UNIVERSITÀ DI PISA

Engineering PhD School "Leonardo Da Vinci"

PhD Course: *REMOTE SENSING*

PhD Thesis

Models and Methods

for Automated Background Density Estimation

in Hyperspectral Anomaly Detection

PhD Candidate

Tiziana Veracini

Advisors

Prof. Giovanni Corsini

Prof. Marco Diani

PhD Course year: 2011
Scientific disciplinary field: SSD ING-INF/03

Abstract

Detecting targets with unknown spectral signatures in hyperspectral imagery has been proven to be a topic of great interest in several applications. Because no knowledge about the targets of interest is assumed, this task is performed by searching the image for anomalous pixels, i.e. those pixels deviating from a statistical model of the background. According to the hyperspectral literature, there are two main approaches to Anomaly Detection (AD) thus leading to the definition of different ways for background modeling: *global* and *local*. Global AD algorithms are designed to locate small *rare* objects that are anomalous with respect to the *global* background, identified by a large portion of the image. On the other hand, in local AD strategies, pixels with significantly different spectral features from a local neighborhood just surrounding the observed pixel are detected as anomalies.

In this thesis work, a new scheme is proposed for detecting both global and local anomalies. Specifically, a simplified Likelihood Ratio Test (LRT) decision strategy is derived that involves thresholding the background log-likelihood and, thus, only needs the specification of the background Probability Density Function (PDF). Within this framework, the use of parametric, semi-parametric (in particular finite mixtures), and non-parametric models is investigated for the background PDF estimation. Although such approaches are well known and have been widely employed in multivariate data analysis, they have been seldom applied to estimate the hyperspectral background PDF, mostly due to the difficulty of reliably learning the model parameters without the need of operator intervention, which is highly desirable in practical AD tasks. In fact, this work represents the first attempt to jointly examine such methods in order to asses and discuss the most critical issues related to their employment for PDF estimation of hyperspectral background with specific reference to the detection of anomalous objects in a scene.

Specifically, semi- and non-parametric estimators have been successfully employed to estimate the image background PDF with the aim of detecting global anomalies in a scene by means of the use of *ad hoc* learning procedures. In particular, strategies developed within a Bayesian framework have been considered for automatically estimating the parameters of mixture models and one of the most well-known non-parametric techniques,

i.e. the fixed kernel density estimator (FKDE). In this latter, the performance and the modeling ability depend on scale parameters, called bandwidths. It has been shown that the use of bandwidths that are fixed across the entire feature space, as done in the FKDE, is not effective when the sample data exhibit different local peculiarities across the entire data domain, which generally occurs in practical applications. Therefore, some possibilities are investigated to improve the image background PDF estimation of FKDE by allowing the bandwidths to vary over the estimation domain, thus adapting the amount of smoothing to the local density of the data so as to more reliably and accurately follow the background data structure of hyperspectral images of a scene.

The use of such variable bandwidth kernel density estimators (VKDE) is also proposed for estimating the background PDF within the considered AD scheme for detecting local anomalies. Such a choice is done with the aim to cope with the problem of non-Gaussian background for improving classical local AD algorithms involving parametric and non-parametric background models. The locally data-adaptive non-parametric model has been chosen since it encompasses the potential, typical of non-parametric PDF estimators, in modeling data regardless of specific distributional assumption together with the benefits deriving from the employment of bandwidths that vary across the data domain.

The ability of the proposed AD scheme resulting from the application of different background PDF models and learning methods is experimentally evaluated by employing real hyperspectral images containing objects that are anomalous with respect to the background.

Index Terms - anomaly detection, hyperspectral images, finite mixture model, kernel density estimation, Bayesian learning, variable bandwidth kernel density estimation.

Acknowledgement

It would not have been possible to write this PhD dissertation without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

Above all, I would like to express my deep and sincere gratitude to my supervisors, Prof. Giovanni Corsini and Prof. Marco Diani, of "Università di Pisa". They provided me with many helpful suggestions, important advice and constant encouragement during the course of this work.

All the people of the Remote Sensing & Image Processing Group, of "Università di Pisa", are also acknowledged for their support and help. Explicit thanks go to Stefania Matteoli for the useful discussions and her assistance in writing reports and papers.

I am most grateful to my family for the support they have provided me through my entire life. My special appreciation goes to my parents, Liliana and Emilio, who always kept me away from family responsibilities and encouraged me to concentrate on my study. I also want to thank my brother Alessandro for helping me with some figures of this dissertation without complaints even when he was on peaks of stress and lack of sleep because of his job. To them I dedicated this thesis work.

Last, but by no means least, I would like to express special thanks to Luca. Without him, I would be a very different person today, and it would have been certainly much harder to finish a PhD. My keen appreciation goes to him for helping me to concentrate on completing this dissertation and supported mentally during the course of this work.

Thank you.

Ringraziamenti

È difficile citare e ringraziare in poche righe tutte le persone che, a vario titolo, hanno contribuito alla nascita e allo sviluppo di questo lavoro. Ciò nonostante, di seguito desidero ricordare brevemente le persone che più mi hanno sostenuto in questo percorso di studi.

Innanzitutto, vorrei esprimere la mia profonda e sincera riconoscenza ai miei Tutori, Prof. Giovanni Corsini e Prof. Marco Diani, dell'Università di Pisa, per avermi saputo indirizzare, consigliare e sostenere nella ricerca. Le nostre conversazioni sono sempre state fonte di suggerimenti utili e consigli importanti.

Desidero inoltre ricordare tutti i ragazzi del Laboratorio di Elaborazione delle Immagini e del Segnale, del Dipartimento di Ingegneria dell'Informazione dell'Università di Pisa, per il loro supporto e aiuto. Un ringraziamento speciale lo dedico a Stefania Matteoli per la sua continua disponibilità e per il suo prezioso supporto.

Dedico questa tesi alla mia Famiglia, che ha sempre avuto un ruolo insostituibile nella mia vita. L'amore e la comprensione dei miei genitori, Emilio e Liliana, sono stati essenziali in questo percorso di studi. Li ringrazio soprattutto per avermi sempre lasciato libera di decidere cosa fare, anche quando ciò comportava dei sacrifici notevoli per me e per loro. Voglio anche ringraziare mio fratello Alessandro per avermi aiutato con alcune delle figure presenti in questa tesi senza lamentarsi, anche nei momenti di nervosismo causati dal suo lavoro.

Infine, ma non per questo meno importante, vorrei esprimere un ringraziamento speciale a Luca, il quale ha saputo aiutarmi e sostenermi durante lo svolgimento di questo lavoro. Senza di lui sarei una persona molto diversa oggi, e sicuramente sarebbe stato molto più difficile finire questo percorso verso il dottorato di ricerca.

Grazie.

Contents

List of Tables

List of Figures

List of Acronyms

AD	*Anomaly Detection*
AIC	*Akaike Information Criterion*
A-RX	*adaptive GKNNE-based AD algorithm*
BGMMS	*Bayesian GMM Split*
BIC	*Bayesian Information Criterion*
BN	*Bors-Nasios*
BStMM	*Bayesian StMM*
EC	*Elliptically Contoured*
EM	*Expectation Maximization*
FKDE	*Fixed Kernel Density Estimator*
FMM	*Finite Mixture Model*
FOV	*Field Of View*
GIC	*Generalized Information Criterion*
GMM	*Gaussian Mixture Model*
iid	*independent and identically distributed*
ISE	*Integrated Squared Error*
K-RX	*kernel RX algorithm*
LNM	*Local Normal Model*
LRT	*Likelihood Ratio Test*
MISE	*Mean Integrated Squared Error*
ML	*Maximum Likelihood*
NP	*Neyman-Pearson criterion*
PDF	*Probability Density Function*
RX	*Reed-Xiaoli algorithm*
SIM-GA	*Sistema Iperspettrale Modulare – Galileo Avionica*
StMM	*Student's t Mixture Model*
VKDE	*Variable-bandwidth Kernel Density Estimator*

Models and Methods for Automated Background Density Estimation in Hyperspectral Anomaly Detection

Chapter 1

1 Introduction

1.1 Motivation and problem statement

Hyperspectral remote sensing is based on the fact that all materials show distinctive amount of reflected, absorbed, and emitted radiation at each wavelength that is related to their molecular composition. Hyperspectral sensors capture the spectra of the observed pixels in hundreds of contiguous and very narrow spectral bands (less than 0.010 μm wide). Accordingly, the resulting hyperspectral image includes both spatial features and very rich information content about the spectral characteristics of the observed materials that can be exploited to detect and identify objects in the image.

Several studies have demonstrated the usefulness of exploiting the information extracted from multiple spectral bands when searching for targets and objects in remotely sensed images [27][36][55]. Many works in this area have focused on the detection of targets with known spectral properties [10][21][24][36][37][49]. Within this framework, laboratory or field measurements of target spectra are typically assumed to be used as known spectral signatures to be detected within the remotely sensed image. In principle, such approaches search for those image pixels whose spectrum exhibits a high degree of correlation with the known target spectral signature. Nevertheless, precise knowledge of what type of paint or camouflage the target is equipped with may not be available in most cases, and the target detection task needs to be addressed by searching the image for those objects that are anomalous with respect to the scene, i.e. Anomaly Detection (AD) [27][42][55].

Spectral Anomaly Detection (AD) is a target detection problem in which no previous knowledge about the spectrum of the object of interest is available. Since no prior knowledge for the targets is assumed, the detection is based on the spectral separation between the anomalous objects and the background. In general, targets of interest can be further divided into *global* and *local* anomalies thus leading to the definition of different background models [42][55]. Global AD algorithms aim at detecting small *rare* objects that are anomalous with respect to the rest of the image (i.e. the *global* background). In local AD strategies, on the other hand, the task of locating pixels with significantly different spectral features with respect to their *surrounding* background is taken into account.

In the literature, several AD algorithms have been developed on the basis of different approaches [27][42][55][75]. In this thesis work, the focus is on AD algorithms based on statistical methods. In this context, strategies are usually formulated by resorting to a binary hypothesis-testing problem solved according to decision rules typical of the detection theory. Within this framework, the foundation of many important AD approaches is the Neyman-Pearson criterion (NP), according to which the optimum decision strategy is given by a Likelihood Ratio Test (LRT) dependent on the probability density functions (PDFs) conditioned to the two hypotheses [30]. In this work, a simplified LRT decision rule is derived. Specifically, the proposed AD strategy involves thresholding the background log-likelihood and, thus, only needs the specification of the background PDF to detect spectral anomalies [27][42]. This AD scheme is able to accommodate the different definitions of anomaly. In particular, for global AD purposes, given the target rarity assumption, the whole scene is used to characterize the background. On the contrary, if a neighboring area around the pixels being tested is used to characterize the background, then the anomalies found are local. In both cases, the background PDF is unknown and has to be estimated from the data. This is general accomplished by assuming a model for the PDF to estimate. Thus, different background models lead to different AD algorithms.

The estimation of PDFs based on representative data samples drawn from the underlying density is a problem of fundamental importance in various fields, such as machine learning, pattern recognition, and computer vision [12][46][59]. Therefore, several different PDF models have been proposed in the literature [6][23][54]. The simplest approach to PDF estimation is *parametric* estimation, which assumes data drawn from a specific parametric unimodal distribution (e.g., the Gaussian one) [3]. The

advantage of the parametric approach is that the model is defined by a small number of parameters. Once those parameters are estimated from the sample data and the estimates are plugged in the assumed model, the whole distribution is obtained. However, while this assumption may reflect reality only in certain situations, it is more generally the case that the image background contains several different types of land-covers. Therefore, multi-modal distributions are more appropriate to capture the complexity of the image background [55]. To this aim, *semi-parametric* approaches, such as the widely employed Finite Mixture Models (FMMs), may provide a more accurate background characterization [6][55]. They model the background PDF as a linear combination of PDFs of the same kind, thus accommodating the multimodality. Semi-parametric estimation procedure entails that the type and number of mixture components are properly chosen to do not compromise the estimator performance. Alternatively, a *non-parametric* PDF estimator can be considered, in which the data are not assumed to be drawn from a specific distribution nor has the PDF to follow a specific statistical model [23][29][50][54]. One of the most well-known techniques for non-parametric PDF estimation is the Fixed Kernel Density Estimator (FKDE), also known as Parzen estimator [23][29][54]. FKDE applies smooth functions (i.e., the kernel functions) at each data sample, and, then, the PDF estimate at a test sample is computed by averaging the values assumed by the kernel functions in correspondence of the given test sample. The performance and the modeling ability of FKDE depend on scale parameters, called bandwidths, that control the degree of smoothing of the resulting estimate [54]. Basically, the bandwidths are the kernel function widths. In the case of homogeneous data statistics, global bandwidths suffice for the analysis. However, the use of fixed bandwidths over the entire feature space, as done in the FKDE, has been shown to be not effective when the data samples exhibit different local peculiarities across the entire data domain [54]. In such cases, the employment of a variable-bandwidth KDE (VKDE) to adapt the amount of smoothing to the local density of data sample in the feature space, so as to more reliably and accurately follow the multivariate background data structure, has been suggested [29][50][54].

Although such PDF estimators are well-known and have been widely applied to different aspects of low-dimensional data analysis, their use in the hyperspectral AD context has been limited, mostly because of the difficulty in learning the underlying models and parameters in a reliable and automatic data-driven fashion. The main idea that inspired this thesis is to jointly investigate different PDF estimators within a common AD

framework. Specifically, this research work refers to the previously mentioned AD scheme in which the well-recognized statistical framework of the LRT decision rule is combined with reliable and automatic data-driven background PDF estimation. Within this detection strategy, several different background models are investigated for both global and local AD purposes. To this aim, different learning methods, each tailored to the specific statistical model considered, are investigated in their application to the background PDF estimation.

Specifically, the analysis of global AD is focused on semi-parametric approach, in particular mixture models, and non-parametric approaches.

In the semi-parametric approach, PDF estimation is carried out through a *model learning* procedure aimed at estimating the parameters required to characterize the Finite Mixture Model (FMM). Model parameter learning allows the PDF estimate to be fully specified. Typically, FMM learning has been conducted within the well-known Expectation Maximization (EM) framework. Nevertheless, the EM algorithm may be impaired by several limitations, such as incurring in singular solutions and the inability to automatically solve the FMM model-order selection. As regards the FKDE non-parametric estimator, the model is entirely learned from the data without resorting to parameter estimation. However, the employment of kernel functions to interpolate the data requires the kernel smoothing degree to be specified in advance. This means that a suitable bandwidth matrix has to be selected. The selection of reliable bandwidths has always been regarded as a major problem in the KDE literature. In this work, the parameter learning of FMMs and FKDE is carried out within a Bayesian framework [4][7][12]. In such a way, limitations inherent to EM for FMM learning are overcome as well as an automatic learning of the bandwidths in FKDE is made possible. In principle, the Bayesian approach involves prior knowledge within the model so that relevant properties of the data generation mechanism can be properly modeled and handled. In particular, by posing the model-learning problem in probabilistic terms through the Bayesian approach, the parameters are learnt in an automatic fashion, thus making the whole AD scheme applicable without the need of operator intervention. Though applied to several computational intelligence applications (such as image segmentation and blind source separation), such Bayesian methods have seldom been investigated as concern their capability of automatically learn the background PDF model parameters in hyperspectral images. In fact, this work represents the first attempt to jointly examine such methods in

order to investigate the issues related to Bayesian-based semi- and non-parametric PDF estimation of hyperspectral image background with specific reference to the detection of anomalous objects in a scene.

Among the considered PDF estimators, the non-parametric approach has been shown to be the most attractive approach to be applied in practical AD tasks. This is mostly due to its independence of specific background distributional assumptions. So, the possibility for improving the FKDE outcomes by varying the bandwidth over the domain of estimation as to hyperspectral image PDF estimation is also explored. In fact, although VKDE has already been employed in some pattern recognition applications [11][54], its potential as regards the background PDF estimation for enabling detection of anomalies in hyperspectral images has not been investigated yet and represents a topic of great interest for the remote sensing and target detection communities. Within this framework, the nearest neighbor class of estimators has been shown to represent a valuable attempt to adapt the amount of smoothing to the local density of data [35][54].

As to local AD strategies, parametric and non-parametric approaches to background PDF estimation are compared. The parametric methods are analyzed by employing the Reed-Xiaoli (RX) algorithm [48], which is considered to be the *benchmark* AD algorithm for multi/hyperspectral images. Within this framework, the data in the two hypotheses are assumed to arise from normal distributions with the same covariance matrix but different mean vectors. Such a Local Normal Model (LNM) is generally forced onto the image by performing a local demeaning using a sliding window. Then, according to the proposed simplified LRT, the decision rule for the RX algorithm can be derived. Nevertheless, most real-world data do not fit the LNM, especially in complex background situations. Starting from this, several AD strategies trying to cope with the problem of non-Gaussian background have been presented [32][36][42]. In this work, in order to benefit from the great potential that non-parametric methods embed, i.e. the ability at modeling complex local backgrounds without making specific distributional assumptions, a novel local AD strategy is proposed. Specifically, the strategy relies upon the proposed LRT decision rule and involves a VKDE to model the local background.

Therefore, in this thesis work the use of different PDF estimators and model learning procedures is jointly explored within a common AD scheme for detecting anomalies by means of the background log-likelihood decision rule. Within this framework, attention will not be devoted only to the capability of detecting the anomalous objects in a scene.

Rather, the analysis of the methodologies will be focused on their ability at providing good AD performance coupled with a reliable estimation of the image PDF. Indeed, this latter encloses a rich information content that can be useful for many unsupervised image analysis tasks and may provide ancillary information about the scene containing the detected anomalies. Two real hyperspectral images encompassing different AD scenarios are employed to evaluate and discuss their most critical issues, such as their modeling ability as well as their actual utility in practical AD tasks, and to evaluate, by means of several different performance measures, experimental detection performance.

1.2 Outline of the thesis

The proposed approach to AD in hyperspectral imaging is dealt with in chapter 2. Brief insights into the physics behind hyperspectral signal acquisition and into hyperspectral image structure is given. Detailed mathematical derivation of the proposed AD strategy is then provided, followed by a rigorous description of how it is used as a detector of global and local anomalies.

The problem of the background PDF estimation is dealt with in chapter 3. In particular, parametric, semi-parametric and non-parametric modeling approaches are reviewed.

Although such models are effective from a theoretical perspective and are used in a variety of multivariate signal processing problems, the difficulty in learning the underlying models both reliably and automatically has made their application in the hyperspectral AD context very limited. Solutions to face these limitations are proposed. The aim is to make the background modeling and estimation procedures robust and automatic. In particular, chapter 4 is dealt with methodologies to cope with global background modeling. To this aim, the use of mixture models and non-parametric PDF estimators are considered. Particular emphasis is placed on parameter selection carried out within a Bayesian framework for mixture models and FKDE. Besides, the *k*-nearest neighbor rule is proven to be an intuitively appealing procedure to adapt the bandwidth to the local density of data in the feature space for AKDEs.

In chapter 5, attention is initially focused on two representative local AD methods: RX and kernel RX. Then, a new approach trying to cope with the problem of non-Gaussian background is presented for improving classical local AD algorithms. Specifically, the use of a locally data-adaptive nonparametric model is proposed for estimating the background

PDF within an AD scheme for detecting anomalies by means of the background log-likelihood decision rule.

Experimental evidence of the actual advantages offered by the proposed solutions is obtained by employing real hyperspectral imagery in chapters 6 and 7.

Chapter 8 concludes the thesis outlining a summary, and providing the final remarks and conclusions.

Chapter 2

2 Hyperspectral anomaly detection methodology

In the past few years, hyperspectral data have been the subject of increasing interest for their very rich information content about the spectral characteristics of the materials in a scene. Such a unique ability to remotely extract features related to spectral content on a pixel-by-pixel basis has made Anomaly Detection (AD) a challenging area of research. In this chapter, the proposed AD strategy is illustrated. Since no previous knowledge about the targets of interest is assumed, the AD process is based on the anomalous nature of the pixels with respect to the statistics of the background samples.

2.1 The hyperspectral concept

The basic idea for hyperspectral imaging stems from the fact that, for any given material (e.g. vegetation pigments, minerals, rock, artificial surfaces), the amount of electromagnetic radiation that is reflected, absorbed, or emitted - i.e., the radiance - varies with wavelength in ways characteristic to its molecular composition. Hyperspectral sensors collect spectral radiance received by the observed scene in hundreds of narrow and contiguous spectral bands. In doing so, the optical system of the imaging sensor divides the imaged surface in pixels. The ground pixel size, which defines the image spatial resolution, is a function of the sensor and the platform altitude that, in turn, depend upon the kind of platform (e.g., space-borne or airborne). Every pixel of hyperspectral images provides an integrated measured spectrum of the materials contained on the ground area covered by the pixel. As a result, a hyperspectral image pixel is actually a column vector with dimension equal to the number of spectral bands.

The spatially and spectrally sampled information captured by hyperspectral sensors can be stored in a three-dimensional structure having two dimensions that represent the spatial coordinates and one associated to the spectral components. Such a data structure is commonly referred as *image cube*. Thus, the hyperspectral image can be seen as a stack of two-dimensional spatial images, each one associated to one of the sensor narrow bands. Alternatively, every pixel in the image can be seen as a discrete signal in the wavelength, with a so dense sampling that is potentially able to reveal even very small features peculiar of a certain material. An exemplification of the resulting structure of a hyperspectral image is illustrated in Fig. 2.1.

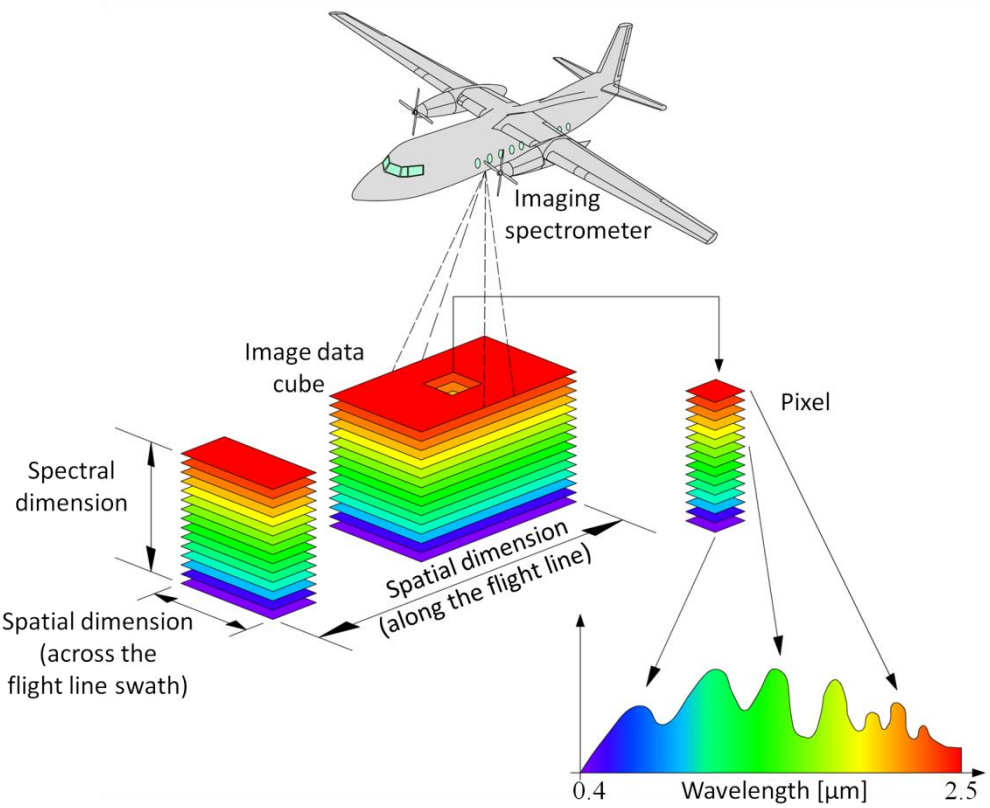


Fig. 2.1. Hyperspectral imaging sensors measure the spectral radiance information in a scene. This information is then processed to form a hyperspectral data set. The hyperspectral image data usually consist of over a hundred contiguous spectral bands, forming a three-dimensional (two spatial dimensions and one spectral dimension) image cube. Each pixel in this data set is associated with a very densely sampled spectrum of the imaged area, which can be exploited to identify the materials present in the pixel.

As is easily understandable, there is a large amount of information included in hyperspectral data, which can be used to detect and identify materials in the image. Such a discrimination capability has made AD an important and interesting area in data analysis.

2.2 The anomaly detection problem in hyperspectral imagery

The task of spectral AD algorithms is to explore the image for locating those image pixels whose spectral content significantly deviates from the background, without any previous knowledge about the targets of interest [42][55]. In principle, AD strategies derive a set of background characteristics and search for pixels that appear to be anomalous in comparison to these.

It is worth noting that there is not an unambiguous way to define an *anomaly*. This is mostly because the background can be identified in different ways. An important distinction is between *global* and *local* anomalies. If the whole image is used to characterize the background, then anomalies found are *global*. In such a way, the target class can be assumed to be scarcely populated, whereas the non-target class is made up by the majority of the image pixels and it encompasses the diverse kinds of background classes present in the examined scene. On the contrary, if the background is identified by a local neighborhood surrounding the observed pixel, the anomalies are *local*. Of course, the kind of anomalies one would like to detect depends on the particular application. Specifically, the local spectral anomaly detector is susceptible to isolated spectral anomalies. For example, consider a scene containing isolated trees on a grass plain. Isolated trees in a locally homogeneous patch of grass may be detected as local spectral anomalies even if the image contains a separate region with many pixels of trees. On the contrary, the global spectral anomaly detection algorithms will not find an isolated target in the open if the signature is similar to that of previously classified background material. In fact, in global AD algorithms, the task of detecting small *rare* objects that are anomalous with respect to the rest of the image is taken into account.

As a matter of fact, since ADs do not use any *a priori* knowledge, they cannot distinguish between legitimate anomalies and detections that are not of interest. Therefore, the detected anomalies may include man-made targets, natural objects, image artifacts, and other interferers. Clearly, a definitive identification of a target cannot be made through a search for anomalies. However, such a detection task can be extremely useful as a prompting device to guide the user in investigation of various kinds.

2.3 Anomaly detector design strategy

Given a test pixel **x**, the main goal of AD strategy is to decide whether a target of interest is present or not in the pixel under test, based on the different spectral characteristics of the pixels and the background. To this aim, AD is formulated as a binary hypothesis-testing problem where each pixel is labeled as anomalous or non-anomalous pixel:

$$\hat{H}(\mathbf{x}) = \begin{cases} H_0 : \ \mathbf{x} \text{ is a non} - \text{target pixel} \\ H_1 : \ \mathbf{x} \text{ is a target pixel} \end{cases} \tag{2.1}$$

where H_0 and H_1 denote the target absent (i.e., the background) and the target present hypothesis, respectively.

The most common approach to the hypothesis-testing problem is the Neyman-Pearson criterion (NP) [30]. Within this framework the *d*-dimensional (where *d* is the number of sensor spectral channels) random vector $\mathbf{X} = [X_1, X_2, \ldots, X_d]^t$ (the notation $(.)^t$ stands for vector transposed), associated to the multivariate pixel **x**, is modeled as:

$$\begin{aligned} H_0 : \quad & \mathbf{X} \sim f_{\mathbf{X}|H_0}(\mathbf{x}) \\ H_1 : \quad & \mathbf{X} \sim f_{\mathbf{X}|H_1}(\mathbf{x}) \end{aligned} \tag{2.2}$$

where $\left\{ f_{\mathbf{X}|H_i}(\mathbf{x}) \right\}_{i=0,1}$ denote the PDFs of **X** conditioned on the target absent (H_0) and target present (H_1) hypothesis, respectively. The NP decision rule has been derived by maximizing the detection probability $P_D = \Pr\left\{ \hat{H}_1 \mid H_1 \right\}$, with the constraint of maintaining a constant false alarm probability $P_{FA} = \Pr\left\{ \hat{H}_1 \mid H_0 \right\}$ at a desired value. According to NP criterion, the decision strategy is given by the Likelihood Ratio Test (LRT), which depends on the conditional PDFs under the two hypotheses, compared with a suitable threshold η:

$$LRT(\mathbf{x}) = \frac{f_{\mathbf{X}|H_1}(\mathbf{x})}{f_{\mathbf{X}|H_0}(\mathbf{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{2.3}$$

As is evident, in order to determine $LRT(\mathbf{x})$, both the conditional PDFs have to be known. However, in many situations of practical interest, there is lack of sufficient information to specify the statistical variability of the target signal, and a detector that exclusively uses background information is in demand.

It should be noted that given the LRT decision rule in (2.3), it is possible to suppose any form for the conditional PDFs. Let us assume $f_{\mathbf{X}|H_1}(\mathbf{x}) = f_{\mathbf{X}|H_0}(\mathbf{x}-\mathbf{s})$ [42]. This means that the following additive model for the two hypotheses is considered:

$$\begin{aligned} \mathbf{X} \mid H_0 &= \mathbf{B} \sim f_{\mathbf{X}|H_0}(\mathbf{x}) \\ \mathbf{X} \mid H_1 &= \mathbf{s} + \mathbf{B} \sim f_{\mathbf{X}|H_0}(\mathbf{x}-\mathbf{s}) \end{aligned} \tag{2.4}$$

where **s** represents the *d*-dimensional unknown deterministic vector associated with the target spectral signature, and **B** is the *d*-dimensional random vector representing the background (comprehensive of the noise). The maximum likelihood (ML) estimate of the deterministic unknown parameter **s** is given by

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \left\{ f_{\mathbf{X}|H_0}(x-\mathbf{s}) \right\} = \mathbf{x} - \arg\max_{\boldsymbol{\psi}} \left\{ f_{\mathbf{X}|H_0}(\boldsymbol{\psi}) \right\} = \mathbf{x} - \boldsymbol{\Psi} \tag{2.5}$$

Replacing **s** in (2.4) with its ML estimate expressed in (2.5), we obtain

$$\mathrm{LRT}(\mathbf{x}) = \frac{f_{\mathbf{X}|H_0}(\boldsymbol{\psi})}{f_{\mathbf{X}|H_0}(\mathbf{x})} \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \eta \tag{2.6}$$

From (2.6), a decision rule can be derived straightforwardly by noticing that the numerator is independent from **x**:

$$\Lambda(\mathbf{x}) = -\log\left\{ f_{\mathbf{X}|H_0}(\mathbf{x}) \right\} \underset{\mathrm{H_0}}{\overset{\mathrm{H_1}}{\gtrless}} \eta' \tag{2.7}$$

being η' the appropriate detection threshold.

The same approximation has been derived in [27] by assuming a uniform distribution of the conditional PDF under hypothesis H_1.

It has been shown that, under simplified assumptions for the two hypotheses, the AD task can be accomplished by thresholding the background log-likelihood. Within this framework, anomalies can be viewed as outliers because of the very different spectral features with respect to the background. Therefore, the corresponding pixels will contribute to the tail of the distribution making it heavier and allowing anomalies to be

detected by searching the deviation from the PDF of the background clutter samples. Since the background PDF $f_{\mathbf{X}|H_0}(\mathbf{x})$ is not known, it has to be estimated from the available data. Any different PDF estimator leads to a different detector. A brief summary of approaches to PDF estimation is given in the next chapter.

2.4 Operational application: the *local* and *global* anomaly detectors

Anomalies are defined with reference to a model of the background. Background models are developed using reference data from either a local neighborhood of the test pixel or the entire image, leading to *local* or *global* anomaly detectors, respectively.

2.4.1 Global anomaly detector

In global AD applications, the targets of interest are small and *rare* objects (i.e. extending over a few pixels and constituting a very small fraction of image) that are anomalous with respect to the rest of the image. In doing so, no previous knowledge is assumed about the nature of anomalies other than they are very sparsely and scarcely represented in the image. Hence, given the target rarity assumption, global AD algorithms are designed to identify small image regions corresponding to anomalies with respect to the global background.

In this thesis work, we refer to the decision rule specified by the equation (2.7), which only needs the specification of the image background PDF for global AD purposes. Therefore, the background PDF is obtained by estimating the image PDF on the basis of all image pixels available, which are indicated with $\{\mathbf{x}_n \in \Re^d | \ n = 1, 2, ..., N\}$. In fact, the heavy population of the background class, in conjunction with the sparseness of the target class, allows the "unclassified" image cube to be used for characterizing the background.

The processing chain of the global AD strategy here proposed is outlined in the graphical model in Fig. 2.2. Such an AD approach consists of two essential steps. First, the image PDF is estimated through one of the methodologies described in Chapter 3. It should be noted that the use of such estimators is coupled with the employment of automatic data-driven model learning methods illustrated in Chapter 4. Once the background PDF is approximated, equation (2.7) is applied to detect anomalous objects within the scene.

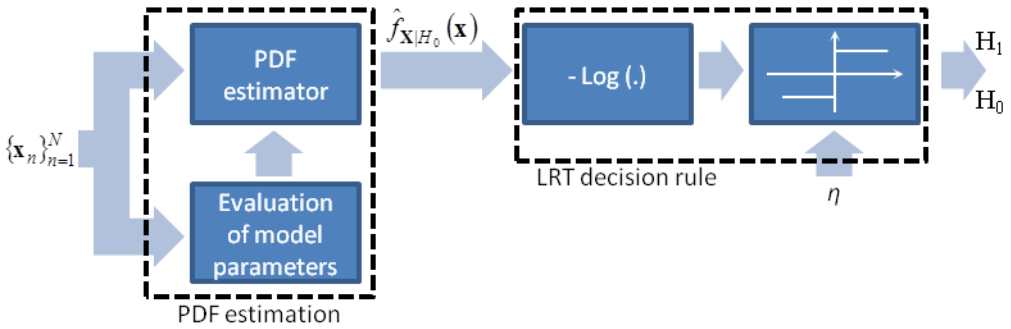


Fig. 2.2. Graphical model describing the stages of the proposed global AD strategy.

2.4.2 Local anomaly detector

Locally AD approaches aim at detecting the targets with respect to their *local* background, embodied by a neighborhood of surrounding pixels. Such *local* AD algorithms capture the local background pixels by sliding a spatial window over the image. In fact, for each test pixel onto which the window is centered, the pixels enclosed in the window are properly processed in order to compare their spectral properties with those of the test pixel. Within this framework, define the N reference background samples to be the pixels in the small surrounding area to the input data sample location employed for the local background characterization.

In order to prevent potential target pixels to affect the local background characterization, the algorithm is applied by sliding a dual concentric rectangular window (a small interior window centered within a larger outer one) over every pixel in the image. A graphical example of such a window is given in Fig. 2.3. The dual concentric windows divide the local area into the potential target region and the background region. The size of the interior window s_{iw} has to be chosen according to the maximum expected target dimension. This approximate size is based on previously knowledge about the Field Of View (FOV) of the hyperspectral sensor and the dimension of the biggest target in the given dataset. Instead, the size of the outer window s_{ow} is set to include sufficient data samples from the neighborhood of the pixels under test for the characterization of the local background. The resulting number of samples employed for the background PDF estimation is $N = s_{ow}^2 - s_{iw}^2$.

