Multiple-feature Kernel-based Probabilistic Clustering for Unsupervised Band Selection

Marco Bevilacqua, Yannick Berthoumieu

Multiple-feature Kernel-based Probabilistic Clustering for Unsupervised Band Selection

Marco Bevilacqua, and Yannick Berthoumieu, *Member, IEEE*.

Abstract

This paper presents a new method to perform unsupervised band selection (UBS) with hyperspectral data. The method provides a probabilistic clustering approach. The band images are clustered in the image space by computing their posterior class probability. Then, for each cluster, the band exhibiting the highest probability of belonging to it is selected as cluster exemplar. More particularly, the proposed method falls into information-maximization clustering methods, where the posterior class probability is modeled and the parameters of the models are derived by maximizing the information between the data and the unknown cluster labels. In this context, we propose a new image representation for hyperspectral images, based on the first and second order statistics of multiple image features. We refer to this representation as multiple-feature local statistical descriptors (MLSD). The descriptors are computed w.r.t. regular grids, and a special pixel selection procedure reduces the number of samples within each block of the grid. A kernel-based model that embeds the MLSD is then proposed for the posterior class probability. The model is finally optimized according to an information-maximization criterion. We conduct several experiments to determine the best parameters for the proposed approach and compare the latter with other state-of-the-art UBS methods. Quantitative evaluations show that, by employing our band selection method, higher performance in terms of classification accuracy and endmember extraction can be achieved in comparison with the state of the art.

Index Terms

Hyperspectral imaging, band selection, image representation, clustering, classification.

I. Introduction

HYPERSPECTRAL imaging offers the possibility of collecting information of a target scene across the electromagnetic spectrum, in narrow and contiguous bands, typically ranging from visible to long-infrared spectra. [1] In practice, we obtain what is named a *hyperspectral cube*, i.e. a collection of hundreds of 2-D images, each one corresponding to a specific spectral band. The availability of a large number of bands with high spectral resolution has been shown to lead to enhanced results in recognition of materials, objects and land cover classes. Working with hyperspectral images has a positive impact in several practical applications, such as remote sensing [2], medical imaging [3] and biological analyzis [4]. However, the large information diversity within a hyperspectral image brings inevitability some drawbacks. First, the computational costs in storing, processing, and analyzing such data can be very high. Second, due to the fact the neighboring bands exhibit a high dependency [5], some undesirable numerical phenomena may conduct to decreasing the accuracy of algorithms. This is linked to multicolinearity, i.e. having redundant information in the predictors. It is in fact established that input features suffering of severe multicolinearity can yield, especially in the case of linear models, solutions that are wildly varying and possibly numerically unstable [6, ch. 9]. Finally, hyperspectral data can suffer of the so-called "curse of dimensionality" [7]. If we consider the spectral signature, each pixel of a hyperspectral image can be seen as a vector lying in a high-dimensional space. It can therefore happen that the available data (let imagine that only a part of the image pixels needs to be processed) become sparse in relation to the volume of the high-dimensional space. This sparsity can assimilate to a problem, as it undermines the statistical significance and reliability of the results.

For all the mentioned reasons, reducing the number of spectral channels is a useful procedure in hyperspectral imaging. To reduce the spectral dimension, two approaches are possible corresponding respectively to feature reduction (or extraction) and band selection. The main purpose of feature reduction is to map the initial high-dimensional spectral vectors to low-dimensional features, so that as much information as possible from the original vectors is kept. Typical feature reduction algorithms employed with hyperspectral data include principal component analysis (PCA) [8], linear discriminant analysis (LDA) [9], and their respective kernelized versions [10], [11]. Other recent feature reduction methods borrow concepts and algorithms from the manifold learning theory [12]. Band selection, rather than transforming the initial spectral vectors, aims at selecting a subset of informative and distinctive spectral bands. Once a subset of bands is identified from a full data set and for a given task, band selection can be immediately performed on new similar data for similar task. A second advantage comes with the physical interpretation of the image that is kept, which is composed of original bands. This is not always the case for feature reduction: although some works have been conducted to perform more "meaningful" extractions [13], results are often

M. Bevilacqua, and Y. Berthoumieu are with Bordeaux INP, Université de Bordeaux and CNRS, IMS laboratory, UMR 5218, F-33405 Talence, France. E-mail: marcobev@inwind.it and yannick.berthoumieu@ims-bordeaux.fr.

