008 – 07/2012 Comity member and webmaster - EPFL Sailing club (voile.epfl.ch), Switzerland

Summ. 2003 - 2005 Sailing instructor - CNM Morges, Switzerland

**LANGUAGES** French: Mother tongue, English: Fluent – C1 Level*, German: High school – A2 Level*

*Levels defined by the « Common European Framework of Reference », see http://www.coe.int/T/DG4/Portfolio/

**OTHERS INTERESTS** Sailing competition, Guitar player, Audio/Visual creations, Snowboard