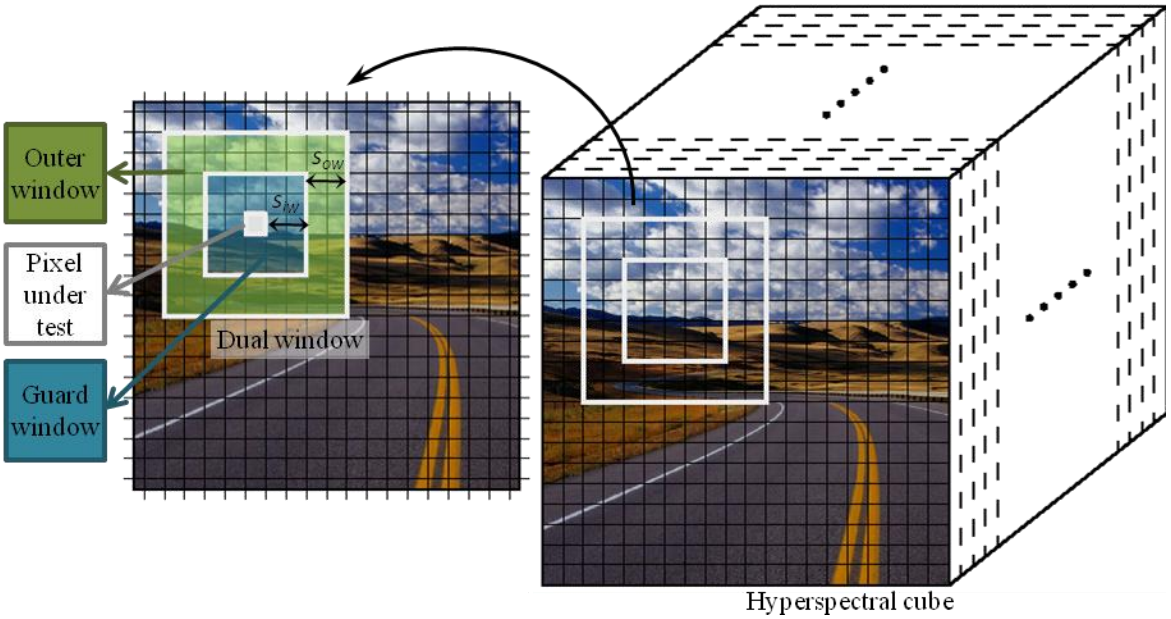


Fig. 2.3. Example of a dual concentric window in hyperspectral images. For 3×3 pixels expected target size, the inner window should be at least 5×5 pixels in order to not include target pixels in background PDF estimation windows.

It is important to note that the choice of the size of the surrounding area around the pixel under test concerning the background is not trivial. In fact, if it is too small could cause problems in computing the PDF estimate, while if it is too large essentially eliminates the locally adaptive nature of the detector [42].

Chapter 3

3 Statistical modeling approaches

This chapter focuses on methodologies for estimating a multivariate PDF. Hereinafter, $\hat{f}_{\mathbf{X}}(\mathbf{x})$ *will be used to denote the estimate of the multivariate PDF* $f_{\boldsymbol{X}}(\mathbf{x})$ *associated to the random vector* $\boldsymbol{X}$. $\hat{f}_{\mathbf{X}}(\mathbf{x})$ *is estimated on the basis of the available sample data, which are indicated with* $\{\mathbf{x}_n \in \Re^d | \, n = 1, 2, ..., N\}$.

3.1 Parametric PDF estimation

The simplest approach to estimate multivariate PDF is to compute it from an assumed parametric model [3]. Specifically, the parametric approach to PDF estimation assumes the data drawn from some specific unimodal distribution (e.g., the Gaussian one) governed by a small number of parameters $\boldsymbol{\theta}$ whose values are to be determined from the available data. Nevertheless, parametric models are very restricted in terms of forms of distribution that they can represent. For instance, if the process that generates the data is multimodal, then this aspect of the distribution can never be captured by a unimodal distribution.

Most AD algorithms in the literature assume that hyperspectral data are represented by the multivariate Gaussian distribution, mainly because of its mathematical tractability [36][43][55]. Typically, such a statistical model can be reliably employed only to characterize background pixels in a homogeneous local neighborhood around the pixel under test [41][42][55][48]. In practice, these assumptions are often violated. In fact, hyperspectral data generally do not closely follow the Gaussian distribution [38][39][40] and, in general, the choice of a rigid parametric model for the PDF to estimate is, indeed,

not appropriate for capturing the complexity of the data, especially for the assessment of the global background PDF [41][48].

3.2 Semi-parametric PDF estimation: finite mixture models

FMMs can approximate arbitrarily closely any continuous PDF provided the model has a sufficient number of components and appropriate model parameters [45]. Such an approach approximates the unknown PDF by a linear combination of J unimodal PDFs of the same kind, as follows:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \sum_{j=1}^{J} \pi_j \, g_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_j) \tag{3.1}$$

where $g(\mathbf{x};\boldsymbol{\theta}_j)$ denotes the multivariate PDF of $\mathbf{X}$ given the component distribution j controlled by the parameters vector $\boldsymbol{\theta}_j$, whereas $\{\pi_j\}_{j=1}^{J}$ are the mixing proportions (or weights). The parameters $\{\pi_j\}_{j=1}^{J}$ are subject to the constraints to be probabilities, i.e.

$$0 \leq \pi_j < 1 \tag{3.2}$$

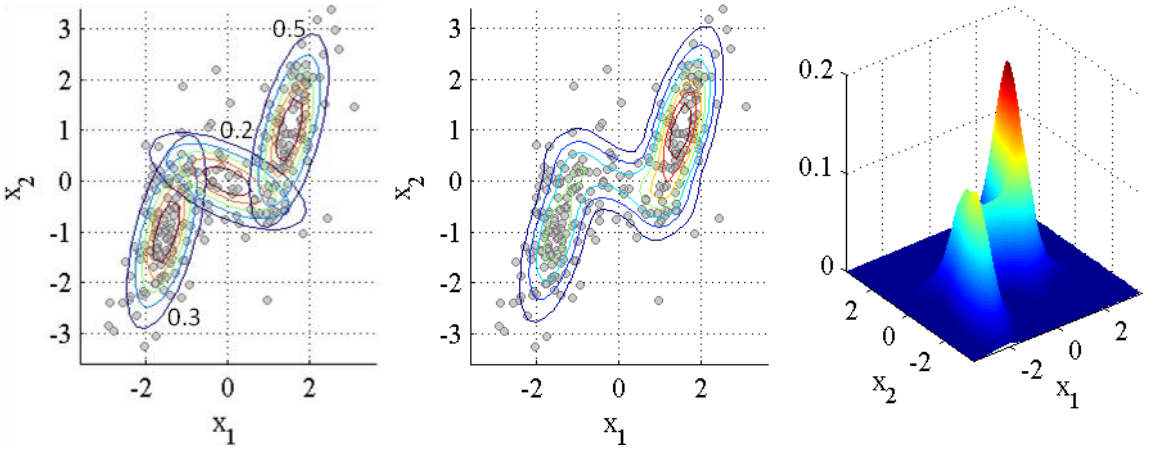
together with

$$\sum_{j=1}^{J} \pi_j = 1 \tag{3.3}$$

in order assure the PDF estimate $\hat{f}_{\mathbf{X}}(\mathbf{x})$ to be a legitimate PDF (non-negative and integrate to one). Fig. 3.1 illustrates using mixture models for PDF estimation. Specifically, contour and surface plots for a mixture model having three components are shown in such a figure.

As it clearly can be seen from the figure, mixture models provide a simple approach that can give rise to very complex densities.

As is evident the estimation procedure involves the choice of the density components $g_{\mathbf{x}}(\,)$ and, then, the estimation of the unknown parameters, $\boldsymbol{\theta}_j$ and π_j for j=1,…, J, based on the available sample data. In the current subsection, we focus on the unimodal PDF choice, whereas the parameters estimation will be addressed in Chapter 4.



(a) (b) (c)

Fig. 3.1. Illustration of using a mixture of three PDFs in a two-dimensional space. (a) Contour surfaces for each of the mixture components. The three components are denoted red, blue and green. The values of the mixing coefficients are indicated near each component. (b) Contours surfaces of the estimated PDF $\hat{f}_{\mathbf{X}}(\mathbf{x})$. (c) A surface plot of the estimated distribution $\hat{f}_{\mathbf{X}}(\mathbf{x})$.

3.2.1 Gaussian mixture model

From the multitude of distributions discussed in the statistics literature, the family of the Elliptically Contoured (EC) distributions has been shown to be suitable in mixture models to characterize hyperspectral data [38][39][40]. In general, the *d*-dimensional random vector **X** is EC distributed if its PDF can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{C}|^{1/2}} h_d(M) \tag{3.4}$$

where $h_d(.)$ is a positive, monotonically decreasing function of M for all d, whereas M corresponds to the square of the Mahalanobis distance, defined by:

$$M = (\mathbf{x}-\boldsymbol{\mu})^t \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{3.5}$$

in which $\boldsymbol{\mu}$ and $\mathbf{C}$ are the mean vector and the covariance matrix, respectively.

EC distributions have some important statistical properties such as [1][38][39]:

i. All EC distributions have elliptical isolevel curves.

ii. All the marginal and the conditional distributions of an EC distribution are also EC distributions.

The class of EC distributions includes the more familiar Gaussian distribution. In fact, the Gaussian is a special case of the EC family given by:

$$h_d(M) = \exp\left(-\frac{M^2}{2}\right) \tag{3.6}$$

The most widely employed FMM makes use of the Gaussian distribution. Such a FMM, which has been often adopted to model global heterogeneous backgrounds in hyperspectral images [9][55], is defined as Gaussian Mixture Model (GMM). Specifically, the GMM assumes $g_{\mathbf{X}}(\mathbf{x};\ \boldsymbol{\theta}_j)$ in (3.1) as a Gaussian distribution with parameters the mean vector $\boldsymbol{\mu}_j$ and the precision (inverse covariance) matrix $\mathbf{T}_j$ in the form:

$$g_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}_j) = g_N(\mathbf{x};\boldsymbol{\mu}_j, \mathbf{T}_j^{-1}) = \frac{\det(\mathbf{T}_j)}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^t\, \mathbf{T}(\mathbf{x}-\boldsymbol{\mu}_j)\right] \tag{3.7}$$

3.2.2 Student's t mixture model

Modeling a hyperspectral image through the GMM implicitly assumes that data from each background class in the image follow a multivariate Gaussian distribution. However, in many experimental studies performed with real hyperspectral images, such a model has been shown not to adequately represent the statistical behavior of the various background classes, which, instead, generally exhibit distributions characterized by heavier tails [38][39][40]. Experimental studies in [40] have suggested that the choice of an EC t-distribution, or Student's t PDF for $g_{\mathbf{X}}(\mathbf{x};\ \boldsymbol{\theta}_j)$, should provide, through equation (3.1), a reliable model for many hyperspectral data sets. In such a case, equation (3.1) develops into the Student's t Mixture Model (StMM) and $g_{\mathbf{X}}(\mathbf{x};\ \boldsymbol{\theta}_j)$ is defined by:

$$g_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}_j) = g_{St}(\mathbf{x};\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j, v_j) = \frac{\Gamma(v_j/2 + d/2)}{\Gamma(v_j/2)} \frac{|\boldsymbol{\Lambda}_j|^{1/2}}{(v_j\pi)^{d/2}} \exp\left[1 + \frac{(x-\boldsymbol{\mu}_j)^t\, \boldsymbol{\Lambda}_j (x-\boldsymbol{\mu}_j)}{v_j}\right]^{-\frac{v_j+d}{2}} \tag{3.8}$$

where $\Gamma(.)$ is the gamma function such that

$$\Gamma(t) = \int_0^\infty u^{t-1} e^{-u}\, du \tag{3.9}$$

whereas v_j>0 is the number of degrees of freedom, $\boldsymbol{\mu}_j$ is the mean vector, and $\boldsymbol{\Lambda}_j$ is the scale matrix, which is related to the covariance matrix $\mathbf{C}_j$ of $\mathbf{X}$ for v_j>2 by the following equation:

$$\mathbf{C} = \frac{v_j}{v_j - 2} \boldsymbol{\Lambda}_j^{-1}, \quad v_j > 2 \tag{3.10}$$

The integer v_j is the number of degrees of freedom, which controls the shape of the distribution tails: the smaller v_j is, the heavier the tails are. In particular, for v_j=1, the Student's t PDF reduces to the multivariate Cauchy distribution which has the heaviest tails, whereas when $v_j \to \infty$ it tends to the multivariate Gaussian distribution with mean $\boldsymbol{\mu}_j$ and precision matrix $\boldsymbol{\Lambda}_j$, characterized by lightest tails, as shown in Fig. 3.2.

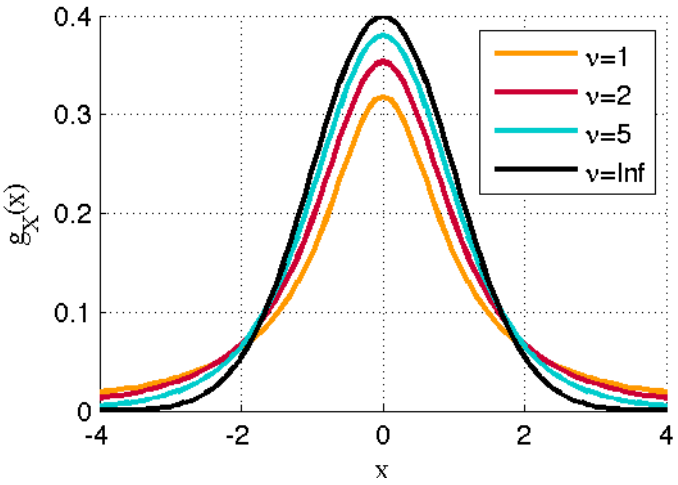


Fig. 3.2. One-dimensional representation of the Student's t probability density function for different values of the number of degrees of freedom v, which controls the shape of the distribution tails: the smaller v is, the heavier the tails are. In particular, for v=1, the Student's t PDF reduces to the multivariate Cauchy distribution. On the contrary, the Student's t PDF converges to the standard normal distribution as the degrees of freedom approaches infinity.

3.3 Non-parametric PDF estimation

Contrary to semi-parametric estimators, non-parametric PDF estimator does not assume any fixed functional form for the unknown PDF [23]. Basically, the unknown PDF is

entirely determined by the data through a kernel function $\kappa(.)$ centered at each different point of the sample data [57]:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\left|\mathbf{H}(\mathbf{x},\mathbf{x}_n)\right|}\kappa\left[\mathbf{H}(\mathbf{x},\mathbf{x}_n)\right] \tag{3.11}$$

In equation (3.11), $\mathbf{H}(.)$ is referred to as the bandwidth matrix, whereas $|.|$ indicates the matrix determinant. The bandwidth matrix $\mathbf{H}(.)$ is a $d \times d$ matrix that includes the bandwidths, i.e. the kernel widths.

The kernel function $\kappa(.)$ defines the shape of the influence region around each data sample location in the feature space and decreases in intensity with the distance from that location depending on the bandwidth values. Many possibilities exist for the kernel function choice in (3.11) [23][29][54]. Popular choices of multivariate kernel functions are radially symmetric unimodal PDFs such as the Gaussian and the Bartlett-Epanechnikov ones [29][23]. In certain situations, a product of univariate kernel functions $\kappa(\mathbf{u}) = \prod_{i=1}^{d}\kappa(u_i)$ may be appropriate. In this latter case, popular choices of the univariate kernel function are the Gaussian distribution and the rectangular and triangular functions [29][23]. In Table 1 and Table 2 the functional forms of common kernel functions are reported.

Table 1. Univariate kernel functions.

Kernel function	$\kappa(u)$ *
Rectangular	$\frac{1}{2}I(\lvert u\rvert \le 1)$
Triangle	$(1-\lvert u\rvert)I(\lvert u\rvert \le 1)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \le 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \le 1)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)I(\lvert u\rvert \le 1)$

*Where $I(\lvert u\rvert \le 1) = \begin{cases} 1 & \lvert u\rvert \le 1 \\ 0 & \textit{elsewhere} \end{cases}$

Table 2. Multivariate kernel functions.

Kernel function	$\kappa(\mathbf{u})$ *
Bartlett-Epanechnikov	$\frac{d+2}{2\pi^{d/2}}\left(1-\mathbf{u}^t\mathbf{u}\right)\Gamma\left(\frac{d}{2}+1\right)I\left(\lvert\mathbf{u}\rvert\le 1\right)$
Gaussian	$\frac{1}{(2\pi)^{-d/2}}e^{-\mathbf{u}^t\mathbf{u}/2}$

*Where $I\left(\lvert\mathbf{u}\rvert\le 1\right)=\begin{cases}1 & \lvert\mathbf{u}\rvert\le 1\\ 0 & elsewhere\end{cases}$

It has been widely recognized that the performance of non-parametric estimators does not depend on the kernel function choice but, rather, on the values of the bandwidths employed. Indeed, it has been shown that the bandwidth matrix **H**(.) in (3.11) influences the degree of smoothing for the resulting PDF approximation [54]. In general, the bandwidth function **H**(.) can be written as a function of both the estimation sample **x** and the observations from the unknown density $\{\mathbf{x}_n \in \Re^d | \ n = 1, 2, ..., N\}$. This form displays the possibility that kernel function shape may change in a large variety of ways. The approaches actually investigated are special cases of this more general bandwidth function **H**(.).

3.3.1 Fixed kernel density estimator

The FKDE is one of the most representative non-parametric techniques for PDF estimation. According to the FKDE [23], the estimation of $f_{\mathbf{X}}(\mathbf{x})$ is given by:

$$\hat{f}_{\mathbf{X}}(\mathbf{x})=\frac{1}{N\lvert\mathbf{H}\rvert}\sum_{n=1}^{N}\kappa\left[\mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}_n\right)\right] \tag{3.12}$$

where the bandwidth matrix **H** is independent of both observations and estimation samples, and, therefore, it has been held constant during the PDF estimation process. An illustration of the PDF estimation procedure is given in Fig. 3.3, where the individual kernel functions are shown as well as the estimate constructed by adding them up.

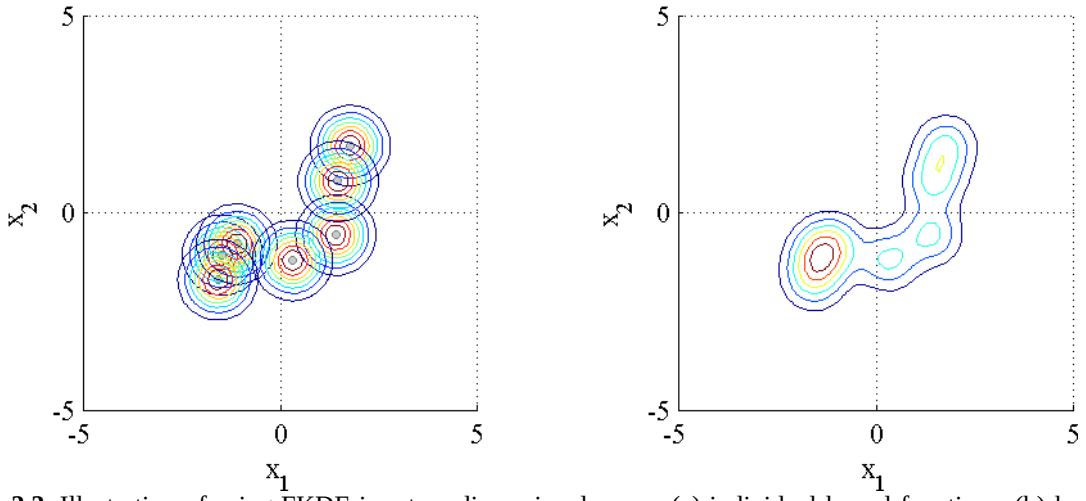


Fig. 3.3. Illustration of using FKDE in a two-dimensional space: (a) individual kernel functions, (b) kernel density estimate.

It is natural to ask that the estimate be a legitimate density function, i.e., that it is nonnegative everywhere and integrates to one:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) \geq 0 \tag{3.13}$$

$$\int_{\Re^d} \hat{f}_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = 1 \tag{3.14}$$

To this aim, any kernel function $\kappa(\mathbf{u})$ can be chosen in (3.11) as long as the following conditions are met:

$$\kappa(\mathbf{u}) \geq 0 \tag{3.15}$$

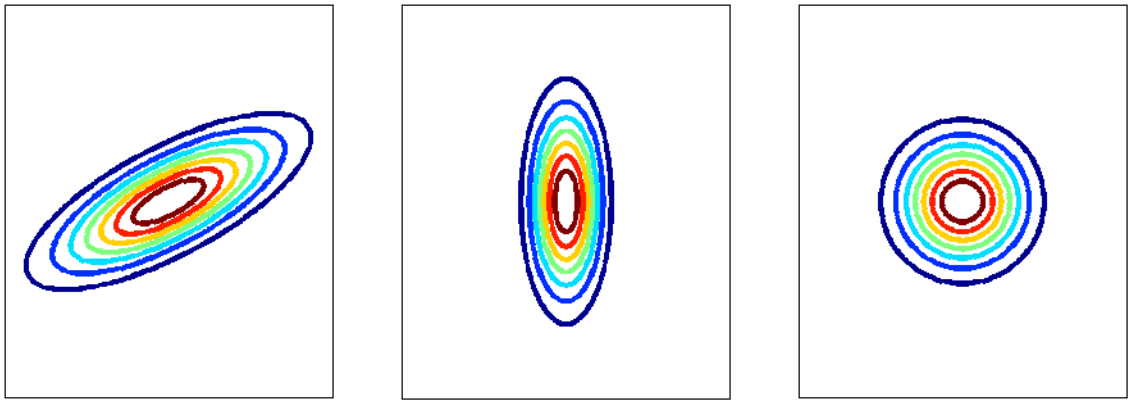
$$\int_{\Re^d} \kappa(\mathbf{u}) d\mathbf{u} = 1 \tag{3.16}$$

This is assured by imposing the kernel function $\kappa(\mathbf{u})$ be a density function [54].

It has been widely recognized that the performance of FKDE does not critically depend on the kernel function choice but, rather, on the bandwidth values employed [23][54].

In general, there are three different forms for the bandwidth matrix. The most general approach is to employ **H** chosen from the set of all symmetric, positive definite, $d \times d$ matrices, which allows ellipsoidal kernel functions of arbitrary orientation. However, this type of matrix involves $d \cdot (d+1)/2$ independent parameters that must be chosen in practice

for the use of the FKDE, and they can be a substantial number even for small dimensions. However, **H** is often parameterized to $\mathbf{H}=diag(h_1, \ldots, h_d)$, (where *diag* indicates a diagonal matrix). In such a way, a different bandwidth value is used for each dimension. Typically, a further simplification, which restricts the contours of the kernel functions to be spherically symmetric, is chosen. This straightforward simplification is obtained by imposing $\mathbf{H}=h\cdot\mathbf{I}_d$, where $\mathbf{I}_d$ denotes the $d\times d$ identity matrix. In fact, more complicated forms of **H** than $\mathbf{H}=h\cdot\mathbf{I}_d$ have been recognized to provide only very little improvements if the data are pre-scaled in order to avoid extreme differences of spread in the various spectral directions [54]. Moreover, a single bandwidth is easier to estimate as well as being also easier to interpret and simpler to control. The bandwidth selection problem will be further explored in the next chapter.



(a) (b) (c)

Fig. 3.4. Comparison of the three main bandwidth matrix parametrization classes in a two-dimensional space: (a) symmetric positive definite matrix, (b) diagonal matrix with positive entries on the main diagonal, (c) positive scalar times the identity matrix.

3.3.2 Variable bandwidth kernel density estimator

The FKDE so far described has been shown to be quite affected by the bandwidth values, which control the degree of smoothing of the resulting PDF approximation [54]. In fact, there may be several situations in which the FKDE leads to poor estimates due to an inappropriate bandwidth choice, which is constrained to be fixed across the estimation domain. In particular, in regions of high data density, choosing large values of bandwidths may obscure many of the structural features characterizing the PDF body, such as de-emphasizing or wiping out significant modes that might otherwise be extracted from the data. However, reducing bandwidth values may lead to noisy estimates elsewhere in data

space where the PDF is smaller [54], as will be detailed in Chapter 4.2.3. These considerations suggest that the amount of smoothing, dictated by the bandwidth values, should be adapted to the local data structure in the feature space.

In the literature, two main approaches can be found that have been proposed to overcome the main limitations of the FKDE. Within such approaches, the bandwidths are allowed to vary across the estimation domain, according to the data density and structure. Specifically, the distinction between the two approaches lies in how the bandwidth is varied. The first approach varies the bandwidth depending on the sample **x** where the PDF value has to be estimated and is referred to as the *balloon estimator* (BE), term used for the first time in [58] on the basis of a suggestion found in [60]. The second strategy varies the bandwidth for each data sample $\{\mathbf{x}_n \in \Re^d | \, n = 1, 2, ..., N\}$ and is referred to as the *sample-point estimator* (SPE) [58]. Both BE e SPE estimators are justified by the fact that for the local smoothness of the PDF evaluation, only those data samples in a small neighborhood of the estimation sample **x** contribute to the PDF value in **x**.

The general form of the BE is:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{N|\mathbf{H}(\mathbf{x})|} \sum_{n=1}^{N} \kappa\left[\mathbf{H}(\mathbf{x})^{-1}(\mathbf{x} - \mathbf{x}_n)\right] \tag{3.17}$$

where **H**(**x**) is the bandwidth matrix, function of the estimation sample **x**. As is evident, the PDF estimate is constructed similarly to the classical FKDE in (3.12), with the difference that the scale parameter of the kernel function placed on each sample data $\mathbf{x}_n$ is allowed to vary from one estimation sample **x** to another. It should be noted that, although this estimator appears reasonable for estimating a PDF at a point, when (3.17) is applied over the whole domain of definition of **x**, the estimate typically fails to integrate to 1 and, thus, it may be not an actual PDF. Nevertheless, this latter aspect is not critical from the detection perspective, which is the purpose of this thesis work, as long as the function in (3.17) follows the data structure in the feature space.

The alternative strategy SPE uses a bandwidth matrix function of the sample point $\mathbf{x}_n$ regardless of the estimation point **x**. The SPE, first considered in [8], is given by:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{|\mathbf{H}(\mathbf{x}_n)|} \kappa\left[\mathbf{H}(\mathbf{x}_n)^{-1}(\mathbf{x} - \mathbf{x}_n)\right] \tag{3.18}$$

where $\mathbf{H}(\mathbf{x}_n)$ is the bandwidth matrix associated with $\mathbf{x}_n$. Unlike BE, the SPE returns PDF estimates that integrate to 1 as long as the kernel function is a PDF itself.

Chapter 4

4 Model learning for global AD approaches

In chapter 3, both semi- and non-parametric estimators have been discussed as to their use for image PDF estimation.

In the semi-parametric approach, PDF estimation is carried out through a model learning procedure aimed at estimating the parameters required to characterize the FMM. In this work, we are interested in model learning procedures that are able to evaluate, without operator intervention, all the parameters necessary to completely specify the mixture.

As regards the non-parametric estimator, the model is entirely learned from the image pixels without resorting to parameter estimation. Nevertheless, FKDE performance strongly depends on the kernel bandwidth. Again, our interest relies in reliable and data-driven bandwidth selection methodologies.

4.1 Model learning for finite mixtures

Parameter estimation is a classical problem in statistics, and it can be approached in several ways. Such a learning procedure involves not only estimating the parameters of each mixture component but also finding the probabilities with which each data point belongs to the components.

Typically, a Maximum Likelihood (ML) [30] formulation is firstly sought. Such an approach views the parameters as deterministic unknown quantities and estimates them by maximizing the likelihood function. According to this approach, the ML estimate is obtained as

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{n=1}^{N} f_{\mathbf{X}}(\mathbf{x}_n; \boldsymbol{\theta}) \tag{4.1}$$

where $\{\mathbf{x}_n | \ \mathbf{x}_n \in \Re^d, \ n=1, \ldots, N\}$ are N independent and identically distributed (i.i.d.) observations coming from a distribution $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ governed by unknown parameters $\boldsymbol{\theta}=\{\boldsymbol{\theta}_j, j=1, ..., J\}$. $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ as a function of the parameters $\boldsymbol{\theta}$ is called likelihood function. Specifically, it describes the probabilistic relationship between the observations and the parameters based on the assumed model that generated the observations. The difficulty here arises from the fact that the unknown parameters often enter the maximization task in a non-linear fashion and, therefore, iterative non-linear optimization techniques have to be adopted.

The Expectation Maximization (EM) algorithm is an iterative method for finding ML estimates of parameters in statistical models that has attracted a great deal of interest in a wide range of applications [45][6]. In practice, given the number of components and an initial set of parameters, the EM algorithm can be applied to compute the optimal estimates of the parameters that maximize the likelihood function. To this aim, the EM algorithm alternates between Expectation (E) and Maximization (M) steps for updating the estimate of the unknown parameter at each iteration:

1. Expectation step produces refined estimates of the response features given the current parameter estimates.
2. Maximization step obtains new estimates of the parameters for the new response features.

These steps are repeated until the improvement in value of the log-likelihood function is less than a tolerance value.

Practical testing has shown that the actual effectiveness of the EM is affected by several limitations seriously restricting its applicability to complex problems [5][61]. First of all, convergence to a global maximum is not guaranteed. In fact, for likelihood functions with multiple maxima, EM may converge to a local maximum depending on initialization values. Another drawback of this approach is that it assumes the user knows the number of mixture components. This is not the case for many practical applications. Typically, in such cases, a set of candidate models is established by applying the EM algorithm for different possible values of the number of components J. The best model is then selected according

to a model-selection criterion, such as the Akaike Information Criterion (AIC), the Generalized Information Criterion (GIC), or the Bayesian Information Criterion (BIC) [56][16]. These methods have already been used for selecting the optimal number of components in a GMM [33][44][73]. However, it has been shown that such criteria typically are likely to fail in selecting the correct number of components. In practice, they tend to favor overly simple models. A further limitation of the EM is that it may lead to singular solutions, i.e. the density of one or more components gets concentrated around one of the data samples so that the corresponding covariance matrix becomes singular. In such a situation, the likelihood function is likely to become unlimited. This latter is the reason why EM is not suitable for estimating the number J of components, for example, by starting with a large number of components and deleting the ones whose weights approach zero.

Appropriate solutions to the aforementioned limitations involving the EM learning procedure may be obtained by adopting a Bayesian framework for estimating the parameters of the mixture [61]. As their name suggests, the hidden variables are variables whose samples are not directly observed. Rather, they can be inferred from data samples. The role of these hidden variables is either to represent hidden causes that explain the observed data samples or be just mathematical artifacts that are introduced into the model in order to simplify it properly. Bayesian approaches usually include some model parameters within the set of hidden variables in order to model them as random variables characterized by adequate priors. Involving prior knowledge makes parameters be matched with physically meaningful values. In such a way, singular solutions often arising in the EM approach where a component becomes responsible for a single data sample are avoided and automated determination of the optimal number of components J is enabled as well.

In this thesis work, Bayesian learning algorithms are considered to learn both GMM and StMM [4][12]. Nevertheless, this task can be computationally heavy and may result in intractable mathematical operations. To this aim, the variational Bayesian approach is adopted for converting the complex inferring problem into a set of simpler calculations [6][61]. The Variational Bayesian framework has been widely employed as an approximation of the Bayesian learning for models involving hidden variables.

4.1.1 The Bayesian model learning approaches

The Bayesian decision theory is a fundamental statistical approach to make inferences about models. In principle, Bayesian strategies allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler distributions. In doing so, they make reference to hidden variables characterized by prior distributions.

Formally, introducing hidden variables within the FMM assumes that for each observation $\mathbf{x}_n$ there exists a hidden variable $\mathbf{z}_n$ denoting the component that generated $\mathbf{x}_n$. Specifically, a set of label indicator vectors $\mathbf{Z}=\{\mathbf{z}_n \in \Re^J | \ n = 1, 2, ..., N\}$ can be constructed, with each $\mathbf{z}_n$ being a binary vector such that if the j-th component is responsible for $\mathbf{x}$ then $z_{nj}=1$, otherwise $z_{nj}=0$. Besides $\mathbf{Z}$, some parameters of the FMM may be absorbed in the hidden variable set, i.e. they can be modeled as random variables characterized by adequate priors, whereas the other ones are still deterministic. In a fully Bayesian model all unknown parameters are handled as random variables that are associated with prior distributions.

At this point, it is important to clarify the difference between the notation $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\varphi})$ and $f_{\mathbf{X}}(\mathbf{x}| \boldsymbol{\varphi})$. Specifically, when we write $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\varphi})$ we imply that $\boldsymbol{\varphi}$ are parameters. In contrast, when we write $f_{\mathbf{X}}(\mathbf{x}| \boldsymbol{\varphi})$, we imply that $\boldsymbol{\varphi}$ are random variables.

Within this framework, since parameters are likely to be modeled as random variables, the log-likelihood function is actually a log-marginal distribution (or marginal likelihood as it is called somewhere [61]). Once hidden variables and their prior distributions have been introduced, the log-marginal distribution is obtained by integrating out the hidden variables of the model [6][61]:

$$L(\mathbf{x}) = \ln \int f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\varphi}) d\mathbf{y} \tag{4.2}$$

where $\mathbf{Y}$ denote the set of all hidden variables and $\boldsymbol{\varphi}$ indicates the vector of deterministic parameters not absorbed into $\mathbf{Y}$.

The above expression of the log-marginal distribution can be decomposed as [6][61]:

$$L(\mathbf{x}) = F\left[q_{\mathbf{Y}}(\mathbf{y})\right] + KL\left[q_{\mathbf{Y}}(\mathbf{y}) \| \ f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} | \mathbf{x}; \boldsymbol{\varphi})\right] = F(q_{\mathbf{Y}}) + KL(q_{\mathbf{Y}} \| \ f_{\mathbf{Y}|\mathbf{X}}) \tag{4.3}$$

where $q_{\mathbf{Y}}(\mathbf{y})$ (denoted for simplicity $q_{\mathbf{Y}}$ hereinafter) is any arbitrary PDF defined over the hidden variables, while the first term $F(q_{\mathbf{Y}})$ consists in the free energy

$$F[q_{\mathbf{Y}}(\mathbf{y})] = F(q_{\mathbf{Y}}) = \int q_{\mathbf{Y}}(\mathbf{h}) \ln\left[\frac{f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{x};\boldsymbol{\varphi})}{q_{\mathbf{Y}}(\mathbf{y})}\right] d\mathbf{y} \tag{4.4}$$

and the second term $KL(q_{\mathbf{Y}}||f_{\mathbf{Y}|\mathbf{X}})$ is the Kullback–Leibler (KL) divergence between $q_{\mathbf{Y}}(\mathbf{y})$ and the posterior PDF $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x};\boldsymbol{\varphi})$ (simply indicated with $f_{\mathbf{Y}|\mathbf{X}}$ hereinafter)

$$KL[q_{\mathbf{Y}}(\mathbf{y}) \| f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x};\boldsymbol{\varphi})] = KL(q_{\mathbf{Y}} \| f_{\mathbf{Y}|\mathbf{X}}) = -\int q_{\mathbf{Y}}(\mathbf{y}) \ln\left[\frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x};\boldsymbol{\varphi})}{q_{\mathbf{Y}}(\mathbf{y})}\right] d\mathbf{y} \tag{4.5}$$

Based on (4.3), $L(\mathbf{x})$ is a functional of the distribution $q_{\mathbf{Y}}$, and a function of the parameter vector $\boldsymbol{\varphi}$. Bayesian inference methodologies are aimed at maximizing $L(\mathbf{x})$ with respect to $q_{\mathbf{Y}}$ and $\boldsymbol{\varphi}$. Since $KL(q_{\mathbf{Y}}||f_{\mathbf{Y}|\mathbf{X}})\geq 0$, it holds that $L(\mathbf{x})\geq F(q_{\mathbf{Y}})$. Therefore, $F(q_{\mathbf{Y}})$ is a lower bound for the log-marginal distribution $L(\mathbf{x})$, as is clear from Fig. 4.1. Equation (4.5) shows that the equality occurs when $KL(q_{\mathbf{Y}}||f_{\mathbf{Y}|\mathbf{X}})=0$, which implies $q_{\mathbf{Y}}= f_{\mathbf{Y}|\mathbf{X}}$. As a result, the lower bound $F(q_{\mathbf{Y}})$ can be maximized by optimization with respect to the distribution $q_{\mathbf{Y}}$, which is equivalent to minimizing the KL divergence. If we allow any possible choice for $q_{\mathbf{Y}}$, then the maximum of the lower bound takes place when the KL divergence vanishes, which occurs when $q_{\mathbf{Y}}$ equals the posterior distribution $f_{\mathbf{Y}|\mathbf{X}}$. However, the model is typically such that working with the true posterior distribution is mathematically intractable. Assuming an appropriate form for $q_{\mathbf{Y}}$ in the decomposition of (4.3) allows the exact knowledge of $f_{\mathbf{Y}|\mathbf{X}}$ to be bypassed. Thus, in Bayesian learning, direct estimation of the model parameters is replaced by the maximization of the lower bound $F(q_{\mathbf{Y}})$ with respect to the density $q_{\mathbf{Y}}$ and therefore leads to approximate posterior distribution as close as possible to the true posterior distribution.