This study has been carried out with the financial support from the PharmaSense project funded by the French Program "Fonds Unique Interministériel – Nouvelle Aquitaine".

Manuscript received July XX, 2018.

difficult to directly interpret. A final advantage for band selection is a consequence of the preserved physical meaning. Once the band selection stage is performed and real wavelengths are obtained as an output, it is possible to design less expensive ad-hoc multispectral systems (3 to 10 bands). Band selection methods for hyperspectral images can be further classified as supervised and unsupervised ones. Supervised methods [14], [15], [16] make use of training labeled samples to train models. Compared to unsupervised methods, they typically lead to better performance, as they leverage prior information. However, *unsupervised band selection* (UBS) methods [17], [18], [19] can be preferable for their flexibility and robustness. They are robust in the sense that they do not rely on the presence and quality of labeled samples, which might be difficult and time-consuming to collect. Instead, they aim at defining class-independent procedures that exploit the inherent content of the hyperspectral image.

In this paper we focus on the UBS problem. In the literature a large variety of UBS methods are proposed. Two main categories can however be identified: *ranking-based* and *clustering-based* methods. Ranking-based methods aim at ranking the bands according to a certain metric that quantifies their importance, e.g. variance [17], information divergence as a measure of non-Gaussianity [20], linear independence [18] and mutual information [21]. Clustering-based UBS methods [19], [22], [23], [24], [25], [26] tend to exploit similar concepts but they follow an optimization strategy to group bands into clusters. One representative band per cluster is finally taken to form the subset of selected bands. UBS methods leveraging graph theory [27], [28] can also be included in this category. Some other studies that attempt at combining ranking and clustering techniques have also been conducted [29], [30].

In this paper we present a novel UBS method based on a clustering approach. It raises from two main observations.

- If we observe the overall appearance of a hyperspectral cube we can denote that bands tend to form clusters (Fig. 1). However the transitions between these clusters are often smooth. This would suggest the adoption of a probabilistic (or soft) clustering strategy. As far as we know, the only UBS method following such approach is presented in [31].
- Conventional UBS methods only consider spectral magnitude as an image feature. However, numerous articles show that classification with hyperspectral images also benefits from taking into account spatial information [32]. This suggests that spatial information could be used in the UBS task. In this context, we propose to use probabilistic clustering method based on local spatial statistics from augmented spatial/spectral feature spaces. For illustration purpose, in our experiments we select spatial filtering outputs as spatial features. However, more advanced spatial features can be considered, such as structural filtering, morphological profiles, random fields, or other spatial priors.

Our proposed method then follows a probabilistic clustering approach and is designed in order to leverage multiple features. It is based on a multi-kernel probabilistic model to incorporate different image features (besides spectral intensity, spatial features).

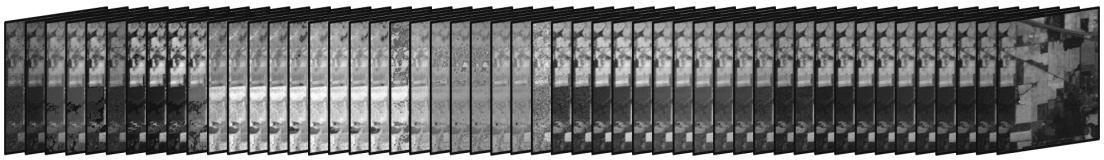


Fig. 1. Subset of the bands (one over three) composing the *Kennedy Space Center* data set. Clusters are recognizable, yet they exhibit smooth transitions.

The main contributions of the paper can be summarized as follows.