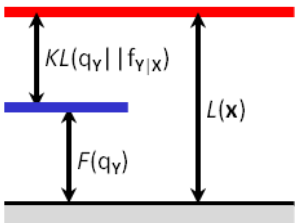


Fig. 4.1. Illustration of the decomposition given by (4.3), which holds for any choice of distribution $q_{\mathbf{Y}}(\mathbf{y})$. Because the Kullback-Leibler divergence satisfies $KL(q_{\mathbf{Y}}||f_{\mathbf{Y}|\mathbf{X}})=0$, we see that the quantity $F(q_{\mathbf{Y}})$ is a lower bound of the log-likelihood function.

In order to simplify the calculation, a variational approximation has been proposed [6][61]. The *variational* framework has been widely employed as an approximation of the Bayesian learning for models involving hidden variables. Such an approximation assumes a specific form for the distribution $q_{\mathbf{Y}}$, with respect to which the optimization is performed. Specifically, we partition the elements of $\mathbf{Y}$ into disjoint groups that we denote by Y_i where $i = 1, \ldots, P$. We then assume that the $q_{\mathbf{Y}}$ distribution is factored in these groups, so that

$$q_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{P} q_i(y_i) \tag{4.6}$$

The use of this factorized form for $q_{\mathbf{Y}}$ within the Bayesian optimization task corresponds to an approximation framework developed in physics called *mean field theory* [47]. It is to note that no further assumptions about the distribution are made. In particular, no restriction on the functional forms of the individual factors $q_i(y_i)$ is placed. Among all distributions $q_{\mathbf{Y}}$ having the form expressed by (4.6), the distribution for which the lower bound $F(q_{\mathbf{Y}})$ is largest is now sought. In principle, a free form (variational) optimization of $F(q_{\mathbf{Y}})$ with respect to all of the distributions $q_i(y_i)$ is desiderate to make. The general expression for the optimal solution $q_j^*(y_j)$ is given by

$$\ln q_{Y_j}^*(\mathbf{h}_j) = E_{i \neq j}\{\ln[f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\varphi})]\} + \text{const} \tag{4.7}$$

where the notation $\mathrm{E}_{i \neq j}(.)$ denotes the expectation with respect to the distributions $q_{Yj}(y_j)$ over all variables $\mathbf{y}_i$ for $i \neq j$, so that

$$E_{i \neq j}\{\ln[f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\varphi})]\} = \int \ln[f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y};\boldsymbol{\varphi})] \prod_{i \neq j} q_{Y_i}(y_i)\, dy_i \tag{4.8}$$

The set of equations given by (4.7) for $j = 1,\ldots,P$ represents a set of consistency conditions for the maximum of the lower bound subject to the factorization constraint. Nevertheless, the optimization task does not have closed-form analytical solutions since the factors $q_{Yi}(\mathbf{y}_i)$ are coupled together in a non-linear fashion [6][61]. So, the variational analysis optimization is carried out by employing an iterative procedure. Specifically, a consistent solution can only be found by initializing all of the factors $q_{Yi}(\mathbf{y}_i)$ and then cycling through the factors and replacing each in turn with the revised estimate given by the right-hand side of (4.7) evaluated using the current estimates for all of the other factors [6][61]. This optimization task requires that prior distribution of the hidden variables to be previously

set. Typically, prior distributions conjugate to the marginal distribution are used for their mathematical tractability. In fact, conjugate prior distributions choice leads to posterior distributions having the same functional form as the prior distributions, and, thus, to a greatly simplified Bayesian analysis.

In this work, Variational Bayesian learning algorithms are considered to learn both GMM and StMM [12][4]. In particular, Bayesian estimation is employed to allow the appropriate number J of mixture components to be automatically determined while the mixture parameters are learnt. Within this framework, the adopted Bayesian learning strategy assumes parameters as random variables with given prior probability distribution.

4.1.1.1 Gaussian mixture model learning

The shape of the GMM PDF, achieved by substituting (3.7) in (3.1), is governed by $\boldsymbol{\pi} = \{\boldsymbol{\pi}_j | j=1, 2, \ldots, J\}$, $\boldsymbol{\mu}=\{\boldsymbol{\mu}_j | j=1, 2, \ldots, J\}$, and $\mathbf{T}=\{\mathbf{T}_j | j=1, 2, \ldots, J\}$.

A fully automated method for learning the GMM by adopting a Bayesian framework was proposed in [6] and [61]. As previously reported, a fully Bayesian analysis treats all parameters as random variables with a given prior probability distribution. As a result, the main task in algorithms using Bayesian inference consists of defining proper distribution functions for modeling the parameters. Then, Bayes's rule provides the framework for combining the prior information with sample data to make inferences about the model. Due the assumption of GMM for the data, conjugate prior distributions from the exponential family are used for their mathematically tractability. That is why Dirichlet prior distribution is used for $\boldsymbol{\pi}$, whereas an independent Gauss-Wishart prior distribution is assumed for both $\boldsymbol{\mu}$ and $\mathbf{T}$ in [6] and [61]. The Dirichlet prior for $\boldsymbol{\pi}$ is given by:

$$f_{\pi}(\boldsymbol{\pi}) = C(\alpha_1,...,\alpha_J)\prod_{j=1}^{J}\pi_j^{\alpha_J-1} \tag{4.9}$$

where, by symmetry, the same α_j is chosen for each component, i.e. $\alpha_j=\alpha_0$ for $j=1, \ldots, J$, and $C(\alpha_1, \ldots, \alpha_J)$ is the normalization constant for the Dirichlet distribution. The Gauss-Wishart prior that governs the mean and the precision of each Gaussian component in equation (3.1) is given by:

$$f_{\boldsymbol{\mu},\mathbf{T}}(\boldsymbol{\mu},\mathbf{T}) = \prod_{j=1}^{J} g_N\left(\boldsymbol{\mu}_j \mid \mathbf{0}, (\beta\mathbf{T}_j)^{-1}\right) g_W(\mathbf{T}_j \mid \varsigma, \mathbf{V}) \tag{4.10}$$

Equation (4.10) is the product of a Gaussian PDF $g_N(.)$ and a Wishart PDF $g_W(.)$, which is defined as follows:

$$g_W(\mathbf{T}_j;\varsigma,\mathbf{V}) = \frac{|\mathbf{T}_j|^{(\varsigma-d-1)/2} \exp\left[-\frac{1}{2}tr(\mathbf{V}\mathbf{T}_j)\right]}{2^{\varsigma d/2}\pi^{d(d-1)/4}|\mathbf{V}|^{\varsigma/2}\prod_{i=1}^{d}\Gamma\left(\frac{\varsigma+1-i}{2}\right)} \tag{4.11}$$

with $\Gamma(.)$ denoting the gamma function defined in (3.9), tr(.) denotes the trace, and parameters ς and $\mathbf{V}$ denote the degrees of freedom and the scale matrix, respectively. Within this framework, α_0, β, ς and $\mathbf{V}$ are called hyperparameters, and they have to be specified in advance. The corresponding graphical model[1] of this learning procedure is shown in Fig. 4.2. It should be emphasized that this is a fully Bayesian GMM. So, if all the hyperparameters (i.e., the parameters α_0, β, ς and $\mathbf{V}$ of the prior distributions) are specified in advance, then the model does not contain any parameter to be estimated, but only the hidden random variables $\mathbf{Y}=(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{T})$ whose posterior distribution $q_{\mathbf{Y}}$ given the data must be computed. It is obvious that such a posterior distribution cannot be computed analytically, thus an approximation for $q_{\mathbf{Y}}$ is computed by applying the variational approximation expressed by (4.6) to this specific Bayesian model. The solution is given by (4.7).

One advantage of the fully Bayesian GMM compared to GMM without prior distributions is that it does not allow the singular solutions often arising in the ML approach where a Gaussian component becomes responsible for a single data point. In addition, this model learning method allows for the optimal number of components to be determined, without resorting to strategies such as the model-selection criterion previously mentioned. In principle, during the optimization procedure, as soon as one of the mixing coefficients converges to zero, the corresponding component is eliminated from the mixture. However, the effectiveness of the fully Bayesian mixture is limited, since the Dirichlet prior

[1] The graphical models are graphs in which nodes correspond to random variables and arrows represent the dependencies among such random variables. In particular, the doubly circled nodes represent observed random variables and nodes denoted as squares correspond to model parameters. The boxes (plates) indicate independent copies of the random variables they enclose, the number of which is depicted in a corner of each plate.

distribution for $\boldsymbol{\pi}$ does not allow the mixing weight of a component to become zero and, hence, the corresponding component to be eliminated from the mixture. Also, the final result highly depends on the hyperparameters characterizing the prior distributions. For a specific set of hyperparameters, it is possible to run the algorithm several times and, then, keep the solution corresponding to the best value of the variational lower bound.

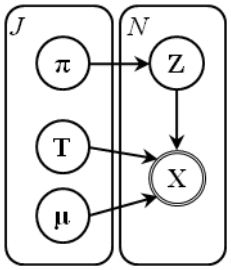


Fig. 4.2. Graphical model for the fully Bayesian GMM in which any unknown parameters are characterized by prior distributions. It is to note that the parameters of the prior distributions on $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, are fixed, thus they are not shown.

In [14], another example of a Bayesian GMM model has been proposed that does not assume a prior distribution over the mixing weights $\{\boldsymbol{\pi}_j | j=1, 2, \ldots, J\}$, which are thus treated as parameters and not as random variables. As a result, the hidden random variable are now $\mathbf{Y}=(\mathbf{Z}, \boldsymbol{\mu}, \mathbf{T})$. The graphical model for this approach is depicted in Fig. 4.3. This approach assumes Gaussian and Wishart prior distributions for $\boldsymbol{\mu}$ and $\mathbf{T}$, respectively, i.e:

$$f_{\mu}(\boldsymbol{\mu}) = \prod_{j=1}^{J} g_N\left(\boldsymbol{\mu}_j \mid \mathbf{0}, \beta\mathbf{I}\right) \tag{4.12}$$

$$f_T(\mathbf{T}) = \prod_{j=1}^{J} g_W\left(\mathbf{T}_j \mid \varsigma, \mathbf{V}\right) \tag{4.13}$$

This Bayesian model is able to estimate the optimal number of components. Specifically, the method starts with a large number of components specified by the user and, as the number of iterations increases, the number of components gradually decreases and, finally, the GMM model for the data set is attained. This happens because the prior distribution on $\boldsymbol{\mu}$ and $\mathbf{T}$ penalizes overlapping components. Thus, during the optimization process following the variational methodology, some of the mixing coefficients converge to zero and the corresponding components are eliminated from the mixture. In general, this

methodology constitutes an effective method exhibiting good performance in the case where the components are well separated [61]. However, its performance exhibits sensitivity on the specification of the scale matrix **V** of the Wishart prior imposed on the precision matrix.

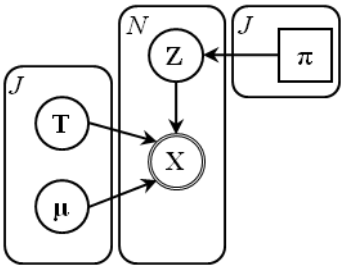


Fig. 4.3. Graphical model proposed in [14]. In this case, $\boldsymbol{\pi}$ is not circled to denote the special treatment of the mixing weights as parameters without prior distributions. The GMM learning procedure fits a mixture initialized with a large number of components and lets competition to eliminate the redundant ones. During the optimization process if some of the components fall in the same region in the data space, then there is strong tendency in the model to eliminate the redundant components (i.e., setting their π_j equal to zero), once the data in this region are sufficiently explained by fewer components. Consequently, the competition between mixture components suggests a natural approach for addressing the model selection problem: fit a mixture initialized with a large number of components and let competition eliminate the redundant.

A recently proposed method that simultaneously trains the mixture, adjusts the number of components, and reduces the sensitivity to **V** was proposed in [12] and [13]. This methodology will be denoted with Bayesian GMM Split (BGMMS) hereinafter. The method follows an incremental structure. Starting with J=1, it progressively adds components to the model. To this aim, the mixture components are partitioned in two groups: the "fixed" components and the "free" components. At each iteration, a *splitting test* is applied to one of the existing mixture components. The outcome of this test controls the procedure for component addition since it decides if the component should be properly split into two sub-components. If the splitting is found to give a better representation of the data, Variational Bayesian learning is applied to the newly added pair of sub-components (the "free" components), while the others remain "fixed"; otherwise, the splitting is not applied since it is considered redundant. Whenever the splitting test provides a positive outcome, the number of mixture components increases and a new round of splitting tests is sequentially applied to all components. The learning procedure ends when all mixture components have been unsuccessfully tested to be split. In order to apply this method, prior distributions have to be imposed on the parameters π_j, $\boldsymbol{\mu}_j$, and $\mathbf{T}_j$ of each component. Again, due the assumption of GMM for the data, conjugate prior distributions from the

exponential family are used. The set of hidden variables and the corresponding conjugate prior distribution characterizing this approach are summarized in Table 3.

Table 3. Hidden variables and corresponding prior distributions for BGMMS.

Hidden variable set	
Y={Z, $\{\boldsymbol{\theta}_j\}_{j=1}^{J} = \{\boldsymbol{\mu}_j, \mathbf{T}_j\}_{j=1}^{J}$, $\{\pi_j\}_{j=1}^{J}$ **}**	
Parameter set	
-	
Hidden variable	**Prior distribution**
Z	Product of multinomials
$\boldsymbol{\pi}$ -"fixed" weights	Dirichlet
$\boldsymbol{\pi}$ -"free" weights	Uniform
$\boldsymbol{\mu}$	Gaussian
T	Wishart

Specifically, this approach assumes a Gaussian and Wishart prior distribution for $\boldsymbol{\mu}_j$ and $\mathbf{T}_j$, respectively. In practice, the Gauss-Wishart prior distribution is the product of a Gaussian PDF $f_N(.)$ and a Wishart PDF $f_W(.)$. It also fixes a uniform prior distribution over the set of "free" mixing coefficients $\tilde{\boldsymbol{\pi}} = \{\tilde{\pi}_j\}$ and a Dirichlet prior distribution over the set of "fixed" mixing coefficients $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_j\}$. These choices allow weights of the "free" components to become zero and be eliminated from the mixture, while prevent the elimination of the "fixed" components from the model. The graphical model of the learning procedure is represented in Fig. 4.4.

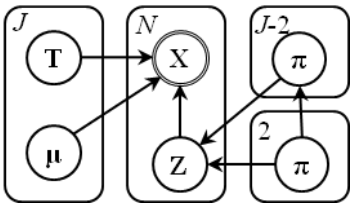


Fig. 4.4. Graphical model. The GMM learning procedure starts with one component and progressively adds components to the model on the basis of a splitting test. At each iteration, the *splitting test* decides if the two sub-components returned by the split provide a much better fit to the data in their influence region. In the case the splitting is found to give a better representation of the data, both components will survive so that the number of model components will be increased and a new round of splitting tests for all the existing components is initialized. Otherwise, the initial component will be recovered. To the learning aim, the mixing coefficients, the mean vectors **μ**, and the precision matrices **T** are defined as random variables characterized by proper prior distributions, as shown in the graph.

4.1.1.2 Student's t mixture model learning

The StMM is controlled by $\boldsymbol{\pi}=\{\pi_j|j=1, 2, \ldots,J\}$, $\boldsymbol{\mu}=\{\boldsymbol{\mu}_j|\ j=1,2, \ldots, J\}$, $\boldsymbol{\Lambda}=\{\boldsymbol{\Lambda}_j|j=1, 2, \ldots, J\}$, and $\mathbf{v}=\{v_j|\ j=1, 2, \ldots, J\}$.

A method developed within a Variational Bayesian framework, was proposed that learns the StMM while simultaneously adjusting the number of components in a fully automatic fashion [4]. This approach will be referred to as Bayesian StMM (BStMM) hereinafter. The model-order is selected according to the maximum of the lower bound $F(q_{\mathbf{Y}})$. Within this framework, in order to implement the maximization with respect to $q_{\mathbf{Y}}$, the *mean field* approximation is adopted. For the optimization task, the variational methodology is followed so that an iterative algorithm is derived.

The BStMM method is based on the fact that the StMM can be viewed as a hidden variable model itself. This can be understood by noting that (3.8) can be re-written as an infinite mixture of scaled Gaussian distributions:

$$St\left(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Lambda}_j,v_j\right)=\int_0^{+\infty} g_N\left(\mathbf{x}\,|\,\boldsymbol{\mu}_j,u_j\boldsymbol{\Lambda}_j\right)g_G\left(u_j\,|\,\frac{v_j}{2},\frac{v_j}{2}\right)du_j \tag{4.14}$$

where u_j is the scaling factor and $g_G(.)$ indicates the Gamma distribution following the expression:

$$g_G(u_j)=\frac{1}{\Gamma\left(v_j/2\right)}\left(\frac{v_j}{2}\right)^{v_j/2} u_j^{\,v_j/2-1}e^{-u_j\cdot v_j/2} \tag{4.15}$$

where $\Gamma(.)$ is the gamma function defined in (3.9). Based on (8), the scaling factor u_j follows a Gamma distribution with parameters depending only on v_j [4]. For each observation $\mathbf{x}_n$ there is a corresponding posterior distribution over the hidden variable u_j specifying the scaling of the precision matrix of the corresponding equivalent Gaussian from which the data sample was hypothetically generated. The scale variable u_{nj} (associated to the n-th data point and the j-th component), given the component label z_{nj}, is unobserved. The employment of the set of hidden variable $\mathbf{U}=\{\mathbf{u}_n \in \Re^J|\ n = 1, 2, ..., N\}$ makes the conditional probability $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ easy to compute.

The Bayesian formulation of the StMM is complete when imposing priors on $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$. Specifically, this method models the parameters of the distribution as random variables

assuming the Dirichlet prior distribution for $\boldsymbol{\pi}$, and an independent Gauss-Wishart prior distribution for both $\boldsymbol{\mu}_j$ and $\boldsymbol{\Lambda}_j$ [4]. Again, the Gauss-Wishart prior distribution corresponds to the product of a Gaussian and a Wishart distribution:

$$f_{\boldsymbol{\mu},\mathbf{T}}(\boldsymbol{\mu},\mathbf{T}) = \prod_{j=1}^{J} f_N\left(\boldsymbol{\mu}_j \mid \mathbf{m}_0, (\eta_0 \boldsymbol{\Lambda}_j)^{-1}\right) f_W\left(\mathbf{T}_j \mid \gamma_0, \mathbf{S}_0\right) \tag{4.16}$$

where $\mathbf{m}_0$, η_0, γ_0, and $\mathbf{S}_0$ are hyperparameters. The set of hidden variables and prior distributions of this method are summarized in Table 4. It should be noted that no prior distribution is imposed on the number of degrees of freedom ν_j of each mixture component, which is assumed as a parameter and not as a random variable. Then, the variational Bayesian learning methodology deriving from (4.7) is followed. Since no prior is imposed on the degrees of freedom, they are updated by maximizing the expected log-likelihood. The corresponding graphical model is depicted in Fig. 4.5.

Table 4. Hidden variables and corresponding prior distributions for BStMM.

Hidden variable set	
$\mathbf{Y}=\{\mathbf{Z}, \mathbf{U}, \{\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j\}_{j=1}^{J}, \{\pi_j\}_{j=1}^{J}\}$	
Parameter set	
$\{\nu_j\}_{j=1}^{J}$	
Hidden variable	**Prior distribution**
Z	Product of multinomials
$\boldsymbol{\pi}$	Dirichlet
$\boldsymbol{\mu}$	Gaussian
T	Wishart

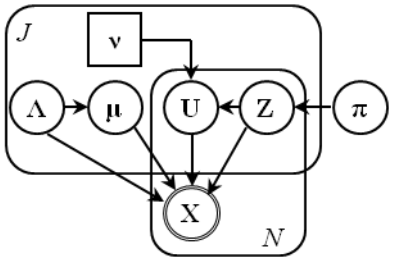


Fig. 4.5. Graphical model. In the StMM learning strategy, each observation $\mathbf{x}_n$ is conditionally dependent on both the label indicator vector $\mathbf{z}_n$ and the scale vector $\mathbf{u}_n$, which are unobserved. The set of scale vectors, included in **U**, are conditionally dependent on set of label indicator variables, included in **Z**. It is important to note that the scale variables in **U** and the label indicator variables in **Z** are contained in both plates, meaning that there is one such variable for each component and each data point. Moreover, according to the Gauss-Wishart prior distribution employed within the Bayesian analysis procedure, the mean vector of each component depends on the precision matrix of the component itself. As regards the numbers of degrees of freedom, they are considered as parameters with no prior distribution and their values are assessed by the maximum likelihood criterion.

It is important to note that, while estimating the background PDF of a hyperspectral image, this method may not be directly applied by combining equations (3.1) and (3.8). Since hyperspectral data distribution is generally characterized by heavy tails [40], the number of degrees of freedom (on which no prior distribution is imposed) of one or more mixture components is very likely to assume values smaller than 2. With a choice like that, the corresponding covariance matrix $\mathbf{C}_j$ is not defined (i.e. when $v_j \leq 2$) and, then, the scale matrix $\mathbf{\Lambda}_j$ cannot be evaluated from the data. Here, such a situation is avoided by assuring that v_j never becomes lower than 2.

4.2 Bandwidth selection for Non-Parametric PDF Estimation

Non-parametric density estimator performance has been widely recognized to be significantly affected by the bandwidths employed, since they control the kernel function smoothing [54]. In fact, as the bandwidths become smaller, the shape of the kernel function becomes narrower and more peaked, so that the influences of each individual kernel function is more localized in the feature space around its mean value. On the other hand, the larger the values are, the broader the kernel function shape becomes and a smoother estimate is obtained. Therefore, the bandwidths should be neither too large nor too small in order to obtain good results. Furthermore, this task becomes even more complicated within the AD framework, since detection should be carried out in a data-driven fully automatic fashion, that is, without operator intervention.

4.2.1 Choosing the bandwidth in the Fixed Kernel Density Estimator (FKDE)

It has been widely recognized that the performance of FKDE suffers very little from the kernel function choice but is significantly affected by the bandwidths employed. If they are too large, the resulting PDF approximation is affected by over-smoothing, which is likely to mask the multimodal nature of the distribution. On the contrary, if the bandwidths are too small, the PDF estimate is likely to result under-smoothed, by exhibiting spurious structures, especially in the distribution tails [54].

In the literature, methods discussing the bandwidth selection problem for multivariate data are very limited. The most frequently used methods of bandwidth selection are the

plug-in methods, in particular rule-of-thumb bandwidths, and the cross-validation [23][51][73].

In general, bandwidth selection strategies approximate the bandwidth by minimizing an error measurement under specified conditions. There are many possible error criteria from which to choose. A common global error criterion is the Mean Integrated Squared Error (MISE) [23], which is defined as follows:

$$MISE = E\left\{ \int_{\Re^d} \left[\hat{f}(\mathbf{x}) - f(\mathbf{x}) \right]^2 d\mathbf{x} \right\} \tag{4.17}$$

with E{.} denoting the expectation operator. The ideal MISE-optimal bandwidth selector is:

$$\mathbf{H}_{MISE} = \arg\min_{\mathbf{H} \in F} MISE \tag{4.18}$$

where F is the space of symmetric and positive definite $d \times d$ matrices. Since MISE does not have a tractable closed form in general [72], $\mathbf{H}_{MISE}$ is extremely difficult to find.

The plug-in bandwidth selection gives a formula for the bandwidth deriving from the minimization of the Approximated Mean Integrated Squared Error (MISE), an approximation of the MISE [23]. For the multivariate FKDE the AMISE formula was derived in [72] as:

$$AMISE = \frac{1}{4}\mu_2^2(\kappa) \int \left\{ tr^2 \left[\mathbf{H}^t \mathbf{H}_H(\mathbf{x}) \mathbf{H} \right] \right\} d\mathbf{x} + \frac{1}{N|\mathbf{H}|} \|\kappa\|_2^2 \tag{4.19}$$

with $\|\kappa\|_2^2$ denoting the d-dimensional squared L_2-norm of the kernel function and $\mathbf{H}_H(\mathbf{x})$ being the Hessian matrix of the second partial derivatives of the function $f(\mathbf{x})$ [23]. In order to make progress under this criterion a reference PDF, the kernel function, and a particular form of $\mathbf{H}$ must be set. Multivariate data-driven full bandwidth selectors based on these plug-in ideas were firstly proposed by [20], focusing on the very simplified bivariate case. Diagonal plug-in bandwidth matrix selectors for bivariate density estimation, for which it is impossible to obtain explicit expressions for the asymptotically optimal bandwidth matrix for general multivariate kernel density estimators, were studied in [73]. Typically, observed data arising from the multivariate normal PDF are assumed so that rule-of-thumb

formulae can be easily derived. In Table 5 are reported rule of thumb formulae for diagonal **H** matrices and normal reference distribution when a multivariate Gaussian kernel function is employed in the FKDE. Such methods are often used in practice despite the fact that most data are typically strongly non-Gaussian. This is especially true as to data from hyperspectral imagery [40][55]. Rule-of-thumb formulae for different distributional assumptions can be found in [72].

Table 5. Rules of thumb formulae.

Kernel function	$\kappa(\mathbf{u})$ [1]
Silverman's rule	$\mathbf{H} = diag(h_j)$ with $h_j = \left[\frac{4}{(d+2)N}\right]^{1/(d+4)} \sigma_j$, j=1, …, d
Scott's rule	$\mathbf{H} = diag(h_j)$ with $h_j = \left(\frac{1}{N}\right)^{1/(d+4)} \sigma_j$, j=1, …, d
Generalization of Scott's rule	$\mathbf{H} = \left(\frac{1}{N}\right)^{1/(d+4)} \mathbf{\Sigma}^{1/2}$

[1] $\mathbf{\Sigma}$ is the $d \times d$ covariance matrix of data whereas $\{\sigma_j, j=1,\ldots,d\}$ indicates the set of standard deviations (one standard deviation for each of each spectral component).

The cross-validation method [23][51] aims at deriving bandwidths that minimize the Integrated Squared Error (ISE) [23]:

$$ISE = \int_{\Re^d} \left[\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2 d\mathbf{x} \tag{4.20}$$

Cross-validation matrix selectors for an arbitrary number of dimensions were studied in [18] and [19]. Diagonal cross-validation-type matrix selectors were instead considered in [51]. However, when the data dimension grows there is an increased difficulty in numerically deriving optimal bandwidths. Also, such a procedure generally leads to large variability in the estimated bandwidths, depending on the selection of specific data samples [34].

Recently, an unsupervised method for estimating the kernel bandwidths, based on a Bayesian approach, has been proposed in [7]. This strategy will be referred to as BN approach from the initials of the two authors of the work [7]. The BN approach does not require any specific assumption for both the data distribution and the kernel function. The foundation of the method is that the degree of smoothing to adopt should be tailored to the

local data spread in the feature space, which is statistically characterized by the data variance. Hence, the bandwidth selection problem is cast in terms of assessing the distribution of the random variable S associated to the data variance. To this aim, a set of realizations s_i of the variance random variable S needs to be extracted from the data. Since the ultimate interest is to estimate the PDF with the proper spectrally local smoothness, the realizations of variance are evaluated from data spectral subsets. For this purpose, a certain number of nearest neighbors to randomly selected *centroids* $\mathbf{x}_{i'}$ are used to evaluate the variance realizations. The nearest neighbors to a specific centroid are evaluated according to their Euclidean distance from the centroid itself. Let us consider k, i.e., the number of nearest neighbors to a specific data sample $\mathbf{x}_i$. All the other data samples are ordered according to their Euclidean distance to $\mathbf{x}_i$ as

$$\left\|\mathbf{x}_{i,1}-\mathbf{x}_i\right\| < \left\|\mathbf{x}_{i,2}-\mathbf{x}_i\right\| < \ldots < \left\|\mathbf{x}_{i,k}-\mathbf{x}_i\right\| < \ldots < \left\|\mathbf{x}_{i,N-1}-\mathbf{x}_i\right\| \tag{4.21}$$

where $\mathbf{x}_{i,j}$ is the j-th ordered data sample according to the Euclidean distance from $\mathbf{x}_i$ and $\mathbf{x}_{i,j} \neq \mathbf{x}_i$ for j=1,…,N-1. For each centroid, the number of nearest neighbors to retain is defined by sampling a uniform distribution limited in the interval $[K_l, K_u]$, which must be chosen by the user. The bounds K_l and K_u for such an interval are usually given as a fraction of the number N of data. Furthermore, in order to make the strategy robust, a number of neighborhoods of various sizes, $\{K_j | \ j=1, \ldots, n\}$, are considered for each selected sample $\mathbf{x}_{i'}$. Thus, the samples s_i are calculated as:

$$s_i = \frac{\sum_{k=1}^{K_j}\left\|\mathbf{x}_{i,(k)}-\mathbf{x}_{i'}\right\|^2}{K_j - 1} \tag{4.22}$$

where $\{\mathbf{x}_{i,(k)}, k=1, \ldots, K_j\}$ are the nearest neighbors to the sampled data $\boldsymbol{x}_{i'}$. Then, within a Bayesian analysis procedure, a prior Gamma distribution is assumed for the random variable S. The Gamma distribution is given by the following expression

$$g_S(s) = \frac{1}{\Gamma(a)} b^a s^{a-1} e^{-bs} \tag{4.23}$$

where the parameters a and b are subject to the constraints $a > 0$ and $b > 0$ in order to ensure that the distribution is a legitimate PDF (i.e. non-negative and integrate to one), and

Γ(.) is the gamma function defined in (3.9). In fact, Gamma distribution is suitable for modeling the distribution of the variance when that of the underlying data is unknown. Once the Gamma distribution parameters are inferred from the variance realizations through the maximum likelihood criterion [30], they are used to compute an equal bandwidth h in all dimensions, corresponding to $\mathbf{H}=h\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. In Fig. 4.6, the graphical model for such an approach is shown.

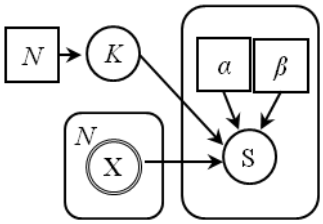


Fig. 4.6. Graphical model. After sampling the uniform distribution limited to the range $[K_l,K_u]$, whose limits are given as a fraction of the number N of data, in order to obtain K, a set of centroids is randomly sampled from the data. A data subset involving the K-nearest neighbors from each centroid (obtained according to their Euclidean distance) is taken into account to assess the realizations of S. A number of neighborhoods of various sizes are considered. The Gamma prior is employed to model S. Specifically, the parameters α and β of the Gamma distribution are inferred from the variance realizations according to the maximum likelihood criterion. The bandwidth h is estimated as the mean of the highlighted Gamma function.

It should be noted that adopting $\mathbf{H}=h\mathbf{I}$ means employing spherically symmetric kernels. By pre-scaling the data in order to avoid extreme difference of spread in the various spectral directions, more complicate forms of the bandwidth matrix (i.e. a diagonal, or a full symmetric semi-positive definite matrix) are not necessary to adopt since they have been recognized to provide very little improvements [54].

4.2.2 Choosing the bandwidth in the Variable-bandwidth Kernel Density Estimator (VKDE)

The FKDE capability of estimating PDFs is strongly influenced by the choice of the bandwidth matrix, which controls the degree of smoothing of the resulting approximation. If the bandwidths are small, each training sample has a significant effect in a small region and no effect on distant points, whereas when the bandwidths are large, there is more overlap of the kernels and a smoother estimate is obtained. Therefore, the use of fixed bandwidths is not effective when the sample data exhibit different local peculiarities across the entire data domain [35]. In fact, regions of high density in the feature space (i.e., highly

populated regions) require small bandwidths so as not to wipe out important details characterizing the PDF body during the estimation process, whereas larger bandwidths are more appropriate in low-density areas where the few sample data available are likely to generate spurious structures. These reasons suggest the employment of a VKDE to adapt the amount of smoothing to the local density of data samples in the feature space, so as to more reliably and accurately follow the multivariate background data structure of multispectral images of a scene [58].

In this thesis work, both the BE and SPE are introduced to improve the performance of FKDEs in estimating the background PDF in hyperspectral images. Similarly to FKDE, in order to apply BE and SPE the kernel function and the bandwidth function **H**(.) should be imposed. The *k*-nearest neighbor method (*k*-NN) [54] represents an attempt to choose both **H**(**x**) and **H**($\mathbf{x}_n$) in order to adapt the amount of smoothing to the local density of the data. Specifically, the k-NN relies upon an integer *k*, chosen to be considerably smaller than the sample size *N*, to control the degree of smoothing of the PDF estimation. Within this framework, the BE formulation is equivalent to take $\mathbf{H}(\mathbf{x})=r_k(\mathbf{x})\mathbf{I}_d$ in (3.17), where $r_k(\mathbf{x})$ is the Euclidean distance of **x** to the k^{th} nearest sample in the set $\{\mathbf{x}_n \in \Re^d | \, n = 1, 2, ..., N, \mathbf{x}_n \neq \mathbf{x}\}$. In this way, the width of the kernel placed on each point $\mathbf{x}_n$ is equal to $r_k(\mathbf{x})$, so that data sample lying in regions where the data are sparse will have flatter kernel functions associated with them, whereas in more populated regions narrower kernel functions will be used. The resulting PDF estimator is called the generalized k^{th} nearest neighbor estimator (GkNNE) and can be written as [54]:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{N \cdot [r_k(\mathbf{x})]^d} \sum_{n=1}^{N} \kappa\left(\frac{\mathbf{x}-\mathbf{x}_n}{r_k(\mathbf{x})}\right) \tag{4.24}$$

Within this framework, the choice of the kernel function affects the precise integrability of the PDF estimation [54]. However, it has been shown [54] that the GkNN may have more reasonable tail behavior if the kernel function is smooth and radially symmetric [54]. This specific kind of PDF estimator was introduced in [35], where a uniform density on the unit sphere in $\Re^d$ was suggested to be used as kernel function. This choice leads to the k^{th} nearest neighbor estimator (kNNE):

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{k-1}{N} \frac{d \cdot \Gamma(d/2)}{2\pi^{d/2} \cdot \left[r_k(\mathbf{x})\right]^d} \tag{4.25}$$

where Γ(.) denotes the Gamma function. The major drawback of this estimator is that when (4.25) is used to estimate a PDF over the extension of the entire domain of **x**, the resulting estimate does not integrate to 1.

Similarly, the use of the k-NN method within the SPE approach results in:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{N} \cdot \sum_{n=1}^{N} \frac{1}{\left[r_k(\mathbf{x}_n)\right]^d} \kappa\left(\frac{\mathbf{x}-\mathbf{x}_n}{r_k(\mathbf{x}_n)}\right) \tag{4.26}$$

where $r_k(\mathbf{x}_n)$ is the distance of $\mathbf{x}_n$ to the *k*-nearest data sample within $\{\mathbf{x}_j \in \Re^d | \, j = 1, 2, ..., N, j \neq n\}$. This is equivalent to choose $h(\mathbf{x}_i) \propto f(\mathbf{x}_i)^{-1/d}$ and, in practice, to use a pilot estimate of the PDF to calibrate the bandwidth matrix [54]. Moreover, the choice of a kernel function as a PDF assures that $\hat{f}_{\mathbf{X}}(\mathbf{x})$ is an actual PDF [54].

4.2.3 Fixed vs. variable bandwidths: evaluation of the kernel PDF estimates on a "toy example"

As mentioned, although the FKDE is undoubtedly the most widely adopted non-parametric technique for modeling data, the variable-bandwidth kernel density estimators have been suggested to improve the PDF estimation reliability. In this section, results on a simple "toy example" are presented in order to investigate the ability of the proposed variable-bandwidth kernel density estimators with respect to the FKDE in assessing the image background PDF [63].

Comparing PDF estimators is a difficult task, especially in the multivariate setting. Moreover, in a spectral dimension higher than *d*=2, only part of the features of a PDF may be graphically displayed. Therefore, for the experiments, we constructed an image of size 500 × 500 pixels consisting in only 2 spectral dimensions thus simplifying results interpretation by enabling graphical representation of the estimation outcome. To this aim, the data were generated following a mixture of two bivariate (*d*=2) Gaussian distributions with parameters related to the Moffett Field data set, which was collected by the AVIRIS sensor and is available online [25], thus making the simulation more realistic. Specifically,

a portion consisting of 571 by 187 pixels of the entire flight line and including the information contained in the green and red bands was considered. In order to select homogenous pixels to build the two mixture components, mixture parameter learning was conducted through the well-known Expectation Maximization (EM) approach [45]. In the EM strategy, the number of mixture components is a user-specified parameter. In this case of study, the number of components was set according to a visual inspection of the spectral diversity of the scene. Next, the parameters of the two more compact and well separated components were selected to be used in the mixture model. Finally, since the approaches to be tested assume kernel functions that are spherically symmetric (i.e., an equal bandwidth across the two spectral dimensions) the data were normalized in order to equally spread the data in all spectral directions. To this aim, the data were linearly transformed to have zero mean and unit covariance matrix, as typically suggested in the literature [26][54] and proposed in [22]. The resulting PDF, which is graphically displayed in Fig. 4.7, resulted in the following form:

$$f_{\mathbf{X}}(\mathbf{x}) = 0.5 \cdot g_N(\mathbf{x};\boldsymbol{\mu}_1,\mathbf{C}_1) + 0.5 \cdot g_N(\mathbf{x};\boldsymbol{\mu}_2,\mathbf{C}_2) \tag{4.27}$$

where $\{g_N(\mathbf{x};\boldsymbol{\mu}_i,\mathbf{C}_i)\}_{i=1,2}$ denote the Gaussian PDFs characterized by mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{C}_i$. Specifically, the model parameters in (4.27) were $\boldsymbol{\mu}_1$=[-0.08; 0.98], $\boldsymbol{\mu}_2$=[0.08; 0.97], $\mathbf{C}_1$=[1.35 -0.20; -0.20 0.06], $\mathbf{C}_2$=[0.63 0.04; 0.04 0.03].

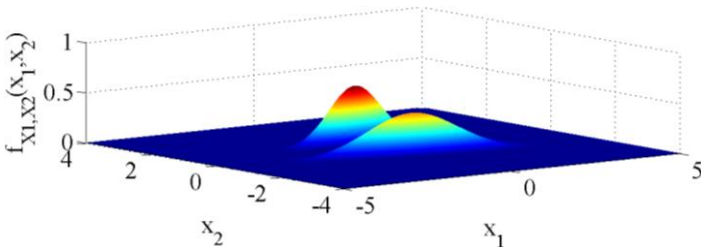


Fig. 4.7. The true PDF of the Gaussian mixture model employed in the "toy-example".

The simulated image background data thus constructed were processed by the PDF estimation techniques described in chapter 3. Since evaluating of the ability to estimate PDFs is not a trivial task, the experimental study involved both qualitative analysis by visual inspection of the estimated PDFs and comparisons based on quantitative error measures. Specifically, three measures of error, usually employed in the statistical literature, were taken into account to evaluate the behavior of the PDF estimators [8]:

- Mean percentage error (MPE)

$$MPE = 100\frac{1}{N}\sum_{n=1}^{N}\frac{\left|f(\mathbf{x}_n) - \hat{f}(\mathbf{x}_n)\right|}{f(\mathbf{x}_n)} \tag{4.28}$$

- Mean absolute percentage error (MAPE)

$$MAPE = 100\frac{1}{N\mu_f}\sum_{n=1}^{N}\left|f(\mathbf{x}_n) - \hat{f}(\mathbf{x}_n)\right| \tag{4.29}$$

where $\mu_f = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$.

- Mean square percentage error (MSPE)

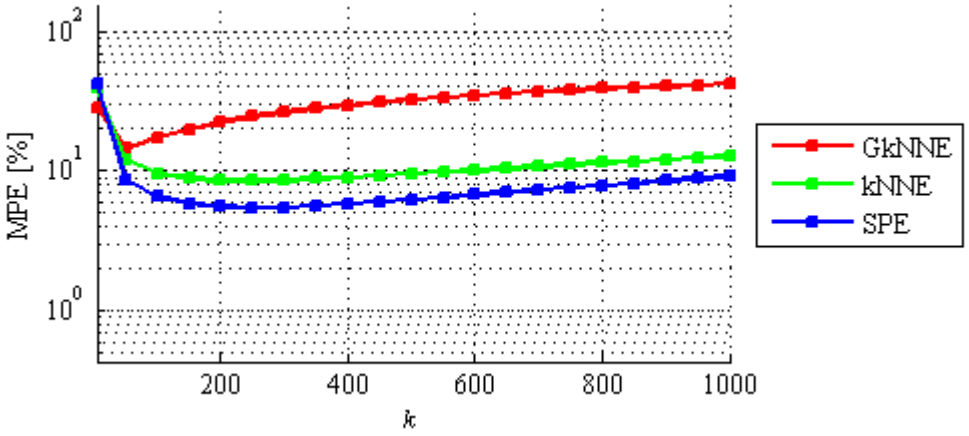
$$MSPE = 100\frac{1}{N\sigma_f^2}\sum_{n=1}^{N}\left[f(\mathbf{x}_n) - \hat{f}(\mathbf{x}_n)\right]^2 \tag{4.30}$$

where $\sigma_f^2 = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \mu_f\right]^2$.

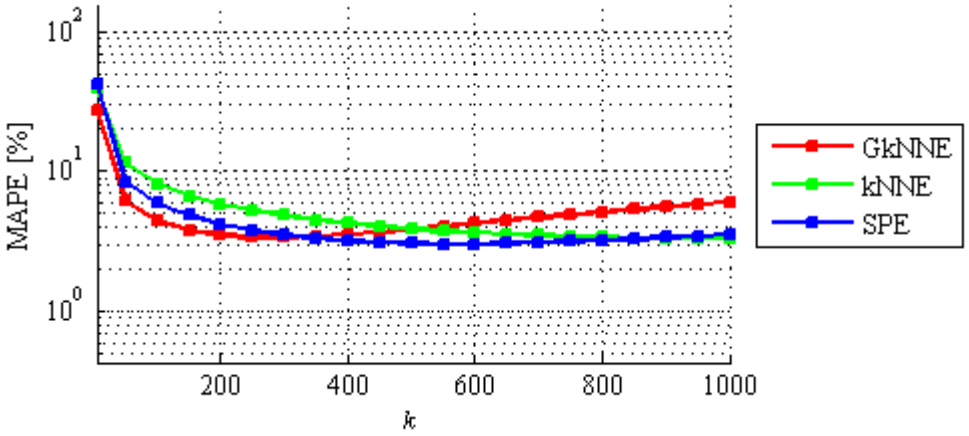
Basically, all these three measures quantify how close the estimates are to the actual values of the PDF being estimated. The smaller the measures are, the better the estimation performance is. Whereas both MAPE and MSPE measure the general behavior of the estimator across the whole estimation domain, the MPE is more sensitive to estimation behavior in the distribution tails. In fact, in eq. (4.28) the reciprocals of the true PDF values provide the weights for the different absolute errors resulted from the set of estimates. In such a way, samples in the distribution tails are associated with weights greater than those of the body of the distribution. Thus, if the estimate has heavy MPE values, the distribution tails are not suitably characterized.

The experimental comparative analysis first involved the VKDEs expressed by equations (4.24) - (4.26), and denoted with GkNNE, kNNE, and SPE, respectively. In this work, as commonly done in the literature [23][54], for both GkNNE and SPE $\kappa(.)$ is taken to be a multivariate Gaussian PDF. The ability of the VKDE to provide reliable PDF estimates was investigated with respect to different choices of the integer k, in order to evaluate the impact of the only user-specified parameter over the estimation performance. To this aim, k was varied from 5 up to 1000.

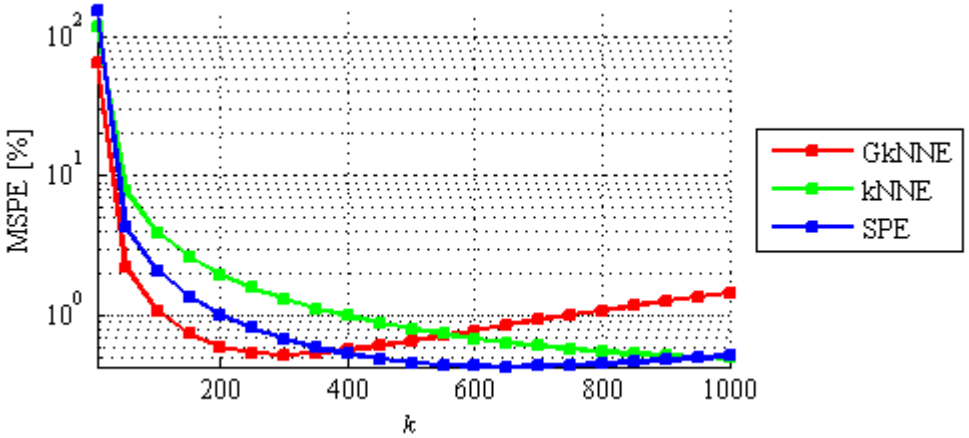
In Fig. 4.8 the plots of the error measures previously mentioned (i.e. MPE, MAPE, and MSPE) associated to the use of GkNNE, kNNE, and SPE are reported for the different configurations of *k* explored.



(a)



(b)



(c)

Fig. 4.8. Graph of MPE (a), MAPE (b), and MSPE (c) measures resulted from the application of GkNNE, kNNE, and SPE over the "toy-example" data set as *k* varies from 5 to 1000.

As is evident, results obtained appear quite insensitive to the choice of *k*, since the use of values of *k* spanning over almost the entire examined range has given comparable error measurements in most cases. Nevertheless, as *k* varies, the error curves behave slightly differently. As *k* increases, both MAPE and MSPE curves decrease for small values of *k* and, then, slowly increase as to GkNNE and SPE. Specifically, the minimum MAPE and MSPE measures were achieved by setting *k*=300 (for GkNNE) and *k*=550 (for SPE) and the region approximately identified by *k*>250 includes MSPE and MAPE values lower than approximately 4% and 1%, respectively. Such outcomes mean that, choosing *k* in a considerably wide region of the examined interval of variation weakly affects the ultimate estimation outcome, which is similarly very good across the different *k* configurations. On the other hand, for kNNE, MSPE and MAPE values are still decreasing at *k*=1000, and we would probably have obtained slightly better results than 0.50% (MSPE@ *k*=1000) and 3.31% (MAPE@ *k*=1000), respectively, by going on to larger values of k. With regard to the MPE measurements, it is observed that kNNE and SPE exhibit similar curve trends, with values generally lower and in some cases a bit higher than about 10% for most *k* values. In contrast, the employment of the GKNNE provides a slightly poorer MPE values than those obtained by employing both the kNNE and the SPE. This means that, in this case, the GKNNE is not able to model the distribution tails as accurately as the other VKDEs. It is also important to note that, although kNNE does not return actual PDFs, in low-density areas its fit capability is only slightly different with respect to the one offered by the GkNNE and the SPE. In general, it should be noted that quite high error measurements are reported only for *k*=5, which represents a very small fraction ($2 \cdot 10^{-5}$) of the sample size, showing that very small values for *k* are not appropriate.

The outcomes of such investigation, where result variability with respect to *k* was explored, suggest the possibility of identifying a common recommendation for the choice of *k* applicable to various different scenarios. Specifically, the use of *k* equal to $N^{1/2}$ (i.e., 500 in this case of study), as proposed in [54], has proven to yield sufficiently low error measures in most cases. Actually, for *k*=500, SPE, which has been shown to perform better on this simulated scenario, provided values of 6.28%, 3.04%, and 0.46% for MPE, MAPE, and MSPE, respectively.

Besides error measures, the resulting PDF estimates are also graphically illustrated as concerns the suggested $k = N^{1/2} = 500$. Specifically, Fig. 4.9 (a) and (b) show the marginal distributions of $\hat{f}_{\mathbf{X}}(\mathbf{x})$ when the estimation procedures were performed for estimating

samples enclosed in a monotonic grid. The corresponding (one-variable) PDFs are denoted by $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ and were obtained by numerical integration of the joint density $\hat{f}_{\mathbf{X}}(\mathbf{x})$ over the two variables x_2 and x_1, respectively. As we could expect from the study of measurement errors, each variable-bandwidth estimator has returned a PDF estimate in good agreement with the true PDF represented in Fig. 4.9 in dashed black. In particular, the kNNE returned a function accurately following the PDF structure, SPE performed similarly to kNNE, and GkNN a bit worse, though providing an estimate still capable of accommodating both body and tails of the true PDF. Notice that the better kNNE behavior occurs despite the fact that, in contrast with GkNNE and SPE, the kNNE returns PDF estimates that do not integrate to one.

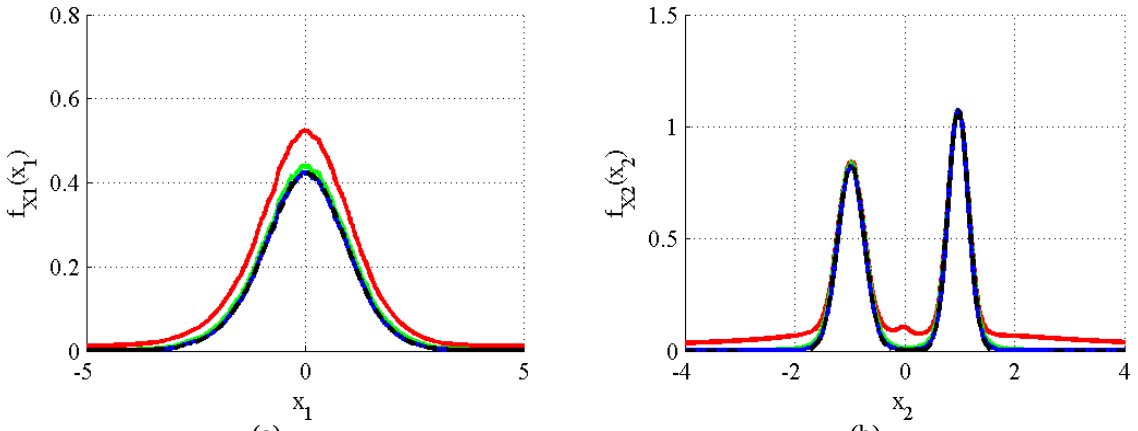


Fig. 4.9. Marginal PDFs obtained through GkNNE, kNNE, SPE. The true marginal PDFs are superimposed in dashed black.

In order to highlight the advantages coming from the employment of data-adaptive variable bandwidths, results obtained are compared to those achieved with the FKDE in (3.12). Again, the kernel function $\kappa(.)$ is taken to be a multivariate Gaussian PDF. Similarly to what done with VKDE, an equal bandwidth h in all dimensions is considered for FKDE. As regard the fixed bandwidth h, several configurations for this parameter were tested. Specifically, the choice of h within the FKDE was done with the aim of exploring values of h spanning the whole range of variable bandwidths tested with VKDE. Thus, h was chosen uniformly sampling 10 values between the lowest (i.e. $9.2 \cdot 10^{-4}$) and the highest (i.e. 2.17) r_k values obtained when k was set equal to 5 and 1000 (which are the extremes of the previously analyzed k range), respectively. The effect of varying h is illustrated in Fig. 4.10 (a), in which plots of the three error measures employed above (MPE, MAPE, and MSPE)

as a function of the parameter *h* in FKDE are shown. As is evident, the FKDE returned in most cases very poor PDF estimates, being characterized by MAPE and MPE measurements around 100% and MSPE values greater than 150 %. Moreover, the VKDE significantly outperformed FKDE even for FKDE results concerning the use of the *h* value yielding the best performance (i.e., h_{best}=0.24) among the examined ones. In fact, FKDE exhibits about as twice as much MAPE and MSPE (whose values for h_{best} are exactly 35.58 and 53.88, respectively), and about 35% more MPE (which is 51.23) than the values obtained, on average, with the VKDEs. Also, none of the error measurements led to any general recommendation as to how the fixed bandwidth *h* should be selected to give the "best" estimate of the unknown PDF.

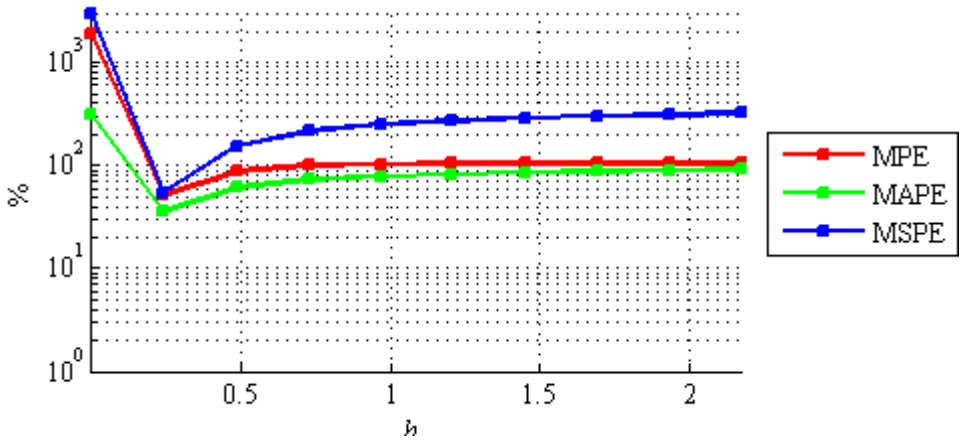


Fig. 4.10. Graph of the three error measures (MPE, MAPE, and MSPE) as a function of the parameter *h* for the FKDE.

In Fig. 4.11 (b) and (c) the marginal distributions of $\hat{f}_{\mathbf{X}}(\mathbf{x})$ for the different values of *h* are plotted. It is clear from the figure that *h* variation has a major impact over FKDE outcome. Moreover, for most values of *h*, FKDE does not respond appropriately to the variations in the magnitude of the PDF being estimated. If *h* is chosen too small, then spurious fine structures become visible since the corresponding estimate exhibits peaks at some data sample locations. On the contrary, if *h* is too large then the bimodal nature of the PDF is obscured due to over-smoothing, mainly occurring in regions where the sample data are more densely packed together. In general, the FKDE outcome reflects an attempt to find some sort of middle ground between what is optimal both for high-density and low-density regions given that the estimation procedure is not data-responsive.

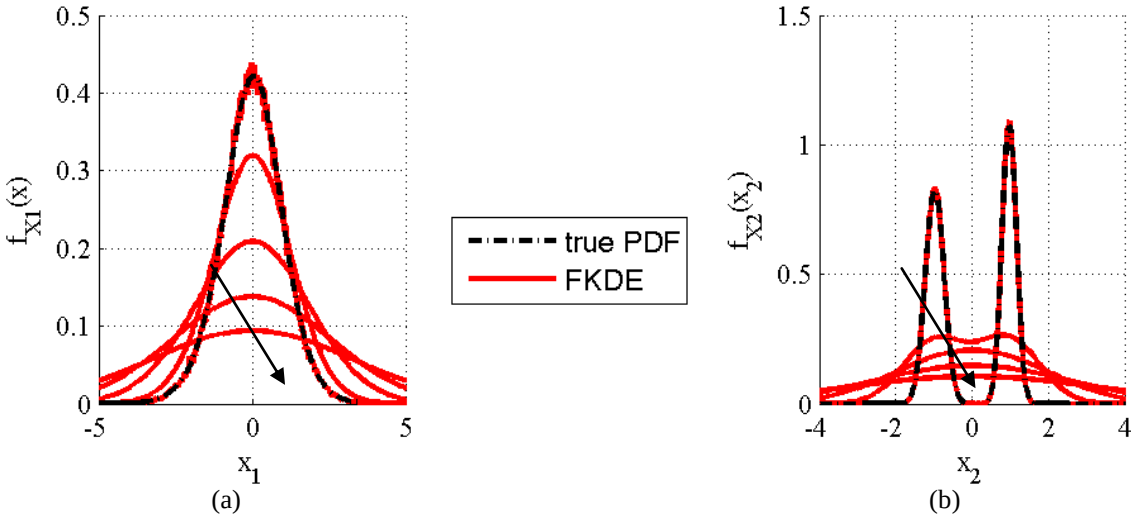


Fig. 4.11. Marginal PDFs (a) $f_{X1}(x_1)$ and (b) $f_{X2}(x_2)$ obtained by numerically integrating the joint PDF $\hat{f}_{\mathbf{X}}(\mathbf{x})$ according to the FKDE at *h* varying (the arrow indicates the direction of increasing *h* values). The true marginal PDFs are superimposed in dashed red.

In summary, this "toy example", though concerning a simple bivariate case, has clearly shown the great advantages provided by the VKDEs employment. First, the specific *k* choice has shown to have a less significant impact over estimation performance: almost all *k* values of the explored range have proven to provide sufficiently low error measures and, thus, a good capability of following the true PDF. On the contrary, the selection of the fixed bandwidth *h* in the FKDE has shown to have a major impact over the estimation performance, with error measures greatly varying within the explored *h* values and, thus, resulting in estimates ranging from under-smoothing conditions up to a heavy over-smoothing of the PDF. In fact, as expected, the parameter *h* plays the role of a smoothing parameter, and we see that there should be a trade-off between sensitivity to noise occurring for small *h* and over-smoothing behavior at large *h*.

Chapter 5

5 Model learning for local AD approaches

Local AD strategies are devoted to locating objects, extending over a few pixels, whose spectral features deviate significantly from those of their surrounding neighbors. To this aim, only a neighboring area of the pixel being tested is used for characterizing its local background.

Within this framework, the conventional AD approach is the popular Reed-Xiaoli (RX) detector. However, such an approach may lead to poor detection performance due to the assumption that the local background is Gaussian and homogeneous. In practice, these assumptions are often violated, especially when the neighborhood of a pixel contains several materials, thus compromising the performance of the algorithm. In the literature, several AD strategies have been presented, most of them trying to cope with the problem of non-Gaussian background. In this thesis work, the use of a locally data-adaptive nonparametric model for estimating the background PDF is proposed within the AD scheme for detecting anomalies by means of the background log-likelihood decision rule.

In this chapter, the new solution has been presented along with the description of existing techniques, thus providing a joint analysis of the different limitations typically met.

5.1 The Reed Xiaoli (RX) algorithm: when data are modeled as a Gaussian non-stationary multivariate random process

In [48], the commonly referred RX algorithm for detecting anomalous objects was established. It is considered the benchmark AD approach for multi-hyperspectral imagery. Basically, parametric models for the data PDFs under the two hypotheses are adopted.

Specifically, the data in the null hypothesis are assumed to arise from a normal distribution. Such a Local Normal Model (LNM) is generally more easily met after application of a local mean-removal procedure using a sliding window. This demeaning window is shown in Fig. 5.1. The demeaning process removes the gross background structures, thus resulting in the following binary hypothesis test:

$$\begin{aligned} \mathbf{X} \mid H_0 &= \mathbf{B} \in N(\mathbf{0}, \mathbf{C}) \\ \mathbf{X} \mid H_1 &= \mathbf{B} + \mathbf{s} \in N(\mathbf{s}, \mathbf{C}) \end{aligned} \tag{5.1}$$

where **s** is the target spectral signature, **B** is the residual background plus noise spectral vector, and **C** is the unknown background covariance matrix, assumed to be the same in the two hypotheses.

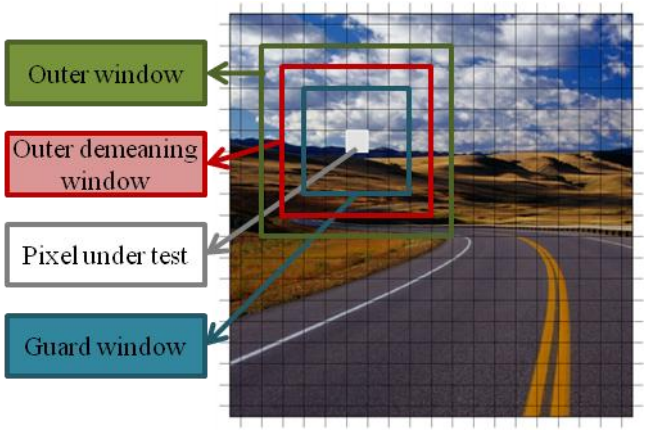


Fig. 5.1. Spatial windows used in the RX implementation: outer demeaning window (red), outer covariance estimation window (green), guard window (blue). The outer window dimension for the demeaning is usually taken to be smaller than the outer window dimension for covariance estimation, since the mean vector is supposed to vary spatially faster than the covariance matrix.

According to the strategy in (2.3), the decision rule for the RX algorithm is the following:

$$\Lambda_{RX}(\mathbf{x}) = \mathbf{x}^t \hat{\mathbf{C}}^{-1} \mathbf{x} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{5.2}$$

where $\hat{\mathbf{C}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_i\mathbf{x}_i^t$, typically referred to as sample covariance matrix, is the ML estimate of **C**, made from the data in a small neighborhood of the pixel under test. This is represented by the dashed green window in Fig. 5.1. It is to be noted that equation (5.2) is simply the square of the Mahalanobis distance between the pixel under test and the local background class is compared to a threshold to detect anomalies. This decision rule can be also derived following the strategy (2.3).

Nevertheless, while the assumption of a multivariate Gaussian distribution is mathematically convenient, LNM has been shown to provide an inadequate representation of the underlying distributions in many environments, leading to poor detection performance [42][55]. This is especially true when the local background contains several materials.

5.2 Kernel - RX: Gaussian model in a high-dimensional feature space

In order to cope with complex local backgrounds, in [32] a nonlinear version of the RX strategy, called kernel RX (and denoted with K-RX hereinafter), was proposed, benefiting from the employment of kernel methods [52]. Specifically, K-RX extends the RX algorithm to a higher-dimensional feature space associated with the original input space via a non-linear mapping function Φ. Within this framework, a LNM is adopted in the higher-dimensional space, which is expected to model a more complex decision boundary in the original input space. The two hypotheses in the feature space are now:

$$\begin{aligned}\Phi(\mathbf{X})\,|\,H_0 &= \mathbf{B}_\Phi \in N(\boldsymbol{\mu}_\Phi, \mathbf{C}_\Phi)\\ \Phi(\mathbf{X})\,|\,H_1 &= \mathbf{B}_\Phi + \Phi(\mathbf{s}) \in N(\boldsymbol{\mu}_\Phi + \Phi(\mathbf{s}), \mathbf{C}_\Phi)\end{aligned} \tag{5.3}$$

In order to maintain the same notation as in [32], the spatial demeaning is not considering. This is obviously taken into account when deriving the decision rule, which introduces the background mean vector $\boldsymbol{\mu}_\Phi$ in the kernel-squared Mahalanobis distance computation. The corresponding RX-algorithm in the feature space is now represented as:

$$\Lambda_{RX_\Phi}\left[\Phi(\mathbf{x})\right] = \left[\Phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_\Phi\right]^t \cdot \hat{\mathbf{C}}_\Phi^{-1} \cdot \left[\Phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_\Phi\right] \tag{5.4}$$

where $\hat{\boldsymbol{\mu}}_{\Phi}$ and $\hat{\mathbf{C}}_{\Phi}$ are the estimated covariance matrix and the mean vector of the background pixel in the feature space, respectively. Nevertheless, the direct implementation of the RX algorithm in the feature space is not feasible, due to the high dimensionality. However, the K-RX decision rule is derived relying upon the kernel theory. In fact, thanks to the "kernel-trick" [52], the K-RX approach implicitly computes the required dot products in the higher-dimensional space by means of kernel functions defined on pairs of input data, without the need of identifying the non-linear mapping. The resulting test statistic is quite complicated, and hence it is not reported, referring to [32] for details. It is worth mentioning that test statistic computation involves the calculation and inversion of large Gram matrices [32][52], which entails a high computational burden.

K-RX has been shown to be equivalent (up to normalizations) to use a FKDE for modeling the background distribution in the original input space [42][32][15]. In FKDE (and, in turn, in K-RX), the smoothness of the approximation and the modeling ability are controlled by scale parameters, which are called bandwidths. Basically, the bandwidths are the kernel function widths. It is well-known that FKDE suffers from the drawback that the bandwidths are assumed constant across the entire feature space [50]. Choosing small bandwidths may lead to PDF estimates exhibiting spurious discontinuities in the tails or in any scarcely populated data region. This effect can be mitigated by increasing the bandwidth values, but at the expense of obscuring structural features characterizing the body of the distribution due to over-smoothing [50]. Moreover, the FKDE outcome has been shown to be very sensitive to even very small variations in the selection of the bandwidth value [54]. These considerations lead to think that similar problems may also affect Kernel-RX behavior, in which the kernel function width parameter plays the role of the FKDE bandwidth.

5.3 A locally adaptive background density estimator: an evolution for RX-based anomaly detectors

Local AD is a topic of great interest in the target detection domain. Within this framework, the conventional AD approach is the popular Reed-Xiaoli (RX) detector. However, such an approach may lead to poor detection performance due to the assumption that the local background is Gaussian and homogeneous. In practice, these assumptions are often violated, especially when the neighborhood of a pixel contains multiple types of

materials, thus compromising the performance of the algorithm. In the literature, several AD strategies have been presented, most of them trying to cope with the problem of non-Gaussian background.

In order to benefit from the great potential that K-RX embeds, such as its ability at modeling complex local backgrounds without making specific distributional assumptions, and – at the same time – trying to overcome problems that are intrinsic to its nature, a novel local AD strategy is here proposed [62]. Specifically, the strategy relies upon the decision rule (2.7) and involves a variable bandwidth kernel density estimator to model the local background. In particular, the GkNNE of equation (4.24) is adopted to better capture the local behavior of the underlying background PDF by allowing the bandwidths to vary over the estimation domain. Such a PDF estimator has proven to adapt the amount of smoothing to the local density of data samples in the feature space, so as to more reliably and accurately follow the multivariate data structure with respect to FKDE [50].

This proposed locally adaptive GKNNE-based AD approach will be denoted with A-RX, hereinafter. The A-RX test statistic is the following:

$$\Lambda_{A-RX}(\mathbf{x}) = -\log\left\{\frac{1}{N \cdot r_k^d(\mathbf{x})}\sum_{i=1}^{N}\kappa\left(\frac{\mathbf{x}-\mathbf{x}_i}{r_k(\mathbf{x})}\right)\right\} \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \eta \tag{5.5}$$

It is important to note that adopting an equal bandwidth across the spectral dimensions means employing spherically symmetric kernel functions. Nevertheless, by pre-scaling the data in order to avoid extreme difference of spread in the various spectral directions, more complicate forms of the kernel functions are not necessary to be adopted since they have been recognized to provide very little improvements [50][54]. To this aim, each input pixel and its surrounding neighboring pixels are linearly transformed to yield data with zero mean and identity covariance matrix prior A-RX application [22][54].

Chapter 6

6 Experimental results: global model learning capabilities

In this chapter, the experiments carried out by applying the proposed global AD strategy are presented and discussed. Two real hyperspectral images characterized by different sizes and background complexity were employed to evaluate the effectiveness of the proposed AD processing chain, applied with the different PDF estimators and learning methods investigated.

It should be noted that the conducted experiments cannot answer the question as to which is the estimator and model-learning combination that better approximates the true image PDFs of the examined data sets, which, of course, are unknown. Rather, the aim of the experimental analysis is to provide further insights into the modeling capabilities of the different methods as well as to evaluate their effectiveness and actual utility in the AD context [63]-[71].

Specifically, this experimental chapter aims at examining three important aspects:

1. *Evaluating and experimentally comparing the ability of both GMM and StMM to represent the statistical behavior of the examined empirical hyperspectral data.*
2. *Evaluating and experimentally comparing the detection performance of the proposed AD strategy when GMM, StMM, and non-parametric estimators, combined with the corresponding learning procedures, are used to estimate the image PDF.*
3. *Evaluating the impact of the user-specified parameters involved in the proposed bandwidth selection strategy on the detection performance when non-parametric density estimators are employed within the proposed AD scheme.*

6.1 Data sets description

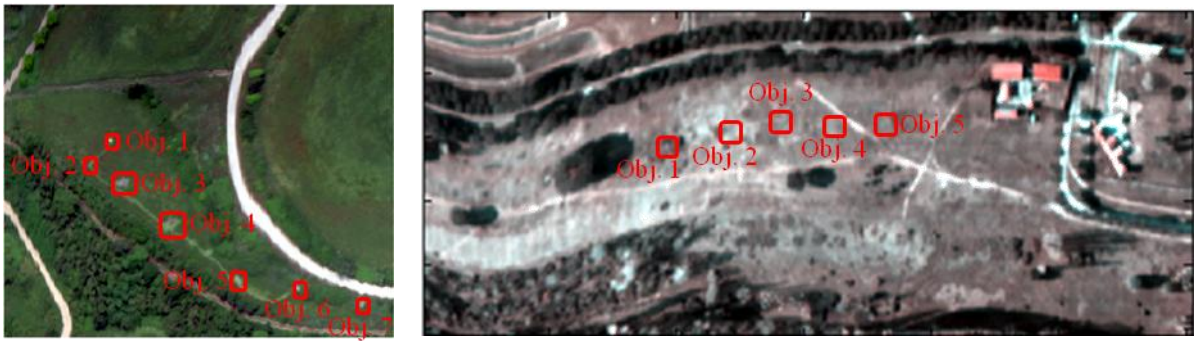
Two hyperspectral images, denoted with *Scene A* and *Scene B*, were analyzed in this research. Both images were collected by the SIM-GA (Sistema Iperspettrale Modulare – Galileo Avionica) hyperspectral sensor, designed and manufactured by Selex-Galileo. The sensor is a push-broom imaging spectrometer operating in the Visible to Near-InfraRed (VNIR) spectral range. The instrument was installed on a micro light aircraft, which has served as experimental remote sensing platform. The main technical characteristics of the sensor are summarized in Table 6. For AD purposes, panels characterized by different sizes and materials were placed within the scenes during the measurement campaigns as targets of interest. Both hyperspectral data were subject to a spectral binning as well as to water-vapor absorption and noisy bands removal. Then, in order to speed-up the computation, a feature reduction method aimed at preserving rare vectors (i.e. anomalies) was used to reduce the dimensionality of the data [2]. Finally, the first principal component was removed in both the data, as it usually addresses the overall scene brightness [27]. The resulting images were processed by the AD scheme discussed above. It should be noted that the feature reduction step also assures more accurate estimates of the mixture parameters and not to incur in dimensionality issues (such as the empty-space phenomenon [54]) during non-parametric estimation.