- New joint spectral/spatial descriptors: for a better representation of hyperspectral images, *multi-feature local statistic descriptors* (MLSD) are proposed. As MLSD, we consider first and second order statistics, i.e. local mean vectors and covariance matrices computed on a finite spatial neighborhood centered on current nodes from regular grids. Covariance is used as a direct technique to fuse several spectral/spatial features. On one hand, the covariance-based fusion technique is exploited to characterize local spectral/spatial geometric structures for texture characterization within hyperspectral images. On the other hand, mean vectors code the local trend of the joint spectral/spatial amplitudes.
- A special *pixel selection* (PS) procedure: this procedure is devoted to robust statistical feature estimation of local mean vectors and covariance matrices. A regular spatial grid along the whole image domain is defined. For each node of this grid, features within the corresponding local spatial neighborhood centered on the node are extracted. The procedure aims at clearing out potential outliers in the feature set. The selected pixels are meant to be the most reliable ones for computing local statistics.
- Kernel-based probabilistic modeling: considering MLSD descriptors, a kernel-based model for the posterior class probability is proposed. The kernel smoothing method is such that each local statistic is embedded in a Epanechnikov kernel according to adapted similarity measure. The kernel-based approach enables the possibility of performing non-linear clustering.

The remainder of the paper is organized as follows. Sections II and III are devoted to related works. In the former, we list existing clustering-based methods for UBS. In the latter, we review, in a general way, probabilistic clustering, which has been seldomly considered in the context of hyperspectral imaging. Section IV details our proposed clustering-based approach, which relies on new statistical descriptors, specially computed, and a kernel-based posterior probability model. Before drawing

conclusions, we consider two experimental settings. In Section V, quantitative evaluations are performed to derive the best parameters for the proposed method and compare it with other state-of-the-art UBS algorithms w.r.t a classification problem. In Section VI, the problem of endmember extraction is considered and our UBS algorithm is compared to other UBS algorithms in terms of spectral similarity.

II. Unsupervised band selection via deterministic clustering

A. Notation adopted

Before revising the clustering approach for unsupervised band selection, let us specify the notation that will be used throughout the paper. We denote a hyperspectral cube as a 3-dimensional matrix $\mathbf{H} \in \mathbb{R}^{h \times w \times n}$, composed by n images (bands) of dimension $h \times w$. The matrix $\mathbf{H}$ can be further arranged as a bi-dimensional matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where each column refers to one particular band and results from stacking all its pixels; i.e. $m = h \times w$ is the total number of pixels in each single band. The band selection task then consists in finding a subset of k columns from $\mathbf{X}$, i.e a subset of bands $\mathcal{B}$ of cardinality $|\mathbb{B}| = k < n$.

B. Existing clustering-based UBS methods

A family of unsupervised UBS methods is based on clustering approaches. All bands composing the hyperspectral cube (the n column vectors of the matrix $\mathbf{X}$) are considered as individual points and clustered in the shared m-dimensional space. Clustering-based UBS methods generally consist of two steps: the proper clustering algorithm and the band selection step.

1) Clustering: the band space is partitioned into k clusters, i.e. each vector (band) $\mathbf{x}_i$ is given a label $y_i \in \{1, \dots, k\}$.
2) Band selection: for each cluster, a unique exemplar that identifies a specific spectral band is kept. Generally, the selected exemplars are the cluster *medoids*.

Clustering-based UBS methods are typically based on existing "general-purpose" clustering algorithms. Martínez-Usó *et al.* [19] apply Ward's method for agglomerative hierarchical clustering to group bands, in order to minimize the intra-cluster variance and maximize the inter-cluster variance. In [23] Ahmad *et al.* adapt the well-known k-means algorithm to the case of hyperspectral data. By proposing a modification of k-means, in [25] Yuan *et al.* present a dual clustering procedure, where novel image features describing pixel contextual information are jointly clustered with the hyperspectral "raw" features. In [26], Zhang *et al.* combine fuzzy clustering with particle swarm optimization to reduce the sensitivity of the clustering algorithm to its initialization. Some other methods do not find clusters explicitly (yet they could be easily outlined), but they aim at directly determining the cluster representatives. Methods based on *affinity propagation* [22], [24] belong to this category. Here, a phase where real-valued messages are exchanged between all data points automatically unveils the data exemplars. Similarly, other methods aim at directly retrieving cluster exemplars by means of empirical measures [33], [30]. Unlike affinity propagation or other clustering algorithms, they do not follow any optimization strategy, but they define indicators to quantify the likelihood of each band to be an exemplar. The rational of these indicators is that in cluster-arranged data the exemplars should be data points with high local density, located sufficiently far away from other high-density points.

While all these methods already allow to perform, explicitly or not, clustering of hyperspectral data and to intelligently select a set of bands as the cluster exemplars, they lack of a probabilistic interpretation. We instead believe that, hyperspectral data, since consisting in whole spectra of contiguous, progressively varying, bands, is more adapted to be clustered in a "soft" way. In Section III, before detailing our method, which relies on a probabilistic clustering approach, we briefly revise the strategies typically adopted in the literature for such methods.

III. Probabilistic clustering

Probabilistic clustering methods belong to the family of clustering methods based on cost function optimization [34]. Here the cost function is constructed on the basis of random vectors, and assignment to clusters follows probabilistic arguments, in the spirit of the Bayesian classification, as opposed to more exploratory, heuristic or algorithmic approaches.

Probabilistic clustering methods can be broadly classified into three categories, according to the driving optimization criterion: *likelihood-maximization* methods, *posterior-maximization* methods, and *information-maximization* methods. Probabilistic methods based on likelihood maximization particularly arise from generative models. We often talk in this case about *generative clustering*, where the probability of the data is defined as a mixture of k probability density functions: $p(\mathbf{x}; \beta, \pi) = \sum_{y=1}^{k} p(\mathbf{x}|y; \beta)p(y; \pi)$, with β and π parameters. The parameters of the model are generally determined by maximizing the marginal likelihood, in order to find the probability distribution $p(x)$ that most likely has generated the observed data. Optimization can be performed with gradient methods, or, by regarding the class labels as latent variables, via the well-known *expectation-maximization* (EM) algorithm [35]. While generative clustering methods, based on likelihood maximization, are statistically well-founded, they present the difficulty that the cluster density models have to be defined in advance. This leads to a lack of flexibility. Moreover, in high dimensions, the class of generative models is basically restricted to mixture of Gaussians. As a results, as the the data to model is rather complex, a large number of mixture components is typically required. [36]

Posterior-maximization probabilistic methods aim at directly maximizing the posterior probability of class labels $p(y|\mathbf{x})$. The latter is generally expressed by means of the cluster-wise data distribution $p(\mathbf{x}|y)$. Parametric or non-parametric Bayesian approaches are possible (e.g. [37]). In both cases, density models for each cluster, as for generative models, need to be preliminarily specified. This often restricts to the use of Gaussian models.

Finally, probabilistic clustering methods based on information maximization sensitively differ from the first two categories of methods, due to their *discriminative* nature. Rather than modeling categories of data explicitly, they leverage conditional-probability models $p(y|\mathbf{x};\boldsymbol{\alpha})$ that depend on a set of parameters $\boldsymbol{\alpha}$. Those models act as *encoders*, in the sense that they define a mapping between the data $\mathbf{x}$ and the latent code, i.e the cluster labels. The optimization criterion is then maximizing the information between data and cluster labels, with respect to the encoder parameters. This effectively defines a discriminative unsupervised optimization framework. Once the parameters of the model are determined, cluster assignments $\{y_i\}_{i=1}^n$ are easily derived as:

$$\{y_i\}_{i=1}^n = \arg\max_y p(y|\mathbf{x};\hat{\boldsymbol{\alpha}}) \ . \tag{1}$$

As for the encoders models used, in [36] the Gaussian class posterior model $p(y=j|\mathbf{x}) \propto \exp\{-\|\mathbf{x}-\mathbf{w}_j\|_2^2/s_j + \mathbf{b}_j\}$ is proposed; mutual information is then chosen as the optimization criterion. In [38], Krause *et al.* propose a logistic regression function for the conditional model ($p(y=j|\mathbf{x}) \propto \exp\{\mathbf{w}_j^\mathsf{T}\mathbf{x} + b_j\}$) , and a regularized version of the mutual information as objective function. Sugiyama *et al.* [39], [40] introduce a novel probabilistic clustering method based on information maximization, where the encoder is a kernel-based conditional probability model. Moreover, they propose to use a variant of mutual information called *Squared-loss Mutual Information* (SMI) as information metric:

$$\text{SMI} := \frac{1}{2}\int \sum_{y=1}^{k} p(\mathbf{x})p(y)\left(\frac{p(\mathbf{x},y)}{p(\mathbf{x})p(y)} - 1\right)^2 \mathrm{d}\mathbf{x} \ . \tag{2}$$