Table 6. Main technical characteristics of the SIM-GA hyperspectral sensor.

VNIR channel	
Spectral range	400-1000 nm
Spectral sampling	≈1.2 nm
# spectral sampling	500
Focal length	17 nm
Nominal IFOV per pixel	0.7 mrad
Spatial resolution @ 1000 m	0.7 m
FOV	±19°
F#	2.0
Quantization bits	12 bits
Detector	Camera CCD
Maximum frame rate	57 fps
Weight	25 Kg

The first image was collected at a flight height of about 850 m, resulting in an approximated Ground Instantaneous Field of View (GIFOV) of 0.6 m. For the experiments, only a portion of size 365 by 430 pixels of the entire flight line including the targets was considered to be processed by the AD procedure discussed above and will be hereinafter denoted as *Scene A*. The resulting scene mostly includes natural vegetation, soil, and two roads running through almost the entire length of the scene. The scenario also includes panels with different sizes and materials as targets of interest. The deployed objects have sizes ranging from 1 m^2 up to 25 m^2. Such targets take up no more than 0.0471% of the image, a percentage that makes them significantly rare in quantity. Moreover, the target pixels show spectra very similar to several background classes from which they have to be distinguished. Such experimental conditions make the detection of these objects not trivial. A true-color image of *Scene A* is shown in Fig. 6.1(a), with highlighted the target locations.

The second image was acquired by the sensor mounted on board an airplane flying at a height of about 1700 m. The resulting GIFOV was of about 1.2 nm. The image processed, indicated hereinafter as *Scene B*, consists of 255 by 605 pixels around the targets of interest. The scene is characterized by a more complex background structure with respect to *Scene A*. In fact, the scene is largely made up of different kinds of natural vegetation, such as trees and grass, soil, but it also includes several lanes and roads and a small group of houses. *Scene B* includes panels having sizes ranging from 1 m^2 up to 16 m^2 that both are rare in quantity (they occupies 0.0136% of the image) and have spectra very similar to those of background pixels. These conditions, together with the more complex background structure, make the detection of the deployed targets challenging. *Scene B*, together with the locations of the targets in the scene, is shown in Fig. 6.1(b).



(a) (b)

Fig. 6.1. True-color representation of the scenes and location of the targets: (a) *Scene A*, (b) *Scene B*.

In this experimental chapter, *Scene A* is employed as the primary dataset, that is it is adopted to thoroughly discuss the conducted experiments. Next, *Scene B* is used as a benchmarking data onto which validate the results obtained.

6.2 On the statistics of hyperspectral imaging data: GMM and StMM modeling capabilities

As mentioned, although the GMM is undoubtedly one of the most widely adopted models for modeling hyperspectral data, the StMM has been suggested to better describe the distribution tails of background classes [64][65]. One of the purposes of this experimental analysis is to investigate and compare the ability of GMM and StMM, learnt with the Bayesian strategies described above and denoted with BGMMS and BStMM, respectively, to represent the statistical behavior of real hyperspectral data. It is expected that StMM provides better modeling capabilities thanks to its mixture components that accommodate longer tails than the Gaussian ones. Other studies have investigated the difference model ability of GMM and StMM for hyperspectral data [1][38][39][40]. However, whereas in those studies the number of mixture components were specified in advance (probably by visual inspection of the spectral diversity of the scene), here the Bayesian model learning has been conducted automatically and without operator intervention.

In order to investigate the GMM and StMM modeling capabilities, the examined image was first segmented into clusters according to such models. In practice, modeling the PDF of the data with FMMs means to assume that each pixel originates from one component of the mixture according to some probability. Therefore, cluster maps were constructed assigning each pixel to the component that has most likely generated it. Once the data were segmented into clusters, the empirical distributions of the pixels within each cluster were analyzed. Such analysis was conducted by computing, for each cluster, the *probability of exceedance* of the Mahalanobis distances between each pixel of the cluster and the cluster itself. Specifically, such a probability of exceedance represents the probability of the Mahalanobis distance exceeding a given threshold.

In general, the Mahalanobis distance M_g of multivariate Gaussian data characterized by mean vector $\boldsymbol{\mu}$ and precision matrix $\mathbf{T}$ is defined as

$$M_g = (\mathbf{x}-\boldsymbol{\mu})^t \mathbf{T} (\mathbf{x}-\boldsymbol{\mu}) \tag{6.1}$$

and follows the Chi-square distribution χ^2 with d degrees of freedom [40]. On the other hand, the Mahalanobis distance M_t of multivariate Student's t distributed data with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Lambda}$, and v degrees of freedom, as already mentioned in [38], is defined as

$$M_t = \frac{(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Lambda} (\mathbf{x}-\boldsymbol{\mu})}{dv} \tag{6.2}$$

and follows the F distribution $F_{v,d}$ with v and d degrees of freedom [40].

On the basis of equations (6.1) and (6.2), the empirical Cumulative Distribution Functions (CDFs) of M_g (for GMM) and M_t (for StMM) were computed for each cluster by employing the pixels belonging to the corresponding mixture component as well as the model parameters estimated for that component (i.e. mean vector and precision matrix for the GMM, and mean vector, scale matrix, and number of degrees of freedom for the StMM). Given the high number of sample data used for estimation, the errors in the estimates of model parameters are assumed to be negligible. By computing the complementary empirical CDFs of M_g and M_t for each cluster, the probabilities of exceedance were obtained. Next, *exceedance plots* were constructed by comparing the empirically evaluated complementary CDFs to the ones that should be obtained theoretically. In this way, such exceedance plots for the Mahalanobis distance statistics show how well each assumed model fits the empirical data distribution.

The Bayesian GMM learning approach, employed on *Scene A*, provided 2 GMM components, and hence a cluster map with 2 clusters, which is shown in Fig. 6.2 (a). Fig. 6.2 (b) depicts the corresponding exceedance plots computed over each cluster of the GMM cluster map. As is evident, results indicate that the empirical distribution of the pixels within each cluster has longer tails than the theoretically expected ones. This means that the data, at least as regards the distribution tails, are not accurately modeled by the multivariate GMM PDF estimated through the learning procedure. In such a case, in order to obtain a more accurate fit, the operator intervention would be necessary to detect the mis-modeling and guide possible subsequent further and more-refined learning procedures. For the sake of comparison, the GMM exceedance plots in Fig. 6.2 (b) also report the theoretical complementary CDFs of the F distributions $F_{v,d}$ obtained with the values of the

v that best fit the empirical curves obtained. Such theoretical curves, obtained with v={4, 10}, show that the empirical exceedance plots significantly resemble those of Student's t distributed data and suggest that the StMM is expected to provide a better fit.

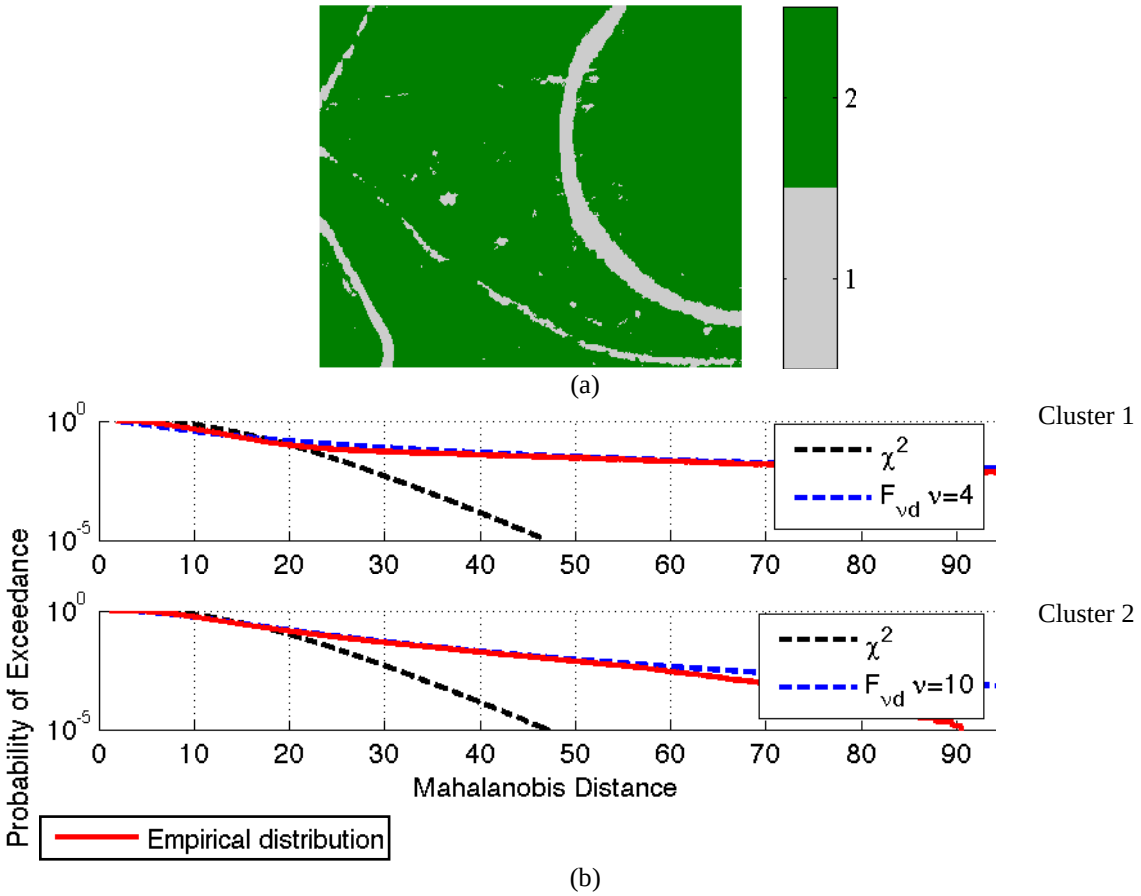


Fig. 6.2. (a) Cluster map and (b) exceedance plots of the Mahalanobis distances for the spectral classes produced by GMM learning strategy. Specifically, the empirical distributions are plotted in red (solid curves), whereas the black and blue (dashed) curves represent the χ^2 and the F distributions considered for comparison to the empirical distributions, respectively. This latter refers to the F distributions $F_{v,d}$ obtained with the values of the v that best fit the empirical curves obtained.

The Bayesian StMM learning strategy applied to the same scenario provided 4 StMM components and hence a cluster map with 4 clusters, which is shown in Fig. 6.3 (a). The corresponding exceedance plots are depicted in Fig. 6.3 (b). In the same plots, the $F_{v,d}$ distributions characterized by the numbers of degrees of freedom returned by the Bayesian learning algorithm are reported for comparison. Not only the empirical StMM exceedance plots exhibit indeed heavy tails, but – on the contrary to the GMM case - they are also in good agreement with the corresponding theoretical F distributions.

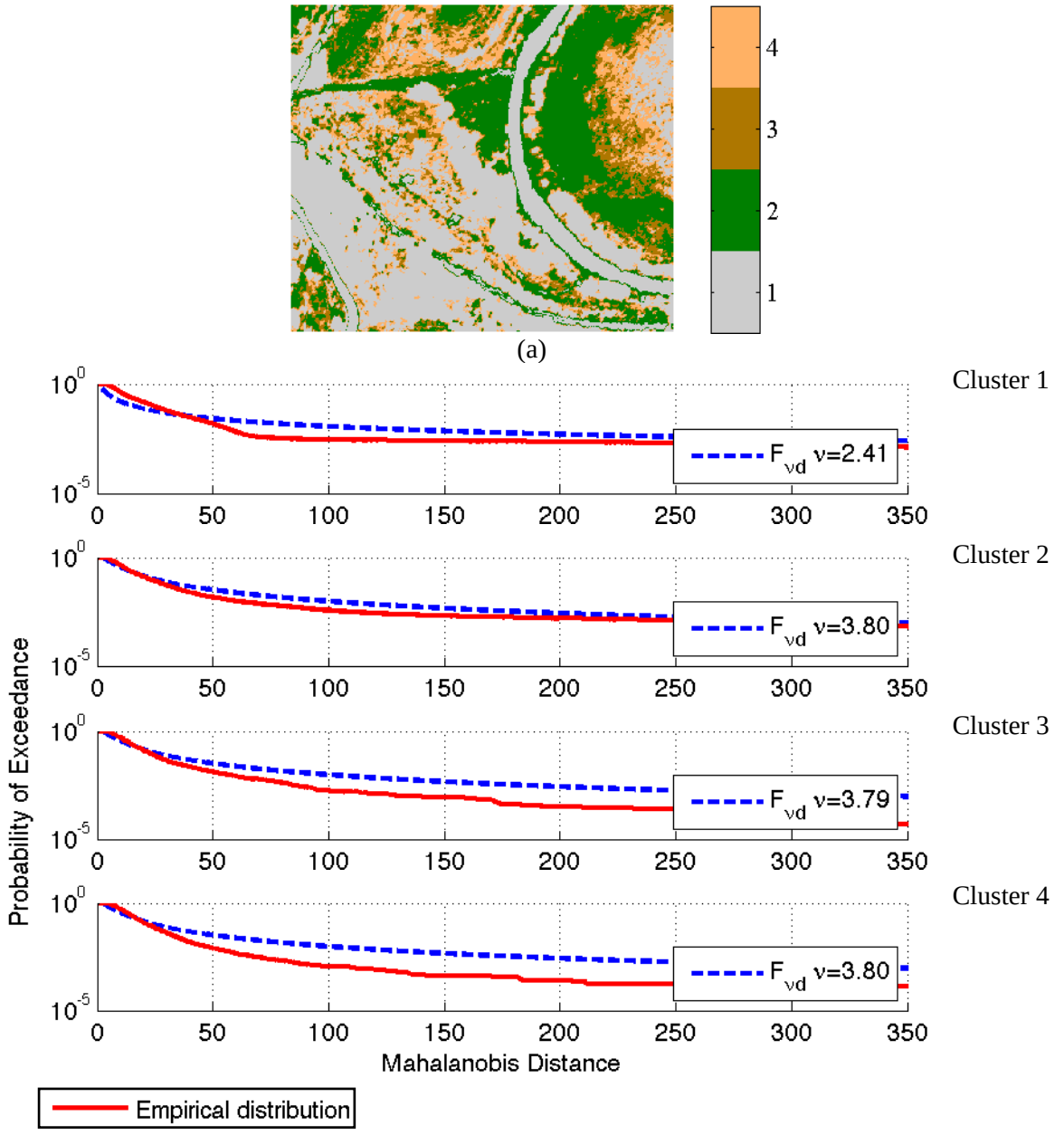


Fig. 6.3. (a) Cluster map and (b) exceedance plots of the Mahalanobis distances for the spectral classes obtained by the StMM learning strategy. Specifically, the empirical distributions are plotted in red (solid curves), whereas the blue (dashed) curves represent the F distributions considered for comparison to the empirical distributions, respectively. This latter refers to the F distributions $F_{\nu,d}$ characterized by the numbers of degrees of freedom returned by the learning algorithm.

6.3 FKDE strategy automation

In the experiments of non-parametric background PDF estimation through the FKDE approach, the classical Gaussian kernel function was used, as commonly done in the literature [3][54]. As previously observed, assessing reliable bandwidths in the FKDE estimator is very important, particularly in the proposed AD scheme where we need to

estimate the PDF with extreme precision. The crucial choice of suitable values for the bandwidths was approached by resorting to the BN Bayesian methodology described in section 4.2.1 [7]. As anticipated, since such an approach assumes an equal bandwidth *h* for each component (by restricting the contours of the kernel functions to be spherically symmetric), the data were normalized so that each spectral component had the same variance.

The goal of FKDE experimental analysis was to evaluate the impact of the choice of K_l and K_u onto the bandwidth selection process and, in turn, on the detection performance [68][69]. A minimal effect of such a choice on the detection performance would suggest that an automatic application of the strategy is not impaired. Specifically, experiments were carried out with respect to different choices for the bounds K_l and K_u and the corresponding effect onto both the background log-likelihood and the detection performance was assessed. To this aim, as in [7], K_l and K_u were expressed as fractions of the number of the available data samples *N*. In particular, K_u was evaluated as $K_l+\Delta$, and a (N/K_l, Δ) space was generated to investigate the effect of the parameter choice. Specifically, N/K_l was varied between 490.5 and 15695, corresponding to K_l ranging in [10, 320], whereas Δ was varied between 0 and 150.

In order to perform a quantitative analysis, two performance measures were considered. The former is aimed at quantifying how much the target pixels emerge from the image background ones in the test statistic (i.e. the background log-likelihood $\Lambda(\mathbf{x})$). Specifically, the Signal to Noise Ratio of target pixels over background pixels was computed over the test statistic. Such evaluation measure is denoted with SNR_Λ and defined as:

$$SNR_\Lambda = \frac{\mathrm{E}\{\Lambda(\mathbf{x})|\,\mathrm{H}_1\} - \mathrm{E}\{\Lambda(\mathbf{x})|\,\mathrm{H}_0\}}{\mathrm{std}\{\Lambda(\mathbf{x})|\,\mathrm{H}_0\}} \tag{6.3}$$

where std{.} indicates the standard deviation. The higher SNR_Λ value is, the better the targets emerge from the background and are more easily detectable. The second performance measure employed is the False Alarm Rate (FAR) corresponding to the maximum threshold value in the detection test statistic at which all target pixels are detected (hereinafter denoted with *Global FAR@100% detection*). Of course, the lower *Global FAR@100% detection* is, the better the detection performance, since less false alarms are required to detect all target pixels.

The SNR_Λ values computed over the *Scene A* on the basis of the available ground truth target map are shown in Fig. 6.4 (a) As is evident, a region can be identified that includes SNR_Λ values not lower than approximately 3dB with respect to its maximum value. This means that all $[K_l, K_u]$ configurations chosen in such a region provide similar enhancement of target pixels in the test statistic. Specifically, on the examined scenario, this region is approximately identified by $N/K_l<600$, that means that choosing a number K_l of nearest neighbors not lower than 2 orders of magnitude with respect to the total number N of pixels is sufficient to exhibit a very good background suppression ability.

Fig. 6.4 (b) displays the *Global FAR@100% detection* values obtained. As is evident, all configurations manage to detect all target pixels with very low, and in most cases equal to 0, FARs. In particular, the region identified above for SNR_Λ mostly corresponds to the $[K_l, K_u]$ configurations yielding the best *Global FAR@100% detection*, i.e. a perfect detection of all target pixels with no false alarms.

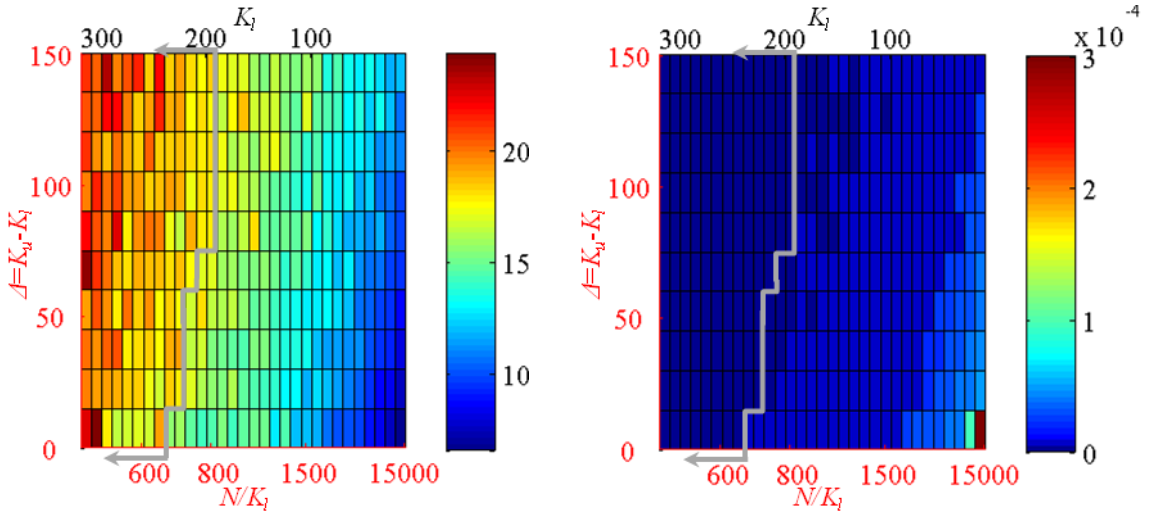


Fig. 6.4. (a) SNR_Λ and (b) *Global FAR@100% detection* measures for different configurations of the interval $[K_l, K_u=K_l+\Delta]$. The red arrow depicts the region including SNR_Λ values not lower than approximately 3dB with respect to its maximum value.

Such outcomes mean that, for the examined scenario, choosing $[K_l, K_u]$ in a considerably wide region of the $(N/K_l, \Delta)$ space examined weakly affects the ultimate detection performance, which is similarly very good across the different configurations. Such a weak effect of the $[K_l, K_u]$ choice can be confirmed by examining Fig. 6.5, which shows the range of bandwidths h obtained. In particular, the attained h values in the highlighted region range in [3.56, 4.33], with an average value of 3.80 and a standard

deviation of 0.17. Such values show a very limited variation of h with respect to significant variations of the $[K_l, K_u]$ in the $(N/K_l, \Delta)$ space.

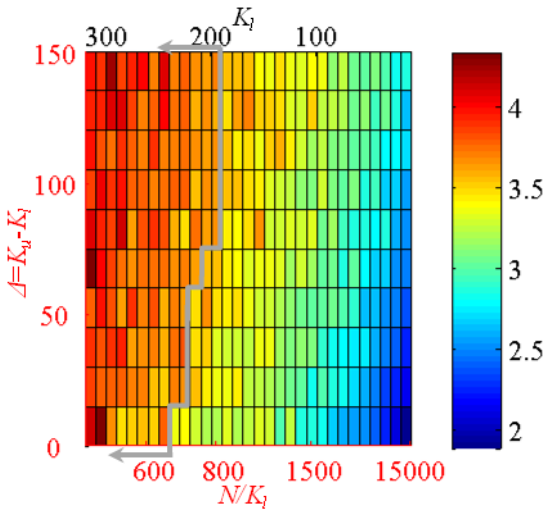


Fig. 6.5. Estimated bandwidths for different configurations of the interval $[K_l, K_u=K_l+\Delta]$. The red arrow depicts the region including SNR_Λ values not lower than approximately 3dB with respect to its maximum value.

The outcomes of such investigation suggest the possibility of identifying a common recommendation for the choice of $[K_l, K_u]$ applicable to various different scenarios. This is appealing since it will allow for operation without "man-in-the-loop", while assuring sufficiently good performance. Further investigations have been being performed by examining the *Scene B*. Outcomes from *Scene B* analysis are reported in the validation section 6.6

6.4 Anomaly detection performance: the Bayesian learning for global background modeling in hyperspectral images

Anomaly detection performance of the proposed global AD strategy with respect to the different image PDF estimation methodologies is here analyzed for *Scene A* on the basis of the ground truth target map. Specifically, the proposed global AD scheme based on equation (2.7) is combined with reliable and automatic data-driven background PDF estimation. Actually, the use of such estimators is coupled only with the employment of the automatic model learning methods developed within a Bayesian framework. Specifically, the BGMMS, BStMM, and BN approaches described in chapter 4 were considered in this

analysis [64][66]-[69][71]. Moreover, the Gaussian kernel function is employed in the FKDE.

It is important to note that the experiments here discussed are not aimed at quantitatively evaluating the ability of the examined PDF estimators to approximate the image background *true* PDF. As a matter of fact, the true image PDF is unknown and the various error measures employed in section 4.2.3 cannot be computed.

Overall detection performance is evaluated by means of the Receiver Operating Characteristics (ROC) curves [42]. ROC curves plot the Fraction of Detected Target pixels (FoDT) versus the FAR, computed by increasing the threshold level from zero to the maximum detection statistic value over the operating scenario analyzed. Specifically, pixel-based ROC curves are reported, i.e., FoDT is computed as the ratio of the number of target pixels properly detected to the total number of target pixels. As the detection threshold is raised, fewer and fewer pixels are classified as anomalies. Thus, a higher threshold leads to a lower FoDT. Nonetheless, decreasing the threshold means that more and more non-anomalous pixels are mistakenly classified as anomalies and, thus, the FAR increase. As a result, plotting FoDT vs FAR at each threshold value builds a curve that summarizes the trade-off for obtaining a high FoDT with a reasonably low FAR.

Then, since evaluation of anomaly detection performance is not a trivial task, several object-wise performance measures are taken into account so as to evaluate algorithm behavior with respect to the different targets deployed in the scene [42]:

- *FAR@1st detection*. The FAR at the first detection provides the FAR for just detecting the presence of the desired target, which is associated with its pixel exhibiting the highest test statistic value.

- *FAR@100% detection*. It is an object-wise version of the *Global FAR@100% detection* measure previously described. This measure assesses the FAR arising from the detection of all pixels of a given target object.

- $TSNR_{\Lambda}$. In order to provide a measure assessing how much, in the AD test statistic $\Lambda(\mathbf{x})$, each target object emerges with respect to the background pixels, a SNR_{Λ} for each target object can be computed as follows:

$$TSNR_{\Lambda} = \frac{\max_{\mathbf{x}\in \text{Obj.}} \{\Lambda(\mathbf{x})|\,\mathrm{H}_1\} - \mathrm{E}\{\Lambda(\mathbf{x})|\,\mathrm{H}_0\}}{\mathrm{std}\{\Lambda(\mathbf{x})|\,\mathrm{H}_0\}} \tag{6.4}$$

The higher this measure is, the better the performance. The assumption for using (6.4) is that it is expected that, after detection, $\Lambda(\mathbf{x})|H_0$ (background) values are tightly concentrated around their mean value whereas $\Lambda(\mathbf{x})|H_1$ (target) exhibits much higher values.

As to the experiments of non-parametric background PDF estimation through the FKDE, since such an approach assumes an equal bandwidth *h* for each component, the data were previously normalized so that each spectral component had the same variance, as done in the previous section.

Fig. 6.6 shows the ROC curves obtained by thresholding the detection test statistics. Such curves show that, on these data, the best overall detection performance is achieved by both the StMM-based and the FKDE-based AD strategy, which provided ROC curves characterized by FoDT=1 and FAR=0 for any value of the detection threshold (and, thus, do not appear in the plot). Specifically, all FKDE configurations within the region including SNR_{Λ} values not lower than approximately 3dB with respect to its maximum value yielded equal ROC curves. As regards the GMM-based AD approach, it provided a ROC curve with lower FoDTs for similar values of FAR. Still, 80% of target pixels are detected with FAR= $3.2\cdot10^{-5}$, 90% of them with FAR=$5.7\cdot10^{-5}$, and detection of all target pixels in the scene corresponds to FAR=$2.5\cdot10^{-4}$.

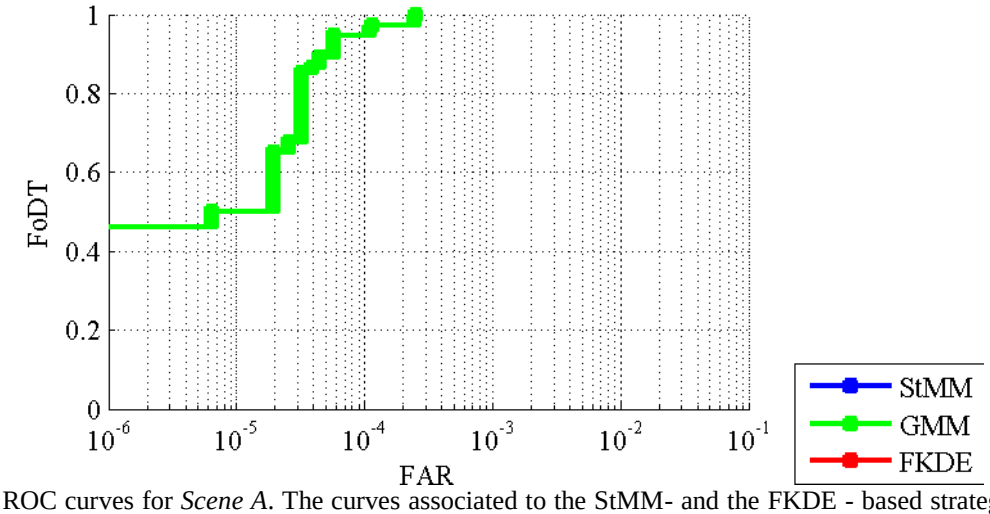


Fig. 6.6. ROC curves for *Scene A*. The curves associated to the StMM- and the FKDE - based strategies do not appear in the plot because they are characterized by FoDT=1 and FAR=0 for any value of the detection threshold.

As to algorithm behavior over each target object, Table 7 and Table 8 report measures of the *FAR@1st detection* and the *FAR@100% detection*, respectively, for the proposed AD strategy employing the GMM, StMM and FKDE PDF estimators. Again, measures related to the use of FKDE as background PDF estimator refer to the [K_l, K_u] configurations falling in the region highlighted previously. These measures confirm the best results obtained by the StMM- and the FKDE -based AD schemes. In fact, both yielded a perfect localization of each target (no false alarms), as is evident from the *FAR@1st detection* measurements. Furthermore, they both succeeded in detecting, with no false alarms, all the pixels within each target object (i.e. *FAR@100% detection*=0 for all objects). As regards the GMM-based AD scheme, very good, though not perfect, object-wise detection is achieved, with all *FAR@1st detection*=0 with the exception of Obj. 3, whose detection makes one false alarm arise. Nevertheless, when detection of all target pixels of each object is concerned, GMM performance exhibits non-null *FAR@100% detection* for three out of seven objects (Obj. 3, 5, and 6), being Obj. 3 the one causing the highest *FAR@100% detection* value (i.e. $2.5 \cdot 10^{-4}$).

Table 7. Measures of *FAR@1st detection* (*Scene A*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7
GMM	**0**	**0**	$6.39 \cdot 10^{-6}$	**0**	**0**	**0**	**0**
StMM	**0**	**0**	**0**	**0**	**0**	**0**	**0**
FKDE	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Table 8. Measures of *FAR@100% detection* (*Scene A*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7
GMM	**0**	**0**	$2.5 \cdot 10^{-4}$	**0**	$1.9 \cdot 10^{-5}$	$1.9 \cdot 10^{-5}$	**0**
StMM	**0**	**0**	**0**	**0**	**0**	**0**	**0**
FKDE	**0**	**0**	**0**-6.4 10^{-6}	**0**	**0**	**0**	**0**

Table 9. Measures of $TSNR_\Lambda$ (*Scene A*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7
GMM	26.82	14.23	12.32	30.16	14.64	16.16	20.89
StMM	5.76	5.22	5.98	6.54	5.87	5.99	6.27
FKDE	22.78 **– 32.00**	19.41 **– 24.24**	16.30 – **23.05**	22.46 **– 31.78**	17.78 **– 25.12**	18.20 **– 25.92**	19.57 **– 27.65**

As regards the background suppression ability, the $TSNR_{\Lambda}$ values computed for each target and with respect to the employment of the different AD schemes are included in Table 9, where, again, the FKDE results reported are the maximum and the minimum values obtained in the region including SNR_{Λ} values not lower than approximately 3dB with respect to its maximum value. As is evident by comparing Table 9 with Table 7 and Table 8, $TSNR_{\Lambda}$ values for StMM apparently seem to be in contrast with the null FAR values. In fact, whereas the StMM-based approach has been shown to perfectly detect all objects and all pixels within each object, the target pixels in its detection test statistic seem not to emerge well with respect to background pixels. $TSNR_{\Lambda}$ values for StMM are actually much lower than those yielded by both GMM and FKDE schemes. More insights into this seemingly unusual behavior can be obtained by examining the AD detection test statistics corresponding to the examined schemes, shown in Fig. 6.7 (a-d). By visual inspection of such detection test statistics, it is clear that the image background structures emerge much more in the StMM case with respect to both the GMM and the FKDE ones.

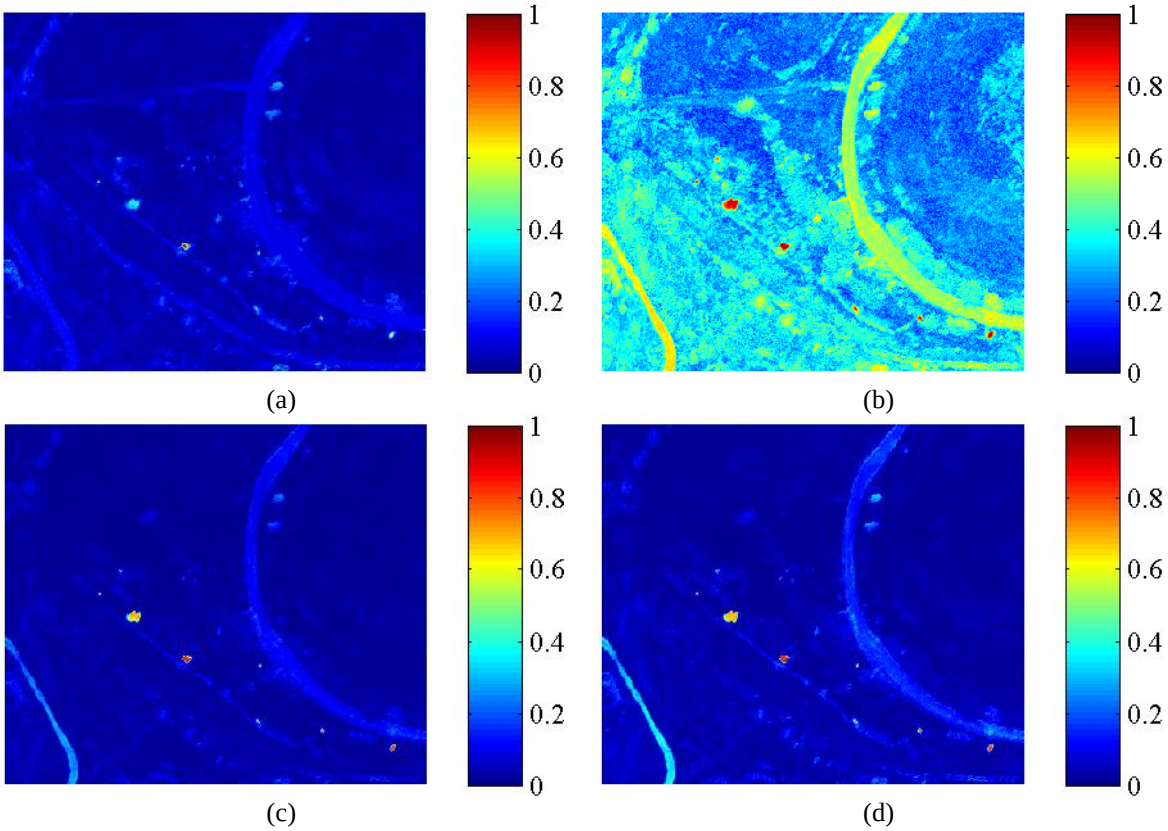


Fig. 6.7. Normalized detection test statistics obtained by using (a) GMM-, (b) StMM- and (c) (d) FKDE -based AD strategy. In the FKDE case, the configuration yielding (c) the highest SNR_{Λ} and (d) the SNR_{Λ} lower than approximately 3dB with respect to the maximum value are taken into account. The statistics have been normalized so that their values range in [0,1].