SMI is the *Pearson divergence* from the joint probability $p(\mathbf{x},y)$ to the product of the two densities $p(\mathbf{x})p(y)$, and exhibits a computational advantage. As easily provable, we can in fact approximate (2) to:

$$\text{SMI} \simeq \frac{1}{2n}\sum_{j=1}^{n}\sum_{y=1}^{k}\frac{1}{p(y)}\left[p(y|\mathbf{x}_j)\right]^2 - \frac{1}{2} \ . \tag{3}$$

The expression obtained is quadratic w.r.t. the posterior probability. This is leveraged in [39], [40] to analytically compute a solution when maximizing SMI. The resulting clustering method is referred to as SMI-based clustering (*SMIC*). To the best of our knowledge the first method that relies on probabilistic clustering to perform hyperspectral band selection is presented in [31]. The method utilizes the same "bricks" as SMIC, by adapting them to the hyperspectral problem: a kernel-based conditional model and SMI as information measure. Our proposed method builds on this previous work, with which it shares the methodology of the approach. New image descriptors based on local statistics (the MLSD) are nevertheless introduced, as well as a novel strategy to incorporate these descriptors into the posterior class probability model.

In general terms, an UBS method based on probabilistic clustering must define three key ingredients:

1) The image descriptors used and a way to compute inter-band distances based on them;
2) The probabilistic model adopted;
3) The optimization strategy chosen to derive the parameters of the model

In the next section we provide the details of our proposed approach, by focusing on each of these three aspects.

IV. Proposed method

In this section we detail our method for unsupervised band selection (UBS). As previously mentioned, the proposed approach consists of three steps:

1) Extraction of local statistical descriptors from multiple features (the MLSD), which will also imply a *pixel selection* procedure;
2) Definition of a kernel-based parametric model for the posterior class probability;
3) Optimization of the model parameters via information maximization.

The steps listed above are detailed, respectively, in the following three subsections. Fig. 2 depicts schematically the proposed approach. The first step allows to compute statistical descriptors (mean and covariance) of multiple-feature vectors. The latter are the input of a probabilistic modeling block that incorporates the extracted features into a kernel-based model with parameters $\boldsymbol{\alpha}$. Optimization via information maximization allows to compute an optimal value for the parameter vector $\hat{\boldsymbol{\alpha}}$, thus determining the class-wise data distribution.

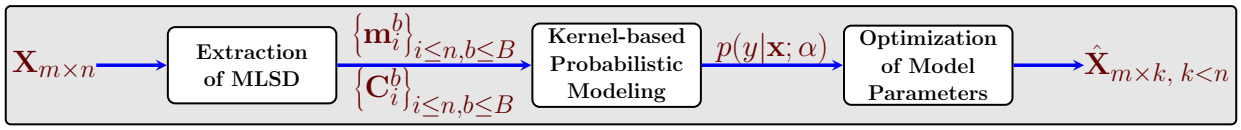


Fig. 2. Scheme of the proposed approach: multiple-feature local statistical descriptors (MLSD) are computed for each band of the hyperspectral image and fed to a posterior class probability model. The last step concerns the optimization of the parameters of the model.

A. *Extraction of multiple-feature local statistical descriptors (MLSD)*

Hyperspectral data, besides spectral redundancy, contains a considerable amount of spatial information. Each band image can count up to hundreds of thousands of pixels. Therefore, to perform unsupervised band selection (UBS), where pixels share the same importance, it is crucial to summarize such information in order to provide a concise representation of each band. In addition, several works in the literature show the benefit of enriching "raw" spectral magnitude with other spatial features, to perform an efficient spectral-spatial classification (e.g. [41], [42], [43]) or, though to a lesser extent, even for the UBS task (e.g. [27], [25]).

In Section IV-A1 we introduce new descriptors for hyperspectral images. They are constructed by computing first and second order statistics over multiple-feature vectors, thus being capable to describe spatial variability too. They are computed according to regular grid sampling. For each band we then have a limited number of descriptors, each one bringing a local information. The section serves a formal presentation of the descriptors adopted; implementation details will then be provided in Section V-C1. In Section IV-A2 we also introduce a crucial procedure for the construction of our descriptors, which we call *pixel selection* (PS). Thanks to that, the sample population for computing statistics is reduced by carefully selecting the most stable pixels of each block.