Such a phenomenon is much more evident by examining Fig. 6.8, where the histograms of the detection statistics associated to target pixels (in red) and background pixels (in blue) are reported for each AD scheme.

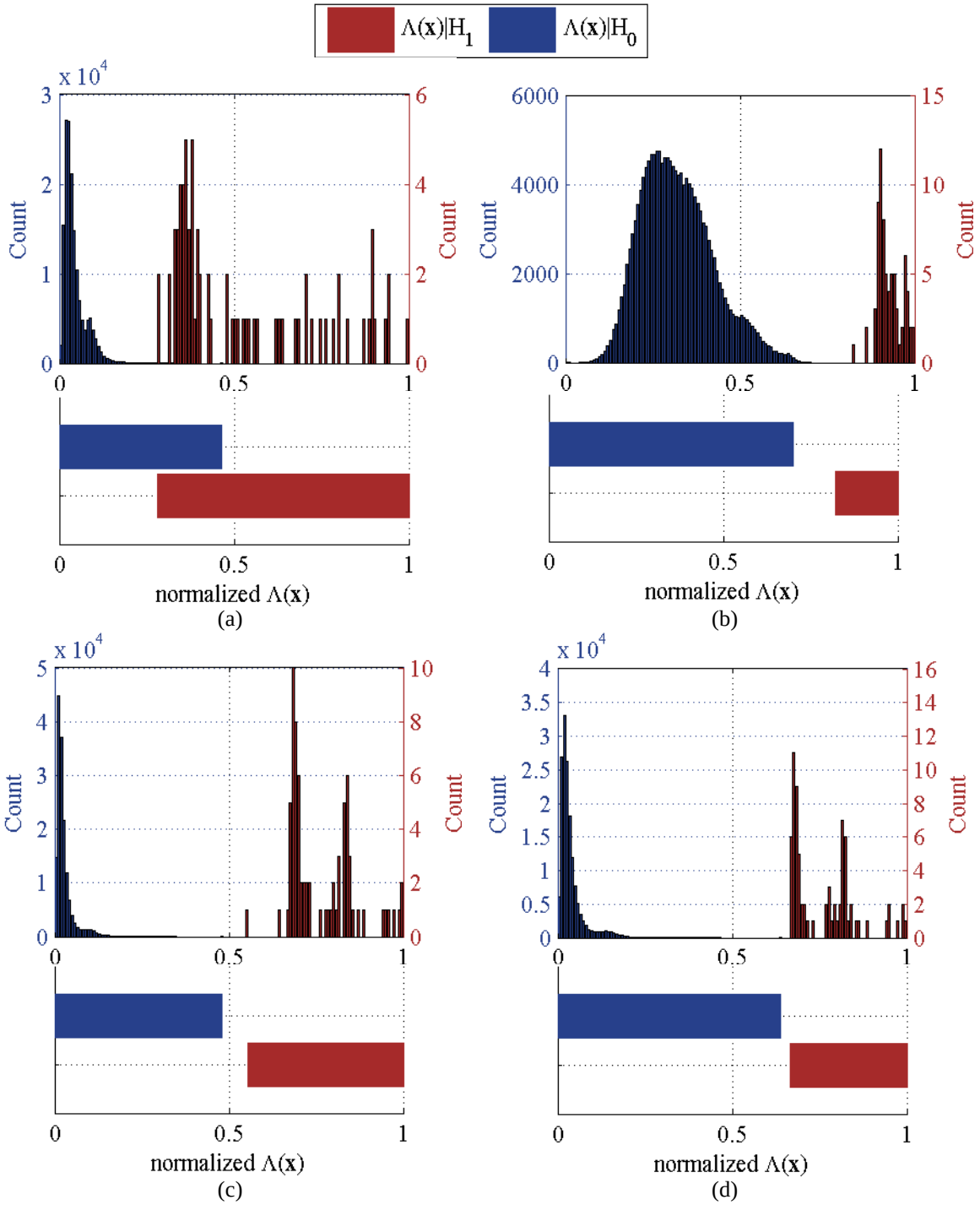


Fig. 6.8. Histograms of the detection test statistics associated to target pixels (in red) and background pixels (in blue) obtained by applying the proposed AD strategy employing (a) GMM, (b) StMM, and (c) (d) FKDE corresponding to the configurations yielding (c) the highest SNR_{Λ} and (d) the SNR_{Λ} lower than approximately 3dB with respect to the maximum value are taken into account. Below each histogram, the interval of variation of target and background test statistic values is represented by means of horizontal bars.

Below each histogram, horizontal bars show the ranges of variation of target and background test statistic values, highlighting the separation of target and background pixels after each AD scheme is applied. As is evident, the histograms computed over the GMM- and the FKDE-based test statistic show $\Lambda(\mathbf{x})|H_0$ values (associated to background pixels) to be almost very concentrated around their mean value (which takes value equal to $2.81 \cdot 10^{-2}$, $3.92 \cdot 10^{-2}$ and $4.34 \cdot 10^{-2}$ for the three analyzed schemes). In fact, their standard deviations range from $3.04 \cdot 10^{-2}$ to $3.17 \cdot 10^{-2}$ and $4.22 \cdot 10^{-2}$. Despite this narrow concentration of $\Lambda(\mathbf{x})|H_0$ values around the mean, the horizontal bars of the GMM-based test statistic clearly show that a given number of target pixels ($\Lambda(\mathbf{x})|H_1$) assumes values comparable to and lower than those of some background pixels, these latter being hence associated to false alarms. This does not occur for the FKDE-based test statistic. As regards the StMM-based histogram, background pixels do not exhibit a similar concentration around their mean value in the test statistic. Rather, $\Lambda(\mathbf{x})|H_0$ values are much more dispersed (with a standard deviation equal to $1.04 \cdot 10^{-1}$) around their mean value, which is also higher ($3.20 \cdot 10^{-1}$) than the one taken in the $\Lambda(\mathbf{x})|\mathrm{H}_0$ case. This widespread character is mainly due to the background structures (mostly the roads) already observed emerging in Fig. 6.7, which are responsible for the higher $\mathrm{E}\{\Lambda(\mathbf{x})|H_0\}$ and $\mathrm{std}\{\Lambda(\mathbf{x})|H_0\}$ that lead to lower $TSNR_\Lambda$ values. Such a behavior makes it clear that $TSNR_\Lambda$ cannot be used as a direct measure of the detection performance and should be coupled with an analysis of background pixels distribution in the test statistic. In fact, horizontal bars for the StMM case clearly show a good separation between target and background pixels in the test statistic, with no overlap of the bars, which means no false alarms. Indeed, by looking at Fig. 6.7, the StMM-based detection test statistic clearly reveals the whole shapes of the objects emerging from the background, here much more evidently than with the other AD schemes. Although in the GMM and FKDE cases most of background structures have been annihilated, the whole target shapes are not as well evident as in the StMM case, especially for the GMM test statistic.

Finally, it is worth noting that the lowest $TSNR_\Lambda$ obtained in the GMM case is that of Obj. 3, which is also the object causing higher *FAR@1st detection* and *FAR@100% detection* values. Obj. 5 and 6, which also yield non-null *FAR@100% detection*, are characterized by low $TSNR_\Lambda$ values as well.

6.5 Exploring the use of variable and fixed bandwidth kernel density estimators for AD purposes

The AD processing chain described in chapter 2.4.1 has been proposed to efficiently explore hyperspectral images for the detection of anomalous objects with the help of reliable image PDF estimation. The goal of the experiments conducted in this section was to provide insights about the background modeling ability of VKDEs, as well as their effectiveness and actual usefulness in the AD context with respect to the FKDE [63][70]. As highlighted before, the use of data adaptive non-parametric background PDF estimators arises from the difficulty of FKDE in following the local structural peculiarities of PDFs in the feature space, mainly due to the inability of a fixed bandwidth to adequately handle both PDF body and tails. With this in mind, design of experiments in this section was performed with the goal of showing the effectiveness of the proposed AD strategy in detecting the anomalous objects in the real hyperspectral image described above while assuring a good capability of following the actual structure of data in the feature space, so that no under-smoothing or over-smoothing phenomena occur.

In order to evaluate the impact of the only VKDE user-specified parameter k on the detection performance, the experiments were conducted with respect to different choices of k. Specifically, k was varied between 5 and 1000, corresponding to k/N ranging in $[3.2\cdot10^{-5}$ $6.4\cdot10^{-3}]$. For each selected k value, the image background PDF was estimated by using the adaptive techniques of equations (4.24), (4.25), and (4.26). In this analysis, the classical Gaussian kernel was used in both GkNNE and SPE, as done in the previous sections and in the "toy-example". Then, the detection of anomalous objects within the image was conducted according to the criterion in (2.7). Special attention was devoted to the behavior of the proposed AD scheme for k values both corresponding to the two extremes of the analyzed interval (i.e., $k \in \{5, 1000\}$) and equal to the suggested one (i.e., $k=N^{1/2}=397$). The choice of this latter k value was suggested in [54], but its validity seems to be confirmed by the analysis conducted in the "toy-example".

Such results were compared to those obtained by using FKDE for non-parametric PDF estimation within the AD strategy in equation (3.12). Again, a Gaussian kernel was employed in the FKDE and an equal bandwidth h in all dimensions was considered, i.e. $\mathbf{H}=h\mathbf{I}_d$, as commonly performed in the literature and similarly to what done with VKDEs. In order to provide a fair comparison, several possibilities for the parameter h were

explored. Specifically, the choice of h within the FKDE was done uniformly sampling the range between the lowest r_k value obtained with the minimum k=5 and the highest r_k value attained for the maximum k=1000, thus exploring h values spanning the whole range of the variable bandwidths being tested.

Since the approaches to be tested assume spherically symmetric kernel functions, the data were normalized in order to avoid extreme difference of spread in the spectral directions so that each spectral component has the same variance.

Performance evaluation over the real image was carried out by analyzing both the estimator capability of following the structure of the given data and as concerns the detection of the small anomalous objects in the scene. To this aim, results are examined by using the available ground truth target map and by means of:

- Quantitative statistical analysis of the VKDE variable bandwidths obtained with the k-NN approach in correspondence of both target and background pixels for the different values of k. It should be noted that the k-NN variable bandwidths are hereinafter indicated simply as r_k (rather than specifying either $r_k(\mathbf{x})$ or $r_k(\mathbf{x}_n)$, as in equations (4.24), (4.25), and (4.26)) since the background PDF value in each tested pixel $\mathbf{x}$ is estimated by employing, as data samples, all the remaining image pixels $\{\mathbf{x}_n \in \Re^d | \ n = 1, 2, ..., N, \mathbf{x}_n \neq \mathbf{x}\}$. Hence, each image pixel is used, in turn, both as estimation pixel $\mathbf{x}$ (where $r_k(\mathbf{x})$ is computed for equations (4.24) and (4.25)) and as observed sample pixel $\mathbf{x}_n$ (where $r_k(\mathbf{x}_n)$ is computed for equation (4.26)), and the corresponding $r_k(\mathbf{x})$ and $r_k(\mathbf{x}_n)$ values actually coincide for a same k.
- Quantitative analysis of the range of test statistic values assumed in correspondence of both target ($\Lambda(\mathbf{x})|H_1$ i.e., mainly on the tails) and background ($\Lambda(\mathbf{x})|H_0$ i.e., over the main body) pixels, performed for different values of k, in conjunction with a visual inspection of detection test statistic (minus logarithm of the estimated PDF) for the three k values mentioned above.
- Evaluation of the distance between the minimum value of the test statistic for the target pixels and the maximum value of the test statistic for the background ones:

$$\delta = \min\left[\Lambda(\mathbf{x}) | H_1\right] - \max\left[\Lambda(\mathbf{x}) | H_0\right] \tag{6.5}$$

Such a measure quantifies the separation between target and background test statistic values. If there is not a perfect separation between target and background test statistics, δ takes negative values.

- Evaluation of *Global FAR@100% detection* measure, which represent the False Alarm Rate (FAR) corresponding to the maximum value of the threshold in the test statistics that allows all target pixels to be detected (100% of detection rate), as already previously described.

The values of k employed led to a wide range of variability for the r_k values obtained for all image pixels, ranging from a minimum value of $r_{k,min}$= 0.61 (obtained for k=5) and a maximum value of $r_{k,max}$= 26.36 (resulting from k=1000). These two specific values represented also the minimum and maximum values for the range of h to be employed within the FKDE approach.

To better analyze how r_k variability translates into an effective ability to adjust the different smoothing requirements of the different structure of the PDF, Fig. 6.9 shows the mean r_k value, computed over both target and background pixels, as a function of the parameter k, with confidence intervals given by its standard deviation. As we expected, target and background r_k mean values are broadly separated for all k configurations and, more importantly, r_k values are typically much higher for the target pixels than the background ones. Such behavior reflects what mentioned, since flatter kernel functions are associated with the target pixels, which we expect to lie in the tails of the PDF, whereas narrower kernel functions are used to model the main body of the PDF. It is important to note that, as k increases, the separation between target and background adaptive bandwidth values becomes wider. Specifically, the r_k mean values associated to the target locations (which, after an increasing behavior for small k values, saturate around a value of 18.0 for k>250) are one order of magnitude lower than the background ones (varying around 2.04). In addition to this large separation, although target r_k values are shown to be much widely dispersed around their mean value with respect to the background ones, the confidence intervals do not overlap for any k configuration.

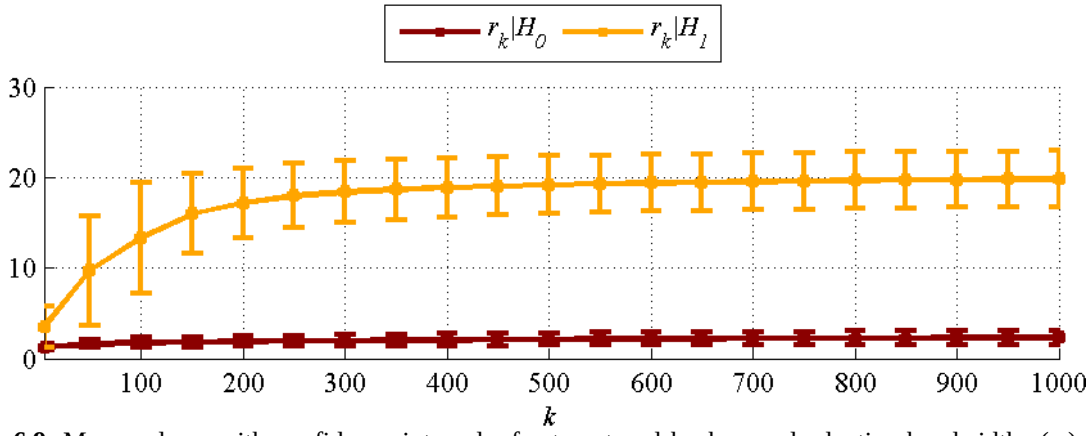


Fig. 6.9. Mean values, with confidence intervals, for target and background adaptive bandwidths (r_k) as a function of the parameter k. The confidence intervals are evaluated as the standard deviation of the corresponding r_k values, so that each bar is symmetric and two times the standard deviation long.

In order to show what the different values of the data-adaptive bandwidths attained in correspondence of target and background pixels mean in terms of adaptability to the data structure, Fig. 6.10 (a-d) show the amplitude (plotted as a vertical bar spanning from the minimum to the maximum test statistic value) of the ranges of variation of the detection test statistic values for the three VKDE strategies employed (a-c) and for the FKDE (d), with respect to the different configurations of k and h, respectively.

The limitations of using the same bandwidth across the whole image appear evident by examining Fig. 6.10 (d). In fact, the larger the bandwidth h, the lower the detection test statistic variability and, in turn, the resulting PDF estimates reliability. Specifically, the range of variation of the detection test statistics obtained with FKDE assumes an initial amplitude of 6.55 for the minimum h employed and gets narrower and narrower as h increases, up to a very small amplitude of 0.60 for the maximum h adopted. This means that the test statistics are likely to take almost the same value across the entire image at higher bandwidths, which indicates a severe over-smoothing. It should also be noted that not only the amplitude of the range of variation of the test statistics gets narrower and narrower, but also the minimum and maximum values assumed (i.e. the vertical bar extremes) move towards higher values as h increases. Since higher values of the detection test statistic mean (according to equation (2.7)) lower PDF values, such behavior clearly shows that, as h increases, most of the range of PDF values were spent to address the lower-density data regions, such as the distribution tails. As regards VKDE, Fig. 6.10 (a-c) show that the variability of the detection test statistic values appears almost constant with respect to variations of k, for all the three VKDE techniques considered. In fact, only slight

fluctuations in both the amplitude range and the specific values taken may be observed for each of the three methods. In particular, such amplitude ranges vary in [11.29, 55.09] for GkNNE, [4.07, 48.83] for kNNE, and [13.19, 50.80] for SPE, with corresponding amplitude mean values of 35.03, 27.40, and 34.36, and small values of standard deviations equal to 1.62, 0.90, and 0.99.

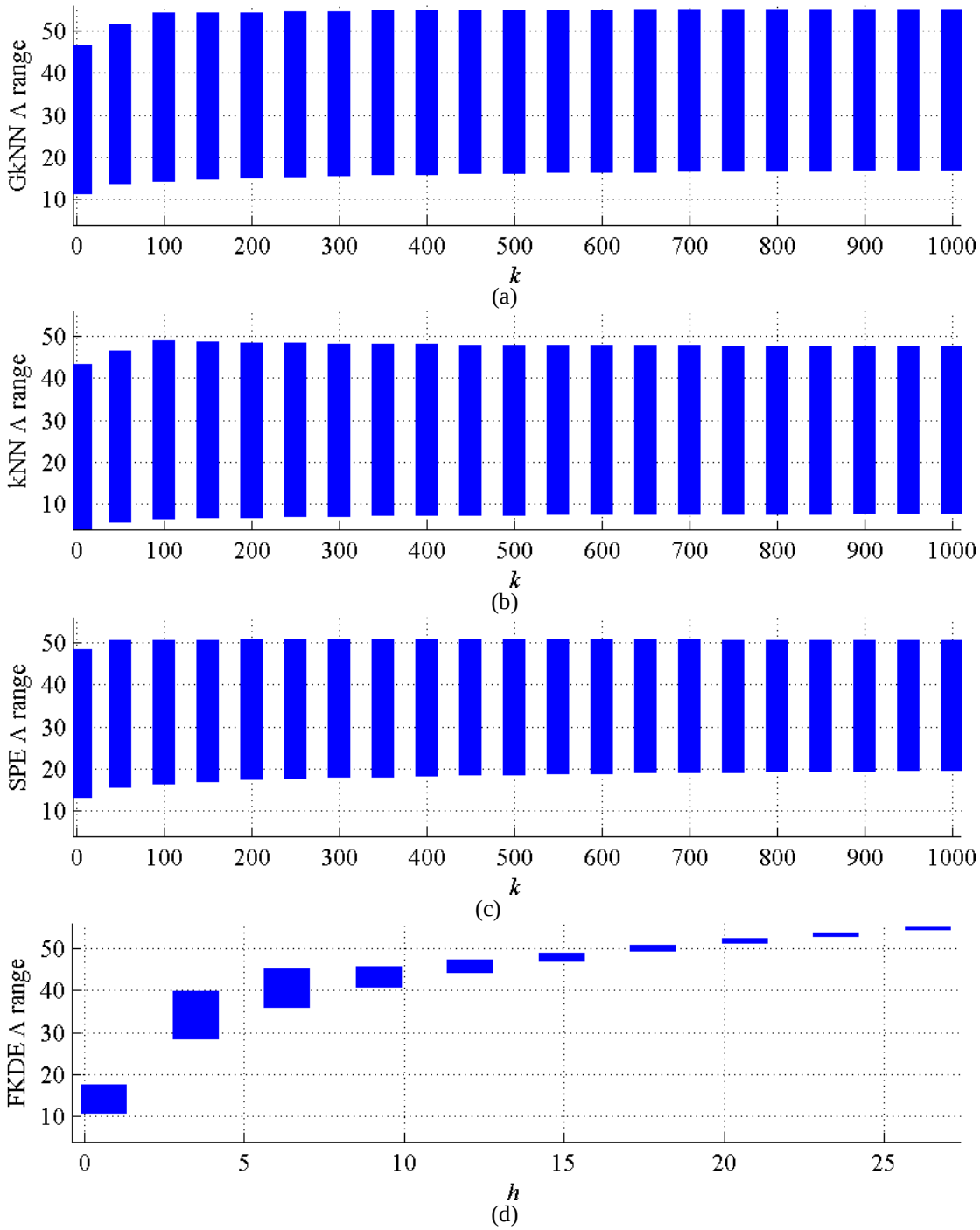


Fig. 6.10. Intervals of variation of the detection test statistics obtained by applying the proposed AD strategy employing (a) GkNNE, (b) kNNE, and (c) SPE for different choices of *k*, and (d) FKDE for different *h* values. Specifically, the vertical bars range from the minimum to the maximum of detection test statistic values.

The behavior in the detection test statistic variability is reflected in the discrimination capability between target and background pixels, which is the task of AD approaches. To this aim, the δ measurements are plotted in Fig. 6.11 (a) and (b). As is evident from Fig. 6.11 (a) concerning FKDE, after an initial increasing trend for low h values, δ starts decreasing with h, assuming lower and lower values as h increases, though never becoming negative again as to the examined image. So, as expected from the positive sign of δ for FKDE in most h configurations, *Global FAR@100%detection* is equal to zero except for an initial value of $9.49\ 10^{-2}$ (corresponding to the minimum h employed). Therefore, in the specific image examined, such a slight separation between target and background in the test detection statistic does not prevent the detection of the targets in the FKDE-based AD approach with high h. Nevertheless, as clearly shown in the "toy-example", such an over-smoothing behavior is not desirable since it severely undermines the reliability of the PDF estimate and, in general, may mask the presence of potential targets and further anomalies. In this context, VKDEs show in Fig. 6.11 (b) a clearer and much wider separation between target and background as compared to that shown by FKDE. In fact, after a slight increasing trend for k values lower than 250 - where δ assumes also negative values for kNNE and GkNNE - VKDEs exhibit δ measurements that stabilize around the value of 3 for the three VKDEs tested. Such positive δ values, obtained for almost the entire range of k tested, give evidence of the good detection ability of the proposed VKDE-based AD strategy, especially for k>100 employing kNNE and GkNNE and for k>5 in SPE. In fact, such configurations correspond to successfully detect, with no false alarms, all pixels within each target object (i.e. *Global FAR@100%detection* =0). As a result, choosing $k \geq 100$, in all configurations, weakly affects the ultimate detection performance, which is similarly very good across the different k configurations for all the proposed VKDEs-based AD strategies. Application of such a recommendation assures good detection performance to be obtained, while reliably estimating the background PDF without incurring in either under-smoothing or over-smoothing issues, as opposed to what happens with FKDE. Such results are in accordance to those obtained in the "toy-example", where small k values have been shown not to provide a good PDF approximation but similarly very good performance have been obtained for higher k values. It should also be noted that the suggested $k=N^{1/2}= 397$ according to [54] is included in the range of k values providing good performance for all the three VKDEs tested.

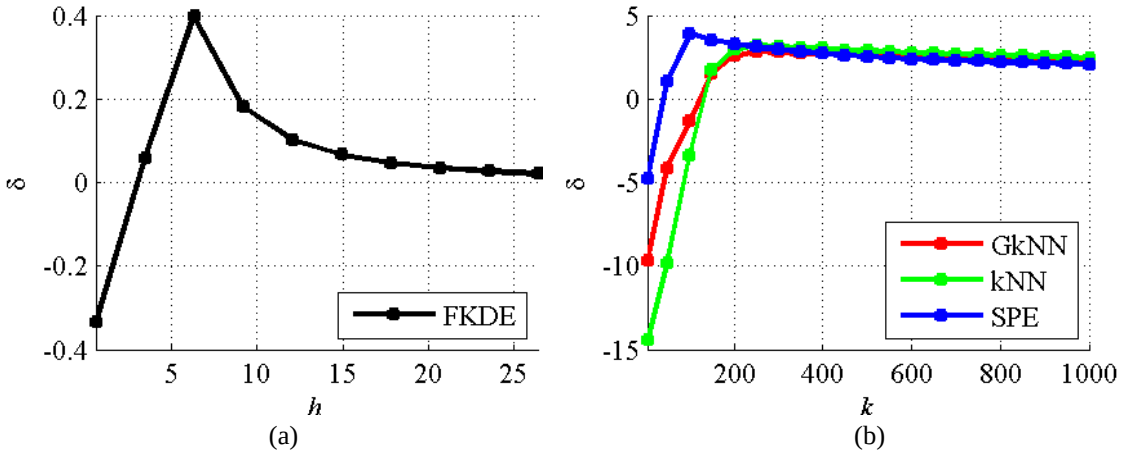


Fig. 6.11. Measures of δ corresponding to the employment of (a) FKDE and (b) the VKDEs within the proposed AD strategy.

Detection maps obtained by the proposed strategy corresponding to the VKDEs under consideration are shown in Fig. 6.12 (a-i) for three different k values. Such maps are specifically pertinent to the configurations employing the minimum (k = 5, Fig. 6.12 (a-c)), the suggested ($k = N^{1/2}$= 397, Fig. 6.12 (d-f)), and the maximum (k = 1000, Fig. 6.12 (g-i) values of the selected k.

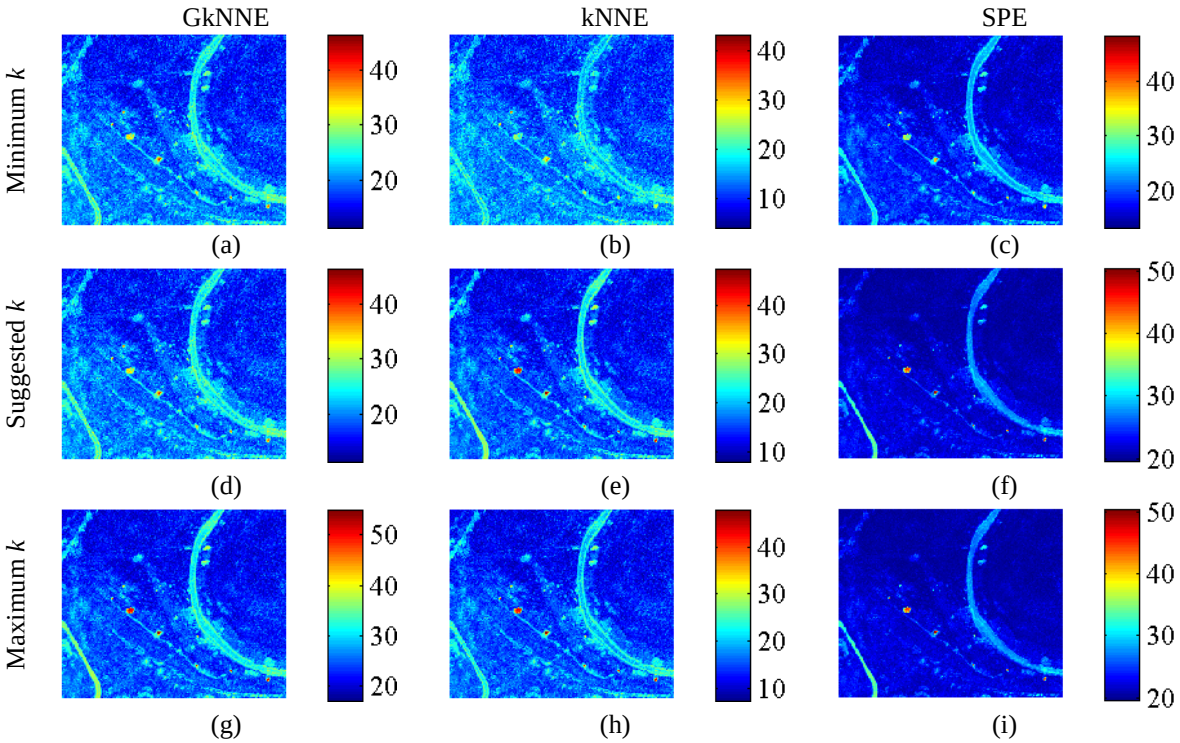


Fig. 6.12. Detection test statistics obtained by using the (a)-(c) minimum, (d)-(f) the suggested and (g)-(i) the maximum k values in the interval of interest when (a) (d) (g) GkNNE, (b) (e) (h) kNNE and (c) (f) (i) SPE are employed within the proposed AD scheme.

These representations give visual evidence of the detection ability with respect to all anomalous objects, especially, as expected, for both the suggested and the maximum k values, which allowed all the target pixels to clearly emerge from the image background in the detection maps.

This effect is much more evident by examining Fig. 6.13 (a-i), where the histograms of the detection test statistics associated to target ($\Lambda(\mathbf{x})|H_1$) and background ($\Lambda(\mathbf{x})|H_0$) pixels are reported for each AD scheme for the same three k values as above. Below each histogram plot in Fig. 6.13, horizontal bars show the ranges of variation of target and background test statistic values, highlighting the separation of target and background pixels after each PDF estimator is applied within the proposed AD scheme. As is evident, all approaches split up the values of the PDF estimate with a good trade-off between the main body and the tails of the distribution for the three configurations of k values under consideration. In particular, the histograms computed over GkNNE- and kNNE-based test statistics show similar behaviors, as we may expect since the two techniques differ only in the employed kernel function whose choice has been recognized not to seriously affect the PDF estimation outcome [23][53]. Those histograms show that the image background test statistic values have been embodied into two bumps: an evident one related to the natural vegetation, and a less pronounced one associated to the roads running through the scene and being responsible for a large number of image pixels. The road pixels, given their number, are associated to quite high test statistic values and, therefore, are characterized by lower PDF values. It is important to note that this did not impair the detection of the anomalous objects in the scene, which exhibits a clear separation from the background in the test detection statistic, as shown by the bars below the histogram plots. Rather, this plays the important role of increasing the modeling accuracy of the image background classes, thus enabling a very high material discriminability. The employment of the SPE within the proposed AD strategy provides slight different histogram plots with respect the previous ones. Specifically, the separation between target and background pixels is still clear, but the image background appears to suffer of slight over-smoothing issues. This may be linked to the phenomenon referred to as “non-locality” [58], i.e. the SPE outcomes at a certain estimation data sample $\mathbf{x}$ may be influenced by observations very far away from the estimation sample itself and not just by the nearby data.

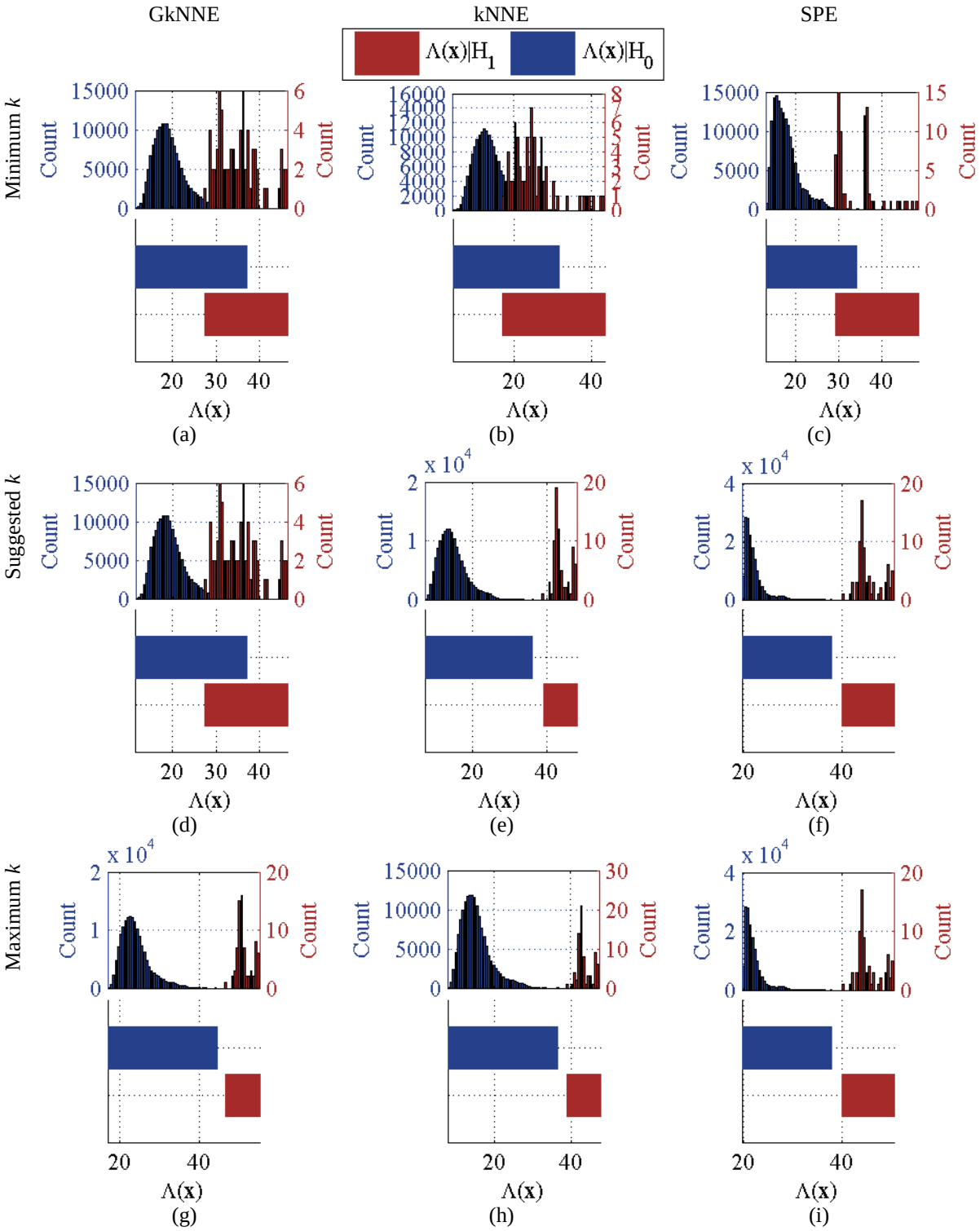


Fig. 6.13. Histogram plots of the detection test statistics associated to target pixels (in red) and background pixels (in yellow) obtained by applying the proposed AD strategy employing the (a)-(c) minimum, (d)-(f) the suggested and (g)-(i) the maximum *k* values in the interval of interest in (a) (d) (g) GkNNE, (b) (e) (h) kNNE and (c) (f) (i) SPE. Below each histogram, the interval of variation of target and background test statistic values is represented by means of horizontal bars.

For the sake of comparison, the detection test statistics resulted by employing FKDE with both the lowest and the highest selected *h* values are depicted in Fig. 6.14 (a) and (b).