1) Multiple-feature descriptors: Let $\mathbf{H}_i \in \mathbb{R}^{h\times w}$ be a single band image composing the hyperspectral cube (before reshaping it to the vector form $\mathbf{x}_i$). If M different per-pixel features can be computed, we then have M feature matrices $\mathbf{G}_i^1, \mathbf{G}_i^2, \dots \mathbf{G}_i^M$. The feature matrix $\mathbf{G}_i^1$ could simply be equal to $\mathbf{H}_i$, i.e. composed of the raw spectral magnitude values, the other matrices being related to other images features (e.g. spatial features to extract contours). For a single pixel located at a certain position $p = (x_p, y_p)$, we can then define the M-dimensional feature vector $\mathbf{f}_{i,p} = \left[G_{i,p}^1, G_{i,p}^2, \dots, G_{i,p}^M\right]^\mathsf{T}$, with all feature values stacked. These feature vectors are the bases of our statistical descriptors, which provide a concise representation of them. We refer to these new descriptors as multiple-feature local statistical descriptors (MLSD)

The computation of statistics is performed locally according to regular grid sampling. We then have per-block descriptors. Let suppose that each band image can be divided into B non-overlapping blocks of size $s \times s$. Let indicate the pixel sets of the B blocks as $\{\mathcal{P}_b\}_{b=1}^B$, i.e. $\mathcal{P}_b$ is the set of pixel locations (2-D coordinates) of the b-th block. The *mean descriptor* related to the b-th block of the i-th band image is then defined as:

$$\mathbf{m}_i^b = \frac{1}{N_b} \sum_{p\in\mathcal{P}_b} \Gamma_{i,p} \mathbf{f}_{i,p} \ , \tag{4}$$

where $\mathbf{\Gamma}_i \in \mathbb{R}^{h\times w}$ is a binary mask, related to the i-th image band $\mathbf{H}_i$, indicating the pixel to take into account in the computation of statistics (if $\Gamma_{i,p} = 1 \quad \forall\, p$, then all pixels are taken into account); and N_b is the number of pixels taken in the b-th block. In terms of the binary mask of activated pixels $\mathbf{\Gamma}_i$, we then have $N_b = \sum_{p\in\mathcal{P}_b} \Gamma_{i,p}$.

With the mean vector defined, we can also provide the definition of a per-block *covariance descriptor*:

$$\mathbf{C}_i^b = \frac{1}{N_b} \sum_{p\in\mathcal{P}_b} \Gamma_{i,p} (\mathbf{f}_{i,p} - \mathbf{m}_i^b)(\mathbf{f}_{i,p} - \mathbf{m}_i^b)^\mathsf{T} \ . \tag{5}$$

With per-block descriptors defined, we need now to be able to define distances between two band images $\mathbf{x}_i$ and $\mathbf{x}_j$ (vectorized versions of, respectively, $\mathbf{H}_i$ and $\mathbf{H}_j$). A *mean-wise inter-band distance* can be defined as the average Euclidean distance between all corresponding mean vectors:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \|\mathbf{m}_i^b - \mathbf{m}_j^b\|_2^2 \ . \tag{6}$$

For the inter-band distance based on covariance descriptors we first need to define a distance between two covariance matrices. Covariance descriptors, as we defined them, are symmetric positive definite (SPD) lying of Riemannian manifolds. Following [44], we select the Log-Euclidean Riemannian Metric (LERM) as a distance metric for such matrices:

$$d_{\mathrm{LE}}(S_1, S_2) = \|\mathrm{Log}\, S_1 - \mathrm{Log}\, S_2\|_F^2 \ , \tag{7}$$

where Log is the principal matrix logarithm and $\|\|_F$ is the Frobenius norm of a matrix defined as

$$\|\mathbf{A}\|_F = \sqrt{\textstyle\sum_i \sum_j |a_{ij}|^2} \ . \tag{8}$$

The metric in (7) approximates the geodesic distance over a Riemannian geometry as an Euclidean distance in the logarithmic domain. It then has all the properties of an Euclidean metric. Secondly, when "injected" into a positive definite kernel, it has been proved that the positiveness of the kernel is preserved [45]. A *covariance-wise inter-band distance* can finally be defined as the average distances between the per-block covariance descriptors:

$$d_C(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \mathrm{d}_{\mathrm{LE}}(\mathbf{C}_b^i, \mathbf{C}_b^j) \ . \tag{9}$$

Further details about the type of features and parameters chosen, i.e. how the M feature matrices $\mathbf{G}_i^1, \mathbf{G}_i^2, \dots \mathbf{G}_i^M$ are defined, are provided in Section V-C1.

2) Pixel selection: In equations (4) and (5), $\mathbf{\Gamma}_i$ represents a binary mask for the i-th band image corresponding to a sub-partition from the considered initial pixel set. The aim is to retain a subset of pixels in order to have a robust estimation of local statistics. The selection is performed at the block level, i.e. for each block we select a subset of "stable" pixels used to compute the statistics of that block. The interest of this selection has been motivated in clustering methods dedicated to high-dimensional vectors (e.g. PROCLUS [46]). The idea is that only a fraction of all dimensions are pertinent and these dimensions vary from one cluster to another.

For such a purpose, we propose to select in each block a fraction of pixels according to a fixed ratio γ. We look for those pixels which are the most stable spectrally. For a given block of the band image $\mathbf{H}_i$, we consider its spectral neighbors in the bands $\{\mathbf{H}_j\}$, with $i - \Delta \leq j \leq i + \Delta$ (Δ is the radius of the neighborhood), and measure the per-pixel dispersion. The $s^2\gamma$ pixels exhibiting the least dispersion of values are then retained. In this way, by proceeding block by block, we can define for each band $\mathbf{H}_i$ an associated binary mask $\mathbf{\Gamma}_i$ of selected pixels. Note that according to the notations used in equations (4) and (5) we have $N_b = s^2\gamma \quad \forall\, b$. Algorithm 1 reports the full *pixel selection* (PS) procedure to compute a binary mask $\mathbf{\Gamma}$ for an entire hyperspectral image, such that each block of each band contains a fixed number of selected pixels.

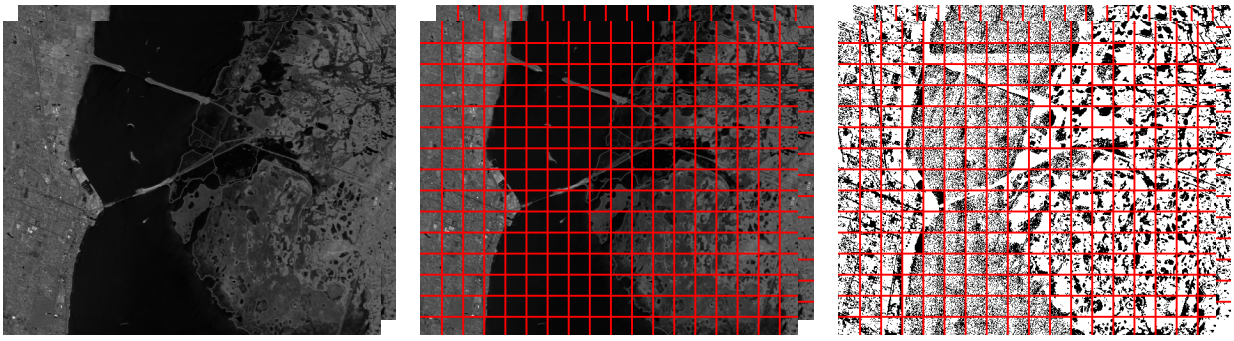
Algorithm 1 Pixel Selection (PS)

Inputs: $\mathbf{H} \in \mathbb{R}^{h \times w \times n}$, s, γ, Δ
Outputs: $\mathbf{\Gamma} \in \mathbb{R}^{h \times w \times n}$ (binary mask)