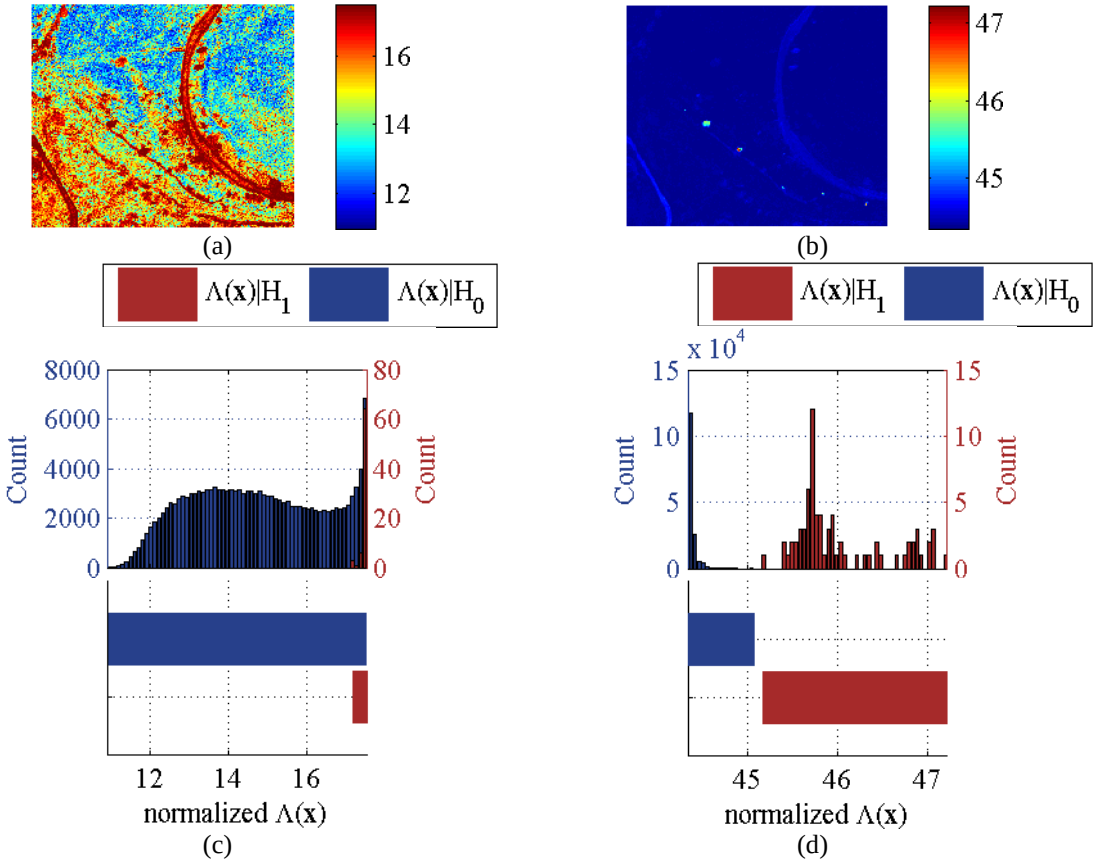


Fig. 6.14. Detection test statistics obtained by using the FKDE with (a) the minimum (i.e. 2.10) and (b) the maximum (i.e. 33.06) *h* values in the interval of interest. Below each map, histograms and horizontal bars showing the intervals of variation of the detection test statistics associated to target and background pixels.

As is evident, the use of a small bandwidth yielded a detection test statistic that exhibits the presence of a quite large number of background structures (mostly but not limited the roads) noticeably high, i.e. characterized by lower PDF values than those of the main body. This is due to the bandwidth employed, which is too small to properly capture enough sample data in the lower-density data regions. Such a phenomenon is much more evident by examining Fig. 6.14 (c), where the histograms of the detection statistics associated to target and background pixels are reported. The background test statistic values appear widely dispersed with respect to the target ones, which is mainly due to the background structures already observed emerging in Fig. 6.14 (a). Moreover, all target pixels have assumed values comparable to and lower than those of some background ones, as shown by the horizontal bars below the histogram plot in Fig. 6.14 (c). Clearly, this superimposition of target and background test statistic values prevents the targets from being detected, as already proven by the negative sign of δ for this *h* configuration in FKDE (Fig. 6.14 (a)). Also, it is worth noting that the entire range of estimated PDF values

was spent to address the background, whereas only a small range of PDF values was devoted to the tails. Fig. 6.14 (b) highlights that the high bandwidth value, on the contrary, resulted in an over-smoothing of the main PDF body, which has been completely suppressed. In fact, most of the range of PDF values was spent to address the distribution tails, whereas a much smaller range is pertinent to the main PDF body, as shown in Fig. 6.14 (d).

Of course, the employment of a bandwidth whose value is intermediate between the lowest and the highest selected may exhibit a more reliable PDF estimate. This is confirmed by the positive, though very low, values of δ in Fig. 6.11 (a) obtained for FKDE on this specific image for most of the examined h values. Nonetheless, despite such a good AD performance, the corresponding estimated image PDFs are significantly affected by the bandwidth value employed. By visual inspection of all results obtained for the different h values, results attained for the second lowest value of h employed (i.e. h= 3.47) can be said to provide the best performance within the FKDE approach. Such a configuration yields the highest δ value as well as a test statistic, which is shown in Fig. 6.15 (a), that is neither too much over-smoothed nor characterized by those background structures emerging in Fig. 6.14 (a-b). It should be noted that this h choice falls almost in the range of bandwidths obtained by employing the BN selection procedure.

(a) **(b)**

Fig. 6.15. Detection test statistics obtained by using the FKDE with h=3.47 and h= 3.47.

However, the immediately higher h value leads to the sever over-smoothing effect already observed in Fig. 6.14 (d) for most of h values, which all exhibit test statistics (Fig. 6.15 (b)) similar to that observed in Fig. 6.14 (b). Hence, an automatic data-driven choice of h is not trivial, since very small variations in the selection of h returned very different FKDE outcomes. This is mostly due to the lack of adaptivity derived by the use of a fixed bandwidth across the whole estimation domain. Conversely, the high flexibility achieved by letting the bandwidth vary according to the local data-density in the feature space has

shown to provide results that not only are very weakly affected by *k* but also combine good detection performance and PDF estimation reliability.

6.6 Experimental results validation over the benchmarking data set

This section gives an overview of the results obtained on *Scene B* with the aim of validating, on the benchmarking data set, the results so far discussed.

As regards FMMs, results obtained on *Scene B* confirm the ability of the StMM to capture the heavy-tail behavior of the examined data and, in turn, to more properly model their structure with respect to the GMM case. This clearly emerge by examining Fig. 6.16 (b), where the exceedance plots computed over the StMM cluster map, which is shown in Fig. 6.16 (a), are reported.

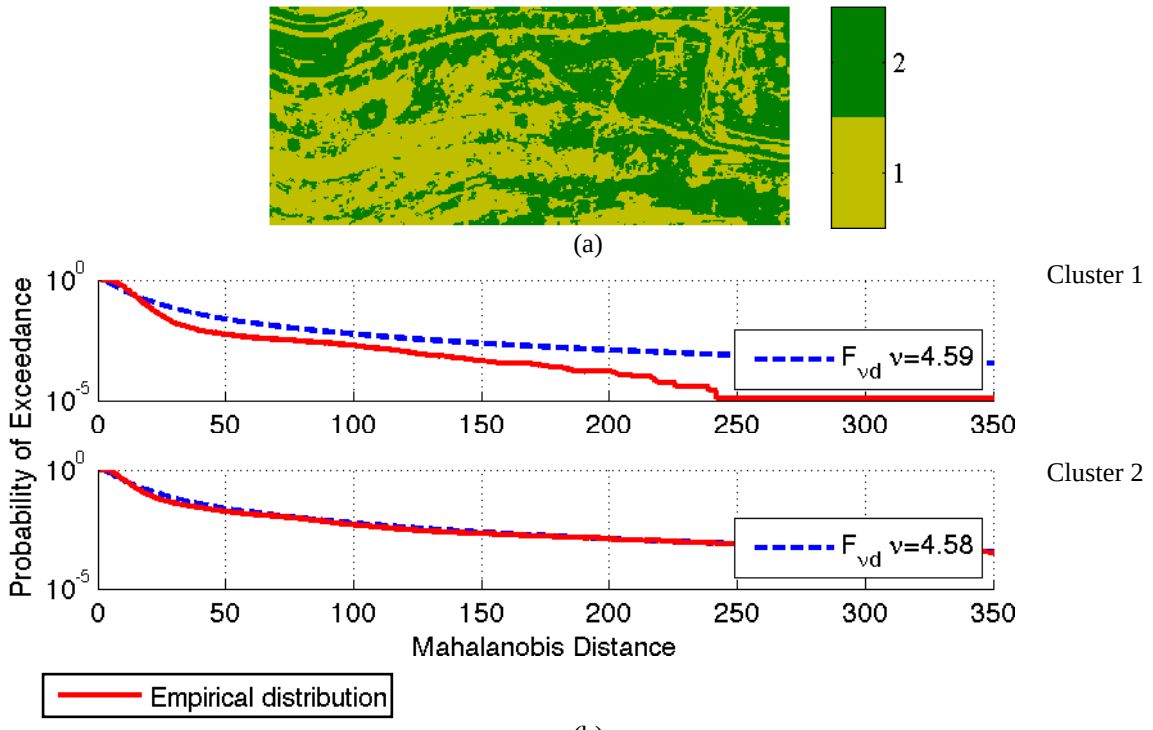


Fig. 6.16. (a) Cluster map and (b) exceedance plots of the Mahalanobis distances for the spectral classes obtained by employing the StMM learning strategy on *Scene B*. Specifically, the empirical distributions are plotted in red (solid curves), whereas the blue (dashed) curves represent the F distributions considered for comparison to the empirical distributions, respectively. This latter refers to the F distributions $F_{\nu,d}$ characterized by the numbers of degrees of freedom returned by the learning algorithm.

Specifically, results indicate that the Gaussian model does not accurately describe the statistical behavior of the majority of background classes. In particular, whereas the χ^2

distribution does a good job in modeling the main body of the Mahalanobis distance distribution, it does not accurately models the tails. In fact, on average, the complementary empirical CDF of the Mahalanobis Distance of each GMM component is better fit by a theoretical F-distributed curve rather than a χ^2 one. As regards the StMM, similarly to what shown for *Scene A*, such a model does a better job than GMM in matching both the main body and the longer distribution tails. Specifically, the empirical distribution of each background cluster resulting from employing the StMM Bayesian learning approach is in good agreement with the theoretical model as to both the distribution body and the tails for most StMM components.

As to the experiments of non-parametric background PDF estimation through the FKDE approach, since such an approach assumes an equal bandwidth *h* for each component, the data were previously normalized so that each spectral component had the same variance. According to the BN methodology for selecting the bandwidth, the number of nearest neighbors is defined by the interval $[K_l, K_u]$, which is the only user-specified parameters. Consistently with the analysis performed over *Scene A*, in order to evaluate the impact of the user-specified parameters on the detection performance, both SNR_Λ and *Global FAR@100% detection* measures obtained for different values of K_l and K_u were considered. For this purpose, a (N/K_l, Δ) space was also generated by varying $[K_l, K_u]$ within all configurations tested for *Scene A*. The SNR_Λ values computed over the *Scene B* on the basis of the available ground truth target map are shown in Fig. 6.17 (a). As is evident, the region including SNR_Λ values not lower than approximately 3dB with respect to its maximum value is approximately identified by $N/K_l<2200$. Therefore, the region depicted by $N/K_l<600$, identified for *Scene A*, places itself within such a plateau, especially in the part characterized by the highest SNR_Λ values. In particular, if the recommendation ($N/K_l<600$) obtained from *Scene A* analysis is followed, the SNR_Λ values obtained on *Scene B* range within [21.48 24.04], with an average value of 23.14 and standard deviation of 0.38. This means that choosing a number K_l of nearest neighbors not lower than 2 orders of magnitude with respect to the total number *N* of pixels is sufficient to exhibit a very good background suppression ability. Fig. 6.17 (b) displays the *Global FAR@100% detection* values obtained. Choosing $[K_l, K_u]$ within $N/K_l<600$ correspond to a *Global FAR@100% detection* never higher than $5.20\cdot10^{-2}$ and ranging in $[5.18\cdot10^{-2}, 5.20\cdot10^{-2}]$. Besides, the best value obtained in the whole (N/Kl, Δ) space is $5.00\cdot10^{-2}$, just slightly lower than the $5.20\cdot10^{-2}$ achievable within the $N/K_l<600$ region. Therefore, results obtained in terms of

both SNR_Λ and *Global FAR@100% detection* confirm the presence of a plateau in the (N/Kl, Δ) space onto which detection performance is similarly good. As regards the corresponding bandwidth values, h ranges in [3.40, 4.18], with an average value of 3.79 and a standard deviation of 0.12. Once again, such values show a very limited variation with respect to significant variations of the [K_l, K_u] in the (N/K_l, Δ) space.

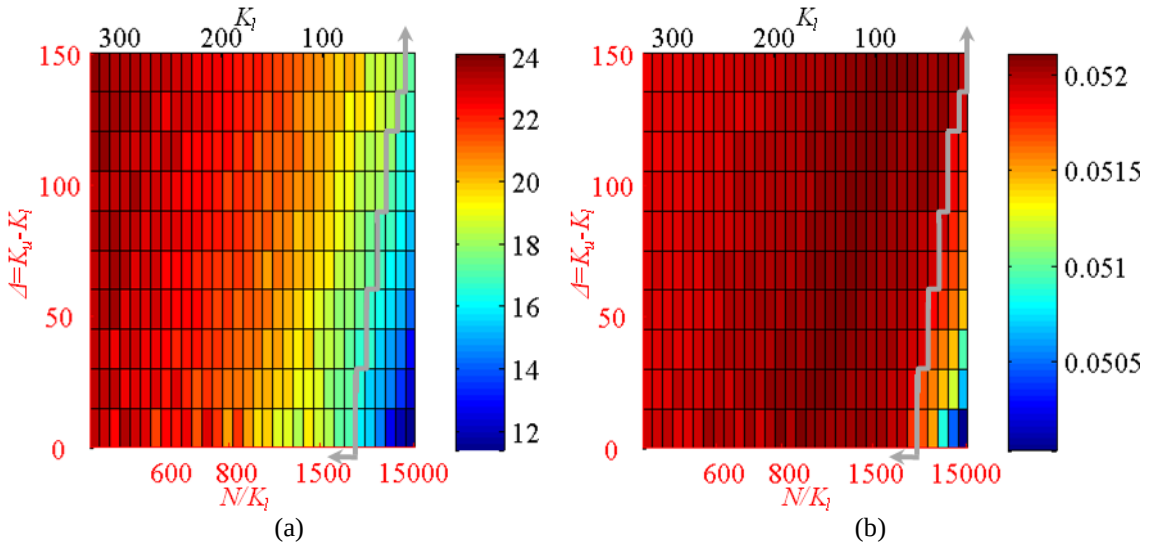


Fig. 6.17. (a) SNR_Λ and (b) *Global FAR@100% detection* measures for different configurations of the interval [K_l, $K_u=K_l+\Delta$]. The red arrow depicts the region including SNR_Λ values not lower than approximately 3dB with respect to its maximum value.

AD performance benchmarking on *Scene B* was carried out evaluating the same performance measures as for *Scene A*.

Fig. 6.18 shows ROC curves for the GMM-, StMM-, and FKDE -based AD schemes. Here, the employment of the Bayesian algorithms BGMMS, BStMM, ans BN are considered for learning the models, similarly to what done for *Scene A*. For the sake of comparison, both FKDE configurations corresponding to the best and the worst SNR_Λ within the recommended region N/K_l<600 were retained for evaluation. Furthermore, the ROC curve associated to the best *Global PFA@100% detection* value obtained in the whole (N/Kl, Δ) space ($5.00\cdot10^{-2}$) is also shown.

As in *Scene A*, StMM yields the best overall detection performance, providing a ROC curve with higher detection probabilities for similar values of FAR with respect to the other curves and assuring a FoDT=0.8 with no false alarms.. In addition, GMM is found to perform worse than StMM, similarly to what found in *Scene A*. However, on *Scene B*,

FKDE is found to perform not as good as in *Scene A*, but rather more similarly to GMM. Specifically, the two ROC curves reported for the recommended region are very similar to each other, showing again that the overall detection performance does not vary significantly when [K_l K_u] is chosen in the recommended region. They are also similar to the GMM curve, though yielding better FAR values for FoDT around 0.70. As regards the ROC curve obtained when the minimum *Global FAR@100% detection* is sought in the whole (*N*/*Kl*, Δ) space, such a curve does not exhibit a considerable improvement of overall detection performance with respect to the other two curves. In fact, though showing lower FAR values for 0.81<FoDT<0.90, it provides equivalent FARs for FoDT>0.90 and even worse FARs for FoDT<0.81. Such outcomes confirm the robustness of the recommendation assessed on *Scene A*, which has allowed good detection performance to be obtained on *Scene B*. In practice, one might have just picked up a [K_l K_u] configuration following the highlighted recommendation and obtained similarly good detection performance.

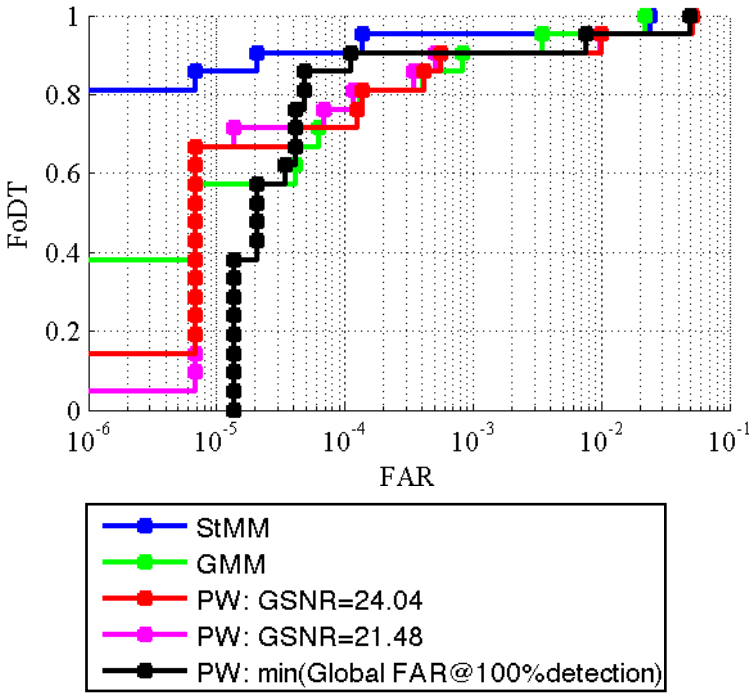


Fig. 6.18. ROC curves for *Scene B*.

It should also be noted that, in general, all approaches have shown a decrease of overall detection performance over *Scene B*, which is very likely to be linked to the more challenging scenario encompassed with respect to *Scene A*.

Object-wise performance measures were also evaluated. Table 10 and Table 11 report *FAR@1st detection* and *FAR@100% detection* measurements, respectively. Here, very few perfect detections are obtained and higher FAR values with respect to *Scene A* can be seen. Again, these are linked to the increased difficulty of the detection task. Perfect detection of Obj. 1 and 2 location is obtained (*FAR@1st detection*=0) by the proposed AD schemes. These two objects are indeed the largest ones, and, hence, their location is more easily detected since some of their pixels in the image are more likely not to be contaminated by background pixels (i.e. full-pixels). Conversely, their lower *FAR@100% detection* is likely to be due to their wider extent as well, since more boundary mixed-pixels have to be properly target-labeled so as to achieve a 100% detection. On the contrary, Obj. 3, 4 and 5 show equal *FAR@1st detection* and *FAR@100% detection* measures, since they consist of only one pixel in the ground truth map.

Table 10. Measures of *FAR@1st detection* (*Scene B*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5
GMM	**0**	**0**	$4.1\cdot10^{-4}$	$3.4\cdot10^{-3}$	**$2.2\cdot10^{-2}$**
StMM	**0**	**0**	**$6.9\cdot10^{-6}$**	**$1.4\cdot10^{-4}$**	$2.4\cdot10^{-2}$
FKDE	**0** –$6.9\cdot10^{-6}$	**0**	$3.5\cdot10^{-4}$– $4.1\cdot10^{-4}$	$9.9\cdot10^{-3}$– $1.0\cdot10^{-2}$	$5.18\cdot10^{-2}$- $5.20\cdot10^{-2}$

Table 11. Measures of FAR@100% detection (*Scene B*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5
GMM	$8.2\ 10^{-4}$	$6.9\ 10^{-5}$	$4.1\ 10^{-4}$	$3.4\ 10^{-3}$	**$2.2\ 10^{-2}$**
StMM	**$2.1\ 10^{-5}$**	**0**	**$6.9\ 10^{-6}$**	**$1.4\ 10^{-4}$**	$2.4\ 10^{-2}$
FKDE	$5.1\cdot10^{-4}$- $5.7\cdot10^{-4}$	$6.9\cdot10^{-5}$- $1.2\cdot10^{-4}$	$3.5\cdot10^{-5}$- $4.1\cdot10^{-4}$	$9.9\cdot10^{-3}$ - $1.0\cdot10^{-2}$	$5.18\cdot10^{-2}$- $5.20\cdot10^{-2}$

Table 12. Measures of $TSNR_{\Lambda}$ (*Scene B*)

Learning strategy	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5
GMM	16.50	25.05	4.66	3.39	**2.34**
StMM	9.96	10.74	6.56	**4.82**	2.16
FKDE	29.88 – **37.69**	31.74 – **43.03**	11.77 – **11.91**	3.68 – 3.75	1.57 – 1.59

$TSNR_{\Lambda}$ values included in Table 12 bring into view what already observed for *Scene A*. Despite the better detection performance, the ability of the StMM-based approach to suppress all image background has been exceeded by both GMM-based and FKDE -based approaches. Also, it is worth noting the objects that are the largest in size (i.e. Obj. 1 and 2)

are those characterized by higher $TSNR_{\Lambda}$, whereas the others, and more specifically the smallest Obj. 5, exhibit much lower $TSNR_{\Lambda}$. Such a behavior is common to all approaches.

Scene B was also employed to compare the use of the AKDEs, provided by equations (4.24), (4.25), and (4.26), with the ability of the FKDE in (3.12) within the proposed AD strategy. Since knowledge of the true background PDF is not available, the analysis was carried out mainly on the basis of comparing the PDF estimates and the detection performance. According to the methodology for selecting the bandwidths in the AKDEs, the only element to be set by the user is *k*. As for *Scene A*, the ability of the proposed AD scheme to detect anomalous objects was investigated with respect to different choices for *k*. To this aim, *k* was varied between 5 and 1000.

As mentioned earlier, the adaptive kernel approaches follow the sparseness of the data by using broader kernel functions over observations located in regions of low density, where we expect that the targets are located. In particular, the AKDEs are able to adjust the different smoothing requirements in the main body of the PDF (i.e. the image background) and in the target locations that we expect to lie in the tails of the PDF. The values of *k* employed led r_k to takes values ranging from 1.32 to 22.05, where the values at the target pixel locations are again much higher than the ones used for the rest of the pixels.

Consistently with the analysis performed over *Scene A*, results obtained by employing the FKDE with different choice for *h* are also examined. For this purpose, as in the experiments involving *Scene A*, a Gaussian kernel was employed. Moreover, in this analysis, the choice of *h* within the FKDE was done selecting 10 values between the lowest (i.e. 1.32) and the highest (i.e. 22.05) r_k values, which should provide a better approximation of the PDF body and tails, respectively.

The limitations of the fixed bandwidth kernel estimation have been confirmed by examining the detection test statistics corresponding to the examined AD schemes. In fact, the use of small bandwidths have yielded PDF estimates that exhibit spiky behaviors, so that a quite high number of regions (characterized by a lower density with respect to the main body) resulted not well represented in the final estimate. On the contrary, high bandwidth values have resulted in an over-smoothing of the main PDF body. Consequently, the detection test statistic has taken almost the same value over the entire image. Also, most of the range of PDF values was spent to address the distribution tails, whereas a much smaller range is pertinent to the main PDF body. This is evident just

examining Fig. 6.19 (d), where vertical bar spanning from the minimum to the maximum test statistic value show the amplitude of the ranges of variation of the detection test statistic values for the FKDE.

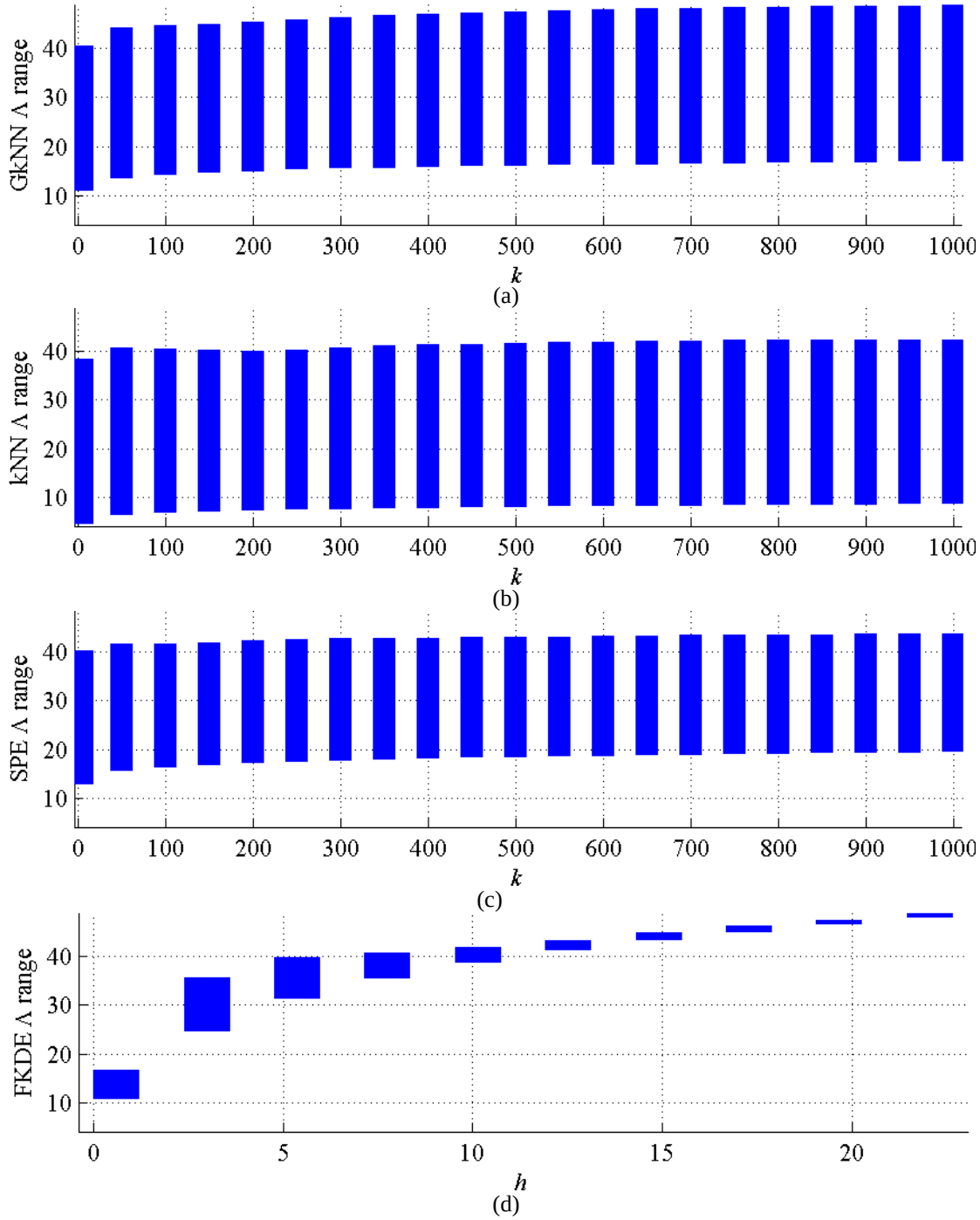


Fig. 6.19. Intervals of variation of the detection test statistics obtained by applying the proposed AD strategy employing (a) GkNNE, (b) kNNE, and (c) SPE for different choices of *k*, and (d) FKDE for different *h* values. Specifically, the vertical bars range from the minimum to the maximum of detection test statistic values.

In particular, it is noted that, again, as *h* increase, the amplitude of the range of variation of the test statistics gets narrower and narrower, as well as moves towards higher and higher values. Nevertheless, such a not desirable over-smoothing behavior undermines the reliability of the PDF estimate and leads to mask the presence of potential targets. As a result, after a slight increasing trend, *Global FAR@100% detection* measurements stabilize around the value of $5.0 \cdot 10^{-2}$.

The employment of a bandwidth whose value is tailored to the local data structure assures a more reliable PDF estimation process, which does not incur in either under-smoothing or over-smoothing issues, and which should provide more robust detection outcomes. In particular, the variability of the detection test statistic values appears almost constant with respect to variations of *k*, as illustrated in Fig. 6.19 (a-c). Within this framework, *Global FAR@100% detection* values is around $6.0 \cdot 10^{-2}$ for SPE. Nevertheless, GkNNE and kNNE does not a good job as SPE. Specifically, both GkNNE and kNNE returned *Global FAR@100% detection* measurements lower than $3.0 \cdot 10^{-2}$ for all *k* configuration tested. Once again, these detection behaviors are related to the increased difficulty of the detection task in this operational scenario.

6.7 Final remarks and conclusions

In this thesis work, a global AD strategy is proposed based on the LRT decision rule, in which reliable estimation of the background PDF is addressed. In this chapter, the ability of semi- and non-parametric approaches have been analyzed with the aim of modeling the statistical variability of hyperspectral data in order to detect spectral anomalies within the proposed AD scheme. Specifically, the employment of StMM, GMM, FKDE and AKDEs has been thoroughly investigated. Although such semi- and non-parametric PDF estimators are used in a variety of multivariate signal processing problems, the difficulty in learning the underlying models both reliably and automatically has made their application in the hyperspectral AD context very limited.

Methodologies developed within a Bayesian framework have been considered for the parameter selection of StMM, GMM, and FKDE. The conducted experimental analysis has focused on three aspects:

- Evaluating the ability of the considered mixtures based distributions, learnt with the Bayesian approach, to represent the statistical behavior of real hyperspectral data

and, specifically, to asses both GMM and StMM behavior as regards the distribution tails.

- Assessing how much the change in the configuration [K_l, K_u] affects the detection performance of the proposed AD scheme applied with FKDE.
- Evaluating and comparing algorithm behavior and detection performance of the proposed AD strategy when GMM, StMM and FKDE are used.

For these purposes, the experimental analysis involved two real hyperspectral images and the evaluation of several different performance measures.

As regards the semi-parametric approaches to PDF estimation, experimental results have indicated that the GMM has difficulty in properly modeling the empirical distribution of background classes in real hyperspectral data, due to the need of addressing longer tails than the Gaussian ones. On the contrary, StMM has been shown to yield a powerful model for the statistical characterization of hyperspectral data, since it benefits from the better fit the EC PDFs provide to the distribution tails. On the other hand, on these data, the use of the StMM learning within the proposed AD scheme has been shown not to properly suppress the predominant background structures in the detection test statistic map, as confirmed by the lower $TSNR_\Lambda$ values obtained with respect to the other methods. Nevertheless, such lower background suppression ability has not resulted in a similarly lower detection capability. Rather, as indicated by the ROC curves as well as by the *FAR@1st detection* and *FAR@100% detection* measures, the StMM has been proven particularly effective at detecting the anomalous targets placed in both the scenes examined.

As to the FKDE approach, though it is supposed to learn the PDF entirely from the data, the bandwidth selection process examined requires the choice of the bounds K_l and K_u to be specified by the user. The problem of making the choice of the bandwidth automatic, which is highly desirable in practical AD tasks, was investigated. Specifically, a common recommendation for the region where [K_l, K_u] should be selected, capable of assuring similarly good performance and applicable to different scenarios, was sought. The identification of such a recommendation will minimize the importance of the operator intervention thus making the strategy automatic. On the first examined data, all the configurations characterized by K_l>N/600 have been shown to assure similarly good performance, in that *Global FAR@100% detection* and SNR_Λ values exhibit – in that

region - the presence of a "plateau" in the (N/K_l,Δ) space. The automatic application of such a recommendation to the secondary data set has allowed good performance to be obtained. Additionally, further analysis has confirmed the presence of the "plateau" in the (N/K_l,Δ) space, which has been shown to include the recommended K_l>N/600 configurations.

As regards comparative AD performance analysis, all three different AD schemes examined that makes use of a Bayesian approach have been proven to be effective at detecting the anomalous objects present in the two scenarios. In particular, on the examined images, the StMM-based scheme has provided the best detection performance in both scenarios, being capable of detecting all anomalous targets with the fewest false alarms.

Though the FKDE-based AD algorithm has provided, on these data, performance not so good when compared to those obtained by using the StMM-based semi-parametric estimator, such a non-parametric approach has been shown to be the most attractive approach to be applied in practical AD tasks. This is mostly due to FKDE independence of specific background distributional assumptions. However, the single smoothing parameter *h* used in the FKDE can be ineffective for modeling complex PDFs. In this work, some of the possibilities for reliability improvement of non-parametric PDF estimation by varying the bandwidth over the domain of estimation have been investigated. Specifically, the BE and the SPE methodologies, in which the bandwidth varies with the sample of estimation and with the sample observation, respectively, were employed in the proposed AD scheme. Within this framework, the *k*-nearest neighbor rule has proven to be an intuitively appealing procedure to adapt the bandwidth to the local density of data. However, application of such a method is inhibited by lack of knowledge about the manner in which it is influenced by the value of *k*, and by the absence of techniques for empirical choice of *k*. Therefore, the ability of the variable bandwidth kernel density estimators within the log-likelihood based AD scheme was investigated with respect to the FKDE. The experiments were conducted with respect to different choices for the bandwidths *k* and *h*, the only user specified parameters of VKDE and FKDE, respectively.

Experimental results obtained have confirmed the great potential of the VKDEs for non-parametric PDF estimation when attention is focused on small rare objects. In fact, although FKDE application has still allowed, in most cases, the anomalous objects in the scene to be detected, the use of a fixed bandwidth over the entire feature space has been

shown to provide very diverse detection test statistics with respect to the choice of h. Specifically, whereas for the lowest fixed bandwidth employed the entire range of PDF values has been spent to address the main image structures and background, with the highest h most PDF values have served the anomalous objects with a major over-smoothing phenomenon as to the PDF body. On the contrary, the VKDEs have been shown to provide detection test static values yielding a good trade-off in allocating PDF values for both the main PDF body and the tails. This has been achieved thanks to the variable bandwidths r_k obtained by the employment of the k-NN approach, which well adapted to the different local peculiarities of data in the feature space.

Also, the low dependence of VKDE on k shown in the "toy example" has been confirmed by the real multispectral data analysis: except for the smallest values of k, most results have provided very low diversity in the range of values of the detection test statistic, all assuring a wider separability between anomalous objects and background with respect to FKDE. The recommendation of choosing $k=N^{1/2}$ has proven to be effective also with the real data tested, providing very good detection performance while preserving the desired adaptability of the estimated PDF to the image data.

Although in this paper the variable-bandwidth PDF estimators are applied only for enhancing the separation of anomalous objects with respect to the image background, their high flexibility and adaptability suggest to employ variable bandwidths to other tasks of multi-hyperspectral image analysis requiring reliable PDF estimates, such as spectral signature based target detection, image clustering, and many others.

Finally, it should be noted that the variable bandwidth PDF estimators can be computationally expensive in practical circumstances and, in order to fully exploit their great potential, attempts to increase the computational efficiency are needed and will be dealt with in future works.

Chapter 7

7 Experimental results: local AD performance

In this chapter, an experimental analysis is provided in order to investigate into the effectiveness of the presented solution to the poor detection performance due to the LNM assumption of the conventional RX approach. Thus, experiment design is intended to analyze the detection capability of A-RX with respect to classical AD algorithms in an operational scenario.

7.1 Data set description

The proposed AD strategy is validated using the same portion of real data. Again, the actual image used for testing refers, for simplicity, i.e. both to speed-up the computation and not to incur any kind of *curse of dimensionality* issues [54], to a spatial and spectral subsets of the original hyperspectral cube. The subset used for testing refers to the portion of size 365 by 430 pixels of the whole flight line denoted with *Scene A* in chapter 6, but otherwise processed. Here, in contrast to what done before, besides water-vapor absorption and noisy bands removal, the spectral subset was obtained resorting to a spectral binning and down-sampling procedures, thus obtaining 23 spectral data samples.

The analyzed data portion is interesting since it includes numerous panels of different sizes and materials embedded in different kinds of local background as targets of interest for AD purposes. A true color image of the scene reporting target locations is shown in Fig. 6.1 (a).

7.2 Design of the experiments

The goal of the conducted experiments was to provide insights about the GkNNE effectiveness and actual usefulness in the local AD context with respect to classical local AD algorithms [62]. Specifically, the comparison was performed between the proposed AD strategy, described in detail in section 5.3, and both the RX and the K-RX detectors (see sections 5.1 and 5.2, respectively).

As commonly performed in local AD (and already mentioned in section 2.4.2), in order to prevent potential target pixels to affect local background characterization, the tested algorithms were applied by sliding a dual concentric window over every pixel in the image. The size of the interior window was assumed to be the largest expected target size in the scene. The size of the outer window was set so as to include at least $10 \cdot d$ samples from the neighborhood of the pixels under test for local background characterization (i.e., $N \approx 10 \cdot d = 264$). Thus, the sizes of the inner and outer windows used for the dual window technique were 19×19 and 25×25, respectively. Also, the local mean-removal procedure in the RX approach was performed by using a sliding window of outer size 21×21.

As to the kernel functions, $\kappa(.)$ in A-RX was taken to be a multivariate Gaussian PDF, as commonly done in the literature, whereas the Gaussian RBF (GRBF) kernel was used to implement the K-RX algorithm, as in [32].

According to the proposed A-RX methodology, the number of nearest neighbors in the GKNNE is defined by k, which has to be set by the user. Therefore, the experiments were conducted with respect to different choices of k, so as to evaluate the impact of k over the detection performance. To this aim, k was varied from 5 up to 260 ($\approx N$). Similarly, K-RX was applied with respect to several configurations for the bandwidth h (i.e., the GRBF kernel function width). Specifically, h was chosen uniformly sampling 36 values between 1 and 70.

Detection performance of the examined algorithms is evaluated on the basis of the available ground truth target map. Since evaluation of anomaly detection performance is not a trivial task, several performance measures are adopted. To this aim, the FAR corresponding to the maximum threshold value in the detection test statistic at which all target pixels are detected (already denoted with *Global FAR@100% detection*) is retained as summary measure of the overall AD performance [42]. In this analysis, *Global FAR@100% detection* measures are evaluated for the different values of both k and h in A-

RX and K-RX, respectively. Then, A-RX and K-RX configurations yielding both the minimum and maximum *Global FAR@100% detection* measurements are retained for further and more in-depth investigation. Within this framework, pixel-wise ROC curves are evaluated [42]. Besides ROC curves, in order to analyze algorithm behavior over specific targets, two kinds of object-wise performance measures are also adopted. The first measure provides the FAR at the first detection (denoted with *FAR@1st detection*), i.e. the FAR for just locating the desired target, being associated with its pixel with highest test statistic value. The second performance measure employed is the FAR at full detection (denoted with *FAR@100% detection*), an object-wise version of the *Global FAR@100% detection*, aimed at assessing the FAR arising from the detection of all pixels within each target object.

7.3 Result discussion

In this section, the multispectral image described in section 7.2 is used in order to evaluate the detection performance of the proposed A-RX strategy as compared to both RX and K-RX detectors.

First, the impact over detection performance of the user-specified *k* parameter in A-RX is evaluated and compared to that of *h* in K-RX. To this aim, the *Global FAR@100% detection* measurements obtained for different values of *k* and *h* are plotted in Fig. 7.1 (a) and (b), respectively. As is evident from Fig. 7.1 (a) concerning the A-RX approach, *Global FAR@100% detection* values are approximately constant for the whole range of *k* tested. In fact, for detecting all target pixels the A-RX strategy takes FAR varying between $2.11 \cdot 10^{-3}$ and $2.28 \cdot 10^{-3}$ with a corresponding mean value of $2.22 \cdot 10^{-3}$ and a very small value of standard deviation equal to $4.35 \cdot 10^{-5}$. Therefore, only slight fluctuations may be observed as *k* varies within A-RX. Furthermore, it is important to note that all these *Global FAR@100% detection* values are one order of magnitude lower than that yielded by the RX algorithm, which provides *Global FAR@100% detection*=$2.23 \cdot 10^{-2}$. Conversely, K-RX exhibits *Global FAR@100% detection* values that strongly vary with *h*, as shown in Fig. 7.1 (b). Specifically, on the examined scenario, *Global FAR@100% detection* for K-RX assumes an initial (for the minimum *h* employed) value of $2.02 \cdot 10^{-2}$, which is comparable to that yielded by RX, ad starts decreasing down to $1.83 \cdot 10^{-3}$ as *h* increases. In this case the mean value is $3.78 \cdot 10^{-3}$ whereas the standard deviation is equal to $3.76 \cdot 10^{-3}$.

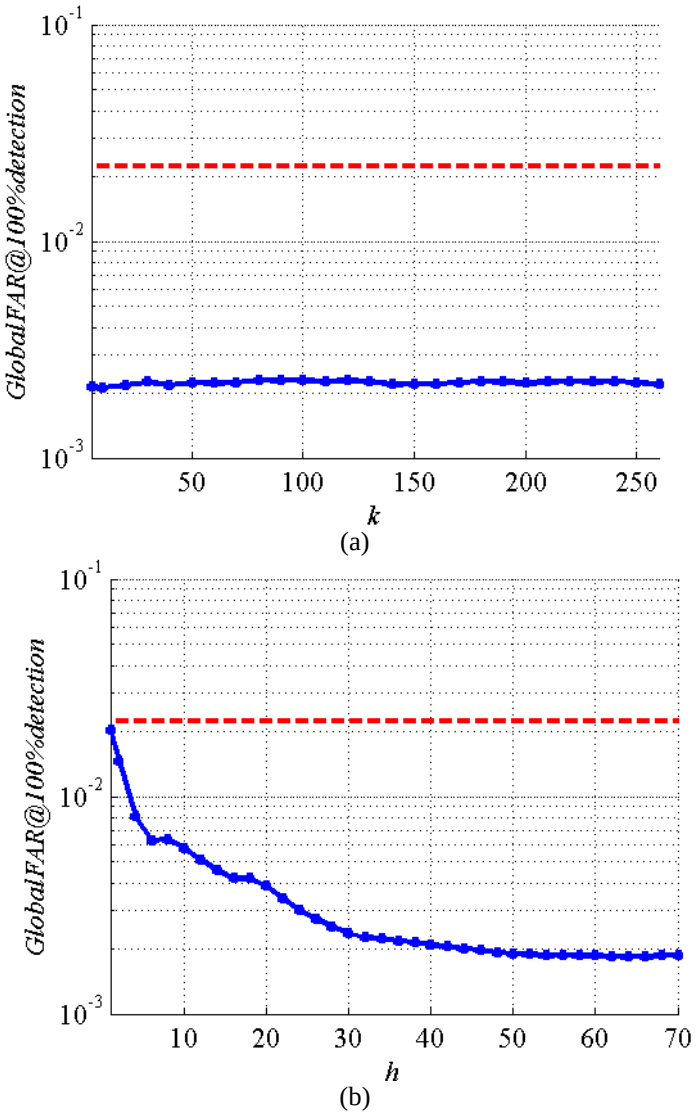


Fig. 7.1. (a-b) *Global FAR@100%detection* measurements for different configuration of (a) *k* and (b) *h* in A-RX and K-RX, respectively. The red dashed line refers to the *Global FAR@100%detection* value obtained with RX.

As anticipated, the A-RX and K-RX configurations corresponding to the minimum and the maximum *Global FAR@100% detection* measures were retained for further evaluation. In this scenario, such configurations are specifically pertinent to k={10, 80} and h={64, 1}. The corresponding ROC curves are shown in Fig. 7.2 together with that of RX. As is evident, despite the large difference between the two k values retained, the ROC curves reported for A-RX are very similar, showing that overall detection performance is similarly very good for all the examined k values. Such curves are also similar to the K-RX curve exhibiting the minimum *Global FAR@100% detection*. This latter yielded a slightly lower FAR value for FoDT=1 with respect to A-RX approach but also provided quite lower

FoDT values than A-RX as to the lowest FAR region. On the other hand, the K-RX ROC curve obtained for the maximum *Global FAR@100% detection* exhibits a considerable degradation of overall detection performance with respect to the other K-RX ROC curve shown. In particular, it is very similar to the one yielded by the RX algorithm. Of course, all K-RX configurations tested yield ROC curves lying between the two K-RX ROC curves shown. This confirms the high sensitivity of K-RX with respect to *h*, which manifests itself not only at the *Global FAR@100% detection* level but also throughout the entire ROC curve.

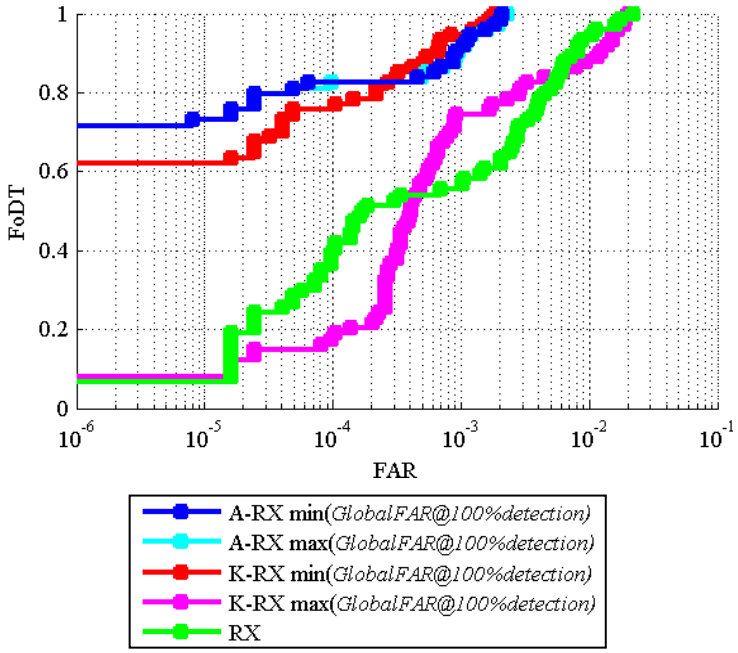


Fig. 7.2. ROC curves.

As to algorithm behavior over each target object, tables 13 and 14 report measures of *FAR@1st detection* and *FAR@100% detection*, respectively, as regards A-RX, K-RX and RX strategies. Specifically, A-RX and K-RX measures are those of the corresponding configurations yielding the minimum *Global FAR@100% detection* (i.e. *k*=10 and *h*=64). As the ROC curves have revealed, very slight differences are expected between A-RX and K-RX measures. In particular, A-RX manages to locate each target object with no false alarms, as is evident from the null *FAR@1st detection* measurements in Table 13. Similarly, K-RX exhibits all *FAR@1st detection*=0 except for Obj. 7, whose detection makes ten false alarms arise. Conversely, RX exhibits only three out of seven null

FAR@1st detection measures, showing a higher difficulty in locating the different target objects. Table 14 show that both A-RX and K-RX succeed in detecting, with no false alarms, all pixels within most of the target objects (i.e. *FAR@100% detection*=0 for five out of seven objects). As could be expected by analysis of ROC curves, K-RX does a (slightly) better job in detecting the most difficult Obj. 3 target, with a *FAR@100% detection* measure slightly lower than that of A-RX, whereas A-RX exhibits better performance in the lowest FAR region, with a *FAR@100% detection* for Obj. 7 one order of magnitude lower than that yielded by K-RX. Nonetheless, both A-RX and K-RX outperform RX, which exhibits non-null *FAR@100% detection* for six out of seven objects. This major difference between RX and both A-RX and K-RX may be linked to the non-homogeneous nature of the local backgrounds surrounding the targets. In this context, natural vegetation, variability in the field, and the two roads make the background surrounding the targets highly cluttered and non-homogeneous. Such conditions clearly violate the LNM assumption of RX, thus resulting in decreased performance. On the contrary, both A-RX and K-RX strategies seem to better model the non-Gaussian and possibly multimodal support of the background pixels, and, thus, provide better detection performance.

Table 13. Measures of *FAR@1st detection*

Method	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7
A-RX	**0**	**0**	**0**	**0**	**0**	**0**	**0**
K-RX	**0**	**0**	**0**	**0**	**0**	**0**	$4.05 \cdot 10^{-5}$
RX	**0**	$1.62 \cdot 10^{-5}$	$1.86 \cdot 10^{-4}$	$1.62 \cdot 10^{-5}$	**0**	**0**	$1.62 \cdot 10^{-5}$

Table 14. Measures of *FAR@100% detection*

Method	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7
A-RX	**0**	**0**	$2.11 \cdot 10^{-3}$	**0**	**0**	**0**	$\mathbf{1.62 \cdot 10^{-5}}$
K-RX	**0**	**0**	$\mathbf{1.83 \cdot 10^{-3}}$	**0**	**0**	**0**	$3.24 \cdot 10^{-4}$
RX	**0**	$1.62 \cdot 10^{-5}$	$2.23 \cdot 10^{-2}$	$1.70 \cdot 10^{-4}$	$1.62 \cdot 10^{-5}$	$1.62 \cdot 10^{-5}$	$8.10 \cdot 10^{-5}$

7.4 Final remarks and conclusions

In this chapter, experimental results related to a new AD approach for detecting small local anomalies in unknown background have been presented. The AD strategy relies upon the background log-likelihood, which is evaluated by making use of the GkNNE. The

employment of the GkNNE allows PDF estimates to follow the local data peculiarities across the data domain.

Experimental results on real multispectral data clearly highlight the benefits deriving from the employment of the proposed locally adaptive GKNNE-based AD approach. First, its non-parametric nature allowed A-RX to obtain, over the examined scenario, performance comparable to and slightly better than that obtained, at its best, by one of the most established non-parametric AD algorithms, namely K-RX. Both non-parametric A-RX and K-RX have shown to provide much better both overall and object-wise detection performance than the parametric RX algorithm, whose outcome is constrained by the validity of the LNM assumption. Secondly, the proposed A-RX has been shown to suffer very little from the variation of *k*, exhibiting similarly good detection performance across the whole range of *k* tested. On the contrary, although K-RX has still allowed, in most cases, the anomalous objects in the scene to be detected (at least comparably to RX), the use of a fixed bandwidth over the entire feature space has been shown to lead to a great sensitivity with respect to the choice of *h*.

Although in this chapter the variable-bandwidth PDF estimators are applied for detecting the anomalous objects in multispectral de facto data, the proposed AD strategy are supposed to have great potential for hyperspectral data analysis. In general, the employment of such PDF estimation methodologies for analyzing hyperspectral data require a spectral dimensionality reduction as a pre-processing step in order to prevent curse of dimensionality issues to occur during PDF estimation. Nevertheless, this latter aspect is not critical, since the hyperspectral signal usually exhibits a high degree of spectral correlation that can be exploited to represent the acquired signal in a more compact and efficient way. In fact, robust algorithms have been proposed in the literature [2][28][31] to represent the hyperspectral signal in a lower dimensional space without losing the information content related to the useful signal component (e.g. preserving both abundant and rare signal components [28][31]).

Chapter 8

8 Summary and conclusion

In this thesis work, an AD strategy is proposed based on the LRT decision rule, in which reliable estimation of the background PDF is essential to a successful detection outcome. Thus, different PDF estimators and model learning procedures have been jointly investigated within the proposed AD framework with the aim of modeling the statistical variability of hyperspectral data. According to the hyperspectral literature, there are two main approaches to background modeling and anomaly detection. In this work, both global and local methodologies have been investigated. The former aims at locating small *rare* objects that are anomalous with respect to the *global* background, which is identified by the whole image pixels. In the latter, pixels with significantly different spectral features from their *surrounding* background are detected as anomalies.

In summary, the proposed AD scheme, applicable with different PDF models and learning methods, has been dealt with in chapter 2. The analysis has been carried out focusing on the operational applicability offered for global and local AD purposes.

Statistical modeling approaches for background characterization have been addressed in chapter 3. In particular, advantages and main limitations of three well-known background models, i.e. the parametric, the semi-parametric and non-parametric, have been investigated.

In chapter 4, how to learn semi-parametric and non-parametric models for global AD purposes has been described. The analysis has been focused on GMM and StMM as mixture models, and FKDE, SPE and BE as non-parametric approaches. Specifically, methodologies developed within a Bayesian framework for automatically conducting

parameter selection of GMM, StMM, and FKDE have been considered in the first section of chapter 4. Then, the use of the *k*-NN has been explored in VKDEs in order to adapt the amount of smoothing to the local density of the data for non-parametric estimation of the multivariate PDF. In the final section of chapter 4, the improvement brought by the employment of data-adaptive VKDE with respect to the conventional FKDE has been investigated through a "toy-example".

A new local AD approach, denoted as A-RX, has been presented in chapter 5 for improving the RX and K-RX algorithms, whose outcomes are constrained by the validity of the LNM in the original input space and in a high-dimensional feature space, respectively. Specifically, the AD process is accomplished by thresholding the background log-likelihood, which is evaluated by making use of a VKDE. This latter has been chosen since it encompasses the potential, typical of non-parametric PDF estimators, in modeling data regardless of specific distributional assumptions together with the benefits deriving from the employment of bandwidths that vary across the data domain. Specifically, the employment of GkNNE has allowed the PDF estimate to be smoothed according to the local density of data samples in the feature space.

Experimental results in chapters 6 and 7 have been obtained by applying the proposed AD scheme to real hyperspectral images encompassing different AD scenarios. The aim was to evaluate and discuss the most critical issues of the different background models tested, such as their modeling ability as well as their actual utility in practical AD tasks, and to evaluate by means of several different performance measures, experimental detection performance.

The present work has shown that the proposed AD scheme is extremely valuable to automatically and adequately solve the task of detecting global anomalous objects in a given scenario. More specifically, it has been shown that different semi- and non-parametric PDF models for the image PDF coupled with specific Bayesian learning methods are effective at properly and automatically capturing the underlying structure of hyperspectral data so that the resulting PDF estimate can be successfully employed to detect spectral anomalies by means of the background log-likelihood decision rule. As regards semi-parametric models (i.e. finite mixtures), experimental results have indicated that the GMM has difficulty in properly modeling the empirical distribution of background classes in real hyperspectral data, due to the need of addressing longer tails than the Gaussian ones. On the contrary, StMM has been shown to yield a powerful model for

statistically characterizing hyperspectral data. As to the non-parametric approach, the BN bandwidth selection method examined was further investigated with regard to the impact of the user-specified parameters [K_l, K_u] on the detection outcome. Experimental results have shown the presence of a region in the (N/K_l,Λ) space where the [K_l, K_u] configurations assure similarly good detection performance. Specifically, this region was approximately identified with K_l values not lower than 2 orders of magnitude with respect to the total number N of pixels. The identification of such a recommendation confirms the robustness of the examined methodology based on selecting the bandwidth according to the distribution of the sample data variance. As regards comparative AD performance analysis, all three different AD schemes examined have been proven to be effective at detecting the anomalous objects present in the two different scenarios examined. In particular, on these data, the StMM-based scheme has provided the best detection performance in both scenarios, being capable of detecting all anomalous targets with the fewest false alarms.

Though not performing, on these data, as good as the StMM-based semi-parametric estimator, the non-parametric FKDE approach has been shown to be the most attractive approach to be applied in practical AD tasks. This is mostly because the FKDE does not rely on specific distributional assumptions. For this reason, further research work has been performed in order to improve its background estimation ability. In particular, the use of VKDEs has been proposed in order to more reliably and accurately follows the multivariate data structure with respect to the use of a fixed bandwidth. Specifically, the BE and the SPE methodologies, in which the bandwidth varies with the sample of estimation and with the sample observation, respectively, were employed in the proposed AD scheme. These methods are attractive since they allow smoothing requirements to be changed by employing small bandwidths to gain insight into highly data-structured regions and larger bandwidths for data lying in low distribution areas. Therefore, such strategies are quite sensitive to local structure peculiarities in the data, such as data clumping in certain regions and data-sparseness in others, such as in the tails. Results over real data have shown the better background PDF estimation ability of the VKDEs with respect to the FKDE, evaluated in the framework of detecting spectral anomalies in hyperspectral images. In particular, experimental analysis has shown that the variable-bandwidth PDF estimators outperform the FKDE in most cases, as to PDF approximation accuracy. Furthermore, results indicate that, whereas FKDE performance is greatly affected by the h choice, the analyzed variable bandwidth PDF estimators suffer very little from the

variation of k. This lower sensitivity in setting this user specified parameter has suggested that an automatic application of the strategy is not impaired, since the application of the proposed recommendation ($k=N^{1/2}$) to the examined data set has yielded good detection performance.

Concerning the local AD strategies, experimental results on real data are strongly in favor of the proposed locally adaptive GKNNE-based AD approach. This outcome is substantiated by performance comparable to and significantly better than that obtained by the classical local AD algorithms tested. Moreover, the proposed A-RX has been shown to suffer very little from the variation of k, the only user specified parameter.

It is worth noting that, although in this work the variable-bandwidth PDF estimators are applied only for enhancing the separation of anomalous objects with respect to the background, their high flexibility and adaptability suggest to employ variable bandwidths in other tasks of remote-sensing image analysis requiring reliable PDF estimates, such as spectral signature based target detection, image clustering, and many others. Nevertheless, it should be noted that the non-parametric PDF estimators can be computationally expensive in practical circumstances and, in order to fully exploit their great potential, attempts to increase the computational efficiency are needed.

In the light of the results achieved, this thesis work has shown that the proposed AD architecture is an extremely effective strategy in detecting the rare anomalous objects present in the scene. From a general point of view, the research carried out in this thesis resulted in the definition of novel methodological and technical contributions in relation to some of the more critical problems present in unsupervised AD literature. Moreover, it resulted in the implementation of processing tools suitable to be adopted in real applications.

Bibliography

[1] Acito N., G. Corsini, and M. Diani, "Statistical Analysis of Hyper-Spectral Data: A Non-Gaussian Approach", EURASIP Journal on Advances in Signal Processing, vol. 2007, pp 1-10, 2007.

[2] Acito, N., M. Diani, and G., Corsini, "A new algorithm for robust estimation of the signal subspace in hyper-spectral images in presence of rare signal components," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3844-3856, 2009.

[3] Alpaydin E., *Introduction to Machine Learning*, The MIT Press, 2004.

[4] Archambeau C. and M. Verleysen, "Robust Bayesian Clustering," *Neural Networks*, vol. 20, pp. 129-138, 2007.

[5] Archambeau C., J.A. Lee, and M. Verleysen, "On Convergence Problems of the EM Algorithm for Finite Gaussian Mixtures," Proc. 11th European Symposium on Artificial Neural Networks (ESANN '03), pp. 99-106, 2003.

[6] Bishop C.M., *Pattern Recognition and machine learning*, Springer, 2006.

[7] Bors A.G., Nasios, N., "Kernel Bandwidth Estimation for Nonparametric Modeling," *IEEE Trans. Systems, Man, and Cybernetics*, Part B: Cybernetics, vol. 39, no. 6, pp. 1543-1555, 2009.

[8] Breiman L., W. Meisel, and E. Purcell, "Variable Kernel Estimates of Multivariate Densities", *Technometrics*, American Statistical Association and American Society for Quality Stable, vol. 19, no. 2, pp. 135-144, 1977.

[9] Carlotto M.J., "A cluster-based approach for detecting man-made objects and changes in imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 374-387, 2005.

[10] Chang C.I and D. Heinz, "Constrained subpixel target detection for remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1144-1159, 2000.

[11] Comaniciu D., V. Ramesh, and P. Meer, "The Variable Bandwidth Mean Shift and Data-Driven Scale Selection," Proc. IEEE Int. Conf. Computer Vision (ICCV'01), vol. 1, pp. 438-445, 2001.

[12] Constantinopoulos C. and A. Likas, "Unsupervised learning of Gaussian mixtures based on variational component splitting," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 745-755, 2007.

[13] Constantinopoulos C., and A. Likas,, "Image modeling and segmentation using incremental Bayesian mixture models", Proc. 12th Int. Conf. Computer Analysis of Images and Patterns, Springer, pp. 596-603, 2007.

[14] Corduneanu A. and C. Bishop, "Variational Bayesian model selection for mixture distributions," Proc. International Conference on Artificial Intelligence and Statistics, pp. 27-34, 2001.

[15] Cremers D., T. Kohlberger, and B. Schölkopf, "Shape statistics in kernel space for variational image segmentation," *Pattern Recognit.*, vol. 36, pp. 1929-1943, 2003.

[16] Davies D. L., and Donald W. Bouldin, "A Cluster Separation Measure", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224-227, 1979.

[17] Duda R. O., P. E. Hart, and D. G. Stork, *Pattern Classification*, New York: Wiley, 2000.

[18] Duong T. and M.L. Hazelton, "Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation," *Journal of Multivariate Analysis*, vol. 93, pp. 417-433, 2005.

[19] Duong T. and M.L. Hazelton, "Cross-validation bandwidth matrices for multivariate kernel density estimation", *Scandinavian Journal of Statistics*, vol. 32, pp. 485-506, 2005.

[20] Duong T., and M.L. Hazelton, "Plug-in bandwidth selectors for bivariate kernel density estimation, Journal of Nonparametric Statistics," vol. 15, pp. 17-30, 2003.

[21] Farrand W. H. and J. C. Harsanyi, "Mapping the distribution of mine tailings in the Coeur D' Alene River Valley, Idaho, through the use of a constrained energy minimization technique," Remote Sens. Environ., vol. 59, pp. 64-76, 1997.

[22] Fukunaga K., *Introduction to Statistical Pattern Recognition*, New York Academic, 1972.

[23] Härdle W., M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*, Berlin: Springer-Verlag, 2004.

[24] Harsanyi J. C. and C. I. Chang, "Detection of low probability subpixel targets in hyperspectral image sequences with unknown backgrounds," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 7, pp. 779-785, 1994.

[25] http://aviris.jpl.nasa.gov/

[26] Hwang J.-N.. S.-R. Lay., A. Lippman, "Nonparametric multivariate density estimation: a comparative study", *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2795-2810, 1994.

[27] Hytla P.C., R. C. Hardie, M. T. Eismann, and J. Meola, "Anomaly detection in hyperspectral imagery: comparison of methods using diurnal and seasonal data," *J. Appl. Remote Sens.*, vol. 3, pp. 033546-033546-30, 2009.

[28] Hyvarinen A., J. Karhunen, and E. Oja, *Independent Component Analysis. Hoboken*, NJ: Wiley, 2001.

[29] Izenman A., "Recent developments in nonparametric density estimation," *J. Amer. Statist. Assoc.*, vol. 86, pp. 205-224, 1991.

[30] Kay S.M., *Fundamentals of Statistical Processing: Detection Theory*, Ed. US: Prentice Hall, 1998.

[31] Kuybeda O., D. Malah, and M. Barzohar, "Rank estimation and redundancy reduction of high dimensional noisy signals with preservation of rare vectors," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5579-5592, 2007.

[32] Kwon H. and N. M. Nasrabadi, “Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388-397, 2005.

[33] Kyrgyzov I. O., Kyrgyzov O. O., Maitre H., Campedel M., “Kernel MDL to Determine the Number of Clusters”, Lecture notes in Computer Science, vol. 4571, pp. 203-217, 2007.

[34] Loader C.L., “Bandwidth selection: classical or plug-in?,” *Ann. Stat.*, vol. 27, no.2, pp. 415-438, 1999.

[35] Loftsgaarden D.O. and C.P. Quesenberry, “A nonparametric estimate of a multivariate density function,” *Ann. Math. Statist.*, vol. 36, pp. 1049-1051, 1965.

[36] Manolakis D. and G. Shaw, “Detection algorithms for hyperspectral imaging applications,” *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 29–43, 2002.

[37] Manolakis D., G. Shaw, and N. Keshava, “Comparative analysis of hyperspectral adaptive matched filter detectors,” Proc. SPIE, vol. 4049, pp. 2-17, 2000.

[38] Manolakis D., Marden D., “Non Gaussian Models for Hyperspectral Algorithm Design and Assessment”, Proc. IEEE International Geosci. Remote Sens. Symp. (IGARSS), vol. 3, pp. 1664-1666, 2002.

[39] Manolakis D., Marden D., Kerekes J., Shaw G., “On the Statistics of Hyperspectral Imaging Data”, Proc. SPIE, vol. 4381, pp. 308-316, 2001.

[40] Marden D. and D. Manolakis, “Modeling hyperspectral imaging data,” Proc. Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX, SPIE, vol. 5093, pp. 253-262, 2003.

[41] Margalit A., Reed I. S., Gagliardi R. M., “Adaptive Optical Target Detection Using Correlated Images”, *IEEE Trans. Aerosp. Electron. Syst.*, vol. 21 no. 3, pp. 46-59, 1985.

[42] Matteoli S., M. Diani, and G. Corsini, “A tutorial overview of anomaly detection in hyperspectral images,” *IEEE Aerosp. Electron. Syst. Mag. Tutorials*, vol. 25, no. 7, pp. 5-28, 2010.

[43] Matteoli S., N. Acito, M. Diani, and G. Corsini, "An Automatic Approach to Adaptive Local Background Estimation and Suppression in Hyperspectral Target Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, 2011.

[44] McKenzie P., Alder M., "Selecting the Optimal Number of components for a Gaussian Mixture Model", Proc. IEEE International Symp. Information Theory, vol. 1, pp. 393, 1994.

[45] McLachlan, G. and D. Peel. *Finite Mixture Models*, John Wiley & Sons, New York, 2000.

[46] Mittal A. and N. Paragios, "Motion-based background subtraction using adaptive kernel den-sity estimation," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2004.

[47] Parisi, G. (1988). Statistical Field Theory. Addison-Wesley.

[48] Reed I.S. and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, 1990.

[49] Robey F.C., D.R. Fuhrman, E.J. Kelly, R. Nitzberg, "A CFAR Adaptive Matched Filter Detector," *IEEE Trans. on Aerosp. Electron. Syst.*, vol. 28, no. 1, pp. 208-216, 1992.

[50] Sain S., "Multivariate Locally Adaptive Density Estimates," *Computational Statistics and Data Analysis*, vol. 39, pp. 165-186, 2002.

[51] Sain S.R., Baggerly, K.A. and Scott, D.W, "Cross-validation of multivariate densities", *J. Amer. Statist. Assoc.*, vol. 89, pp. 807-817, 1994.

[52] Schölkopf B., and A.J. Smola, *Learning with kernels: Support vector machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2001.

[53] Scott, D.W., *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York, 1992.

[54] Silvermen B. W., *Density Estimation for Statistics and Data Analysis*, New York Chapman and Hall, 1986.

[55] Stein D.W.J., S.G. Beaven, L.E. Hoff, E.M. Winter, A.P. Schaum, and A.D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58-69, 2002.

[56] Stoica P. and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, pp. 36–47, 2004.

[57] Terrell G. R., D. W. Scott, "Variable Kernel Density Estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236-1265, 1992.

[58] Terrell, G.R., Scott, D.W., "Variable kernel density estimation," *Ann. Statist.*, vol. 20, pp. 1236-1265, 1992.

[59] Theodoridis S., K. Koutroumbas, and A. Pikrakis, *Introduction to Pattern Recognition: A Matlab Approach*, Academic Press.

[60] Tukey P.A. and Tukey J.W., *Data-driven view selection: agglomeration and sharpening*, V. Burnett Ed., Interpreting Multivariate Data. Wiley, Chichester, 1981.

[61] Tzikas D.G., A.C. Likas, and N.P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol.25, no. 6, pp. 131-146, 2008.

[62] Veracini T., S. Matteoli, Diani M., and G. Corsini, "A locally adaptive background density estimator: an evolution for RX-based anomaly detectors", submitted to *IEEE Trans. Geosci. Remote Sens. Lett.*, 2011.

[63] Veracini T., S. Matteoli, Diani M., and G. Corsini, "Background Density Non-Parametric Estimation with Data-Adaptive Bandwidths for the Detection of Anomalies in Multispectral Imagery", submitted to *IEEE Trans. Geosci. Remote Sens.*, 2011.

[64] Veracini T., S. Matteoli, Diani M., and G. Corsini, "Models and Methods for Automated Background Density Estimation in Hyperspectral Anomaly Detection", submitted to *IEEE Trans. Geosci. Remote Sens.*, 2011.

[65] Veracini T., S. Matteoli, Diani M., and G. Corsini, "Robust Hyperspectral Image Segmentation Based on a Non-Gaussian Model," Proc. International Workshop on Cognitive Information Processing (CIP), pp. 192 – 197, 2010.

[66] Veracini T., S. Matteoli, Diani M., and G. Corsini,"A Novel Anomaly Detection Scheme for Hyperspectral Images Based on a Non-Gaussian Mixture Model," Proc. IEEE Gold, 2010.

[67] Veracini T., S. Matteoli, Diani M., and G. Corsini,"Fully Unsupervised Learning of Gaussian Mixtures for Anomaly Detection in Hyperspectral Imagery," Proc. IEEE Conference on Intelligent Systems Design and Applications (ISDA), pp. 596–601, 2009.

[68] Veracini T., S. Matteoli, M. Diani, and Corsini, G., "Non-parametric Framework for Detecting Spectral Anomalies in Hyperspectral Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 4, 2011.

[69] Veracini T., S. Matteoli, M. Diani, and G. Corsini, "A non-parametric approach to anomaly detection in hyperspectral images", Proc. SPIE's International Symposium, Remote Sensing Europe (ERS), SPIE, pp. 192-197, 2010

[70] Veracini T., S. Matteoli, M. Diani, and G. Corsini, "An Anomaly Detection Architecture based on a Data-Adaptive Density Estimation", Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-4, 2011.

[71] Veracini T., S. Matteoli, M. Diani, G. Corsini, and U. de Ceglie "A spectral anomaly detector in hyperspectral images based on a non-Gaussian mixture model", Proc. IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-4, 2010.

[72] Wand M. P. and M. C. Jones, "Kernel Smoothing," *Monographs on Statistics and Applied Probability*, vol. 60, Chapman and Hall, London, 1995.

[73] Wand M.P and M.C. Jones, "Multivariate plug-in bandwidth selection," *Comput. Statist.*, vol. 9, pp. 97-117, 1994.

[74] Yeung K. Y., Fraley C., Murua A., Rafery A. E., Ruzzo W. L., "Model- Based Clustering and Data Transformations for Gene Expression Data", *Bioinformatics*, vol. 17 no. 10, pp. 977-987, 2001.

[75] Yu X. and I.S. Reed, "Comparative performance analysis of adaptive multispectral detectors," *IEEE Trans. Signal Process. Mag.*, vol. 41, no. 8, pp. 2639–2656, 1993.