$\{\mathcal{P}_b\}_{b=1}^{B} \leftarrow \mathbf{H}$ ▷ pixel sets for $s \times s$ blocks
for $i = 1, \dots, n$ **do** ▷ for all bands
  $\mathbf{\Gamma}_i = \mathbf{0}_{h \times w}$
  for $b = 1, \dots, B$ **do** ▷ for all blocks
    $\mathbf{Y}_{\text{ref}} \in \mathbb{R}^{s \times s} \leftarrow \mathbf{H}_i(\mathcal{P}_b)$
    ▷ current block
    $\mathbf{Y}_{\text{ngb}} \in \mathbb{R}^{s \times s \times 2\Delta} \leftarrow \{\mathbf{H}_j(\mathcal{P}_b)\}_{|i-j| \leq \Delta,\ i \neq j}$
    ▷ spectral neighbor blocks
    $\mathcal{I}$ = SELECTBLOCKPIXELS($\mathbf{Y}_{\text{ref}}, \mathbf{Y}_{\text{ngb}}, \gamma$)
    $\mathbf{\Gamma}_i(\mathcal{P}_b(\mathcal{I})) \leftarrow 1$
  end for
end for

procedure SELECTBLOCKPIXELS($\mathbf{Y}_{\text{ref}}, \mathbf{Y}_{\text{ngb}}, \gamma$)
  $(s, s, n_B) \leftarrow \mathbf{Y}_{\text{ngb}}$ ▷ retrieve dimensions
  $\mathbf{Z} = \dfrac{1}{n_b} \displaystyle\sum_{j=1}^{n_B} |\mathbf{Y}_{\text{ngb},j} - \mathbf{Y}_{\text{ref}}|$
  $n_P = s^2\gamma$ ▷ number of pixels to keep
  $\mathcal{I} \leftarrow$ locations corresponding to the n_P lowest values
  return $\mathcal{I}$
end procedure

For each image band, the MLSD are computed on the spatial supports defined by the PS procedure. The dimension of the MLSD depends on the number of image features M; in particular, we have $\mathbf{m}_i^b \in \mathbb{R}^M$ and $\mathbf{C}_i^b \in \mathrm{Sym}_M^+$ (the space of $M \times M$ symmetric positive definite matrices). As for the number of descriptors per band image, by adopting the MLSD, we pass from $m = h \times w$ (the number of pixels) to $2B$, where $B \approx m/s^2$ is the number of blocks. This represents a substantial reduction in the number of descriptors used; thus we can effectively note that the MLSD offers a way to concisely represent the content of a hyperspectral image. Fig. 3 summarizes visually the process prior to generating the MLSD: a regular grid is considered and a fraction of pixels, according to fixed ratio per block are selected (those ones that are not whitened).



(a) (b) (c)

Fig. 3. Visual example of construction of the MLSD. For a given band, two or more features can be considered (a). A uniform grid is applied to all band images (b). The per-block statistics are calculated on a subset of pixels (those ones that are not whitened (c)).

B. Kernel-based probabilistic modeling

In this section we have defined our new descriptors for hyperspectral images, the MLSD, and a way to compute inter-band distances based on the latter (see equations (6) and (9)). These distances can be straightly used when clustering bands in the image space.

The second step of the proposed approach consists in defining a probabilistic model. We adopt an information-maximization clustering approach for its flexibility and discriminative capability, as pointed out in Section III. The model to define will then be a conditional-probability model of the type $p(y|\mathbf{x};\boldsymbol{\alpha})$, defining a mapping between the data $\mathbf{x}$ and the unknown cluster labels y.

Kernel smoothing methods are largely employed to perform nonparametric density estimation; the use of kernel functions is even more crucial when date is scarce [47]. This is particularly the case for hyperspectral data, where the number of samples (the number of bands) is small compared to the dimensionality of the data. As in [31], we use a kernel-based density model for the posterior class probability:

$$p(y|\mathbf{x};\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_{y,i} \mathcal{K}_S(\mathbf{x}, \mathbf{x}_i) \ , \tag{10}$$

where $\mathcal{K}_S$ is a kernel smoothing function, and $\boldsymbol{\alpha}_y \in \mathbb{R}^n$ is a parameter vector related to the cluster y that weights each data point (each band). The set of vectors $\{\boldsymbol{\alpha}_y\}$ is the set of parameters to be determined. Each vector indicated how the different bands contribute to the density of a certain cluster.

The MLSD we have defined in the previous subsection (IV-A2) exhibits two types of descriptor. The kernel $\mathcal{K}_S$ will then be the combination of two kernels dedicated respectively to mean and covariance descriptors, $\mathcal{K}_M$ and $\mathcal{K}_C$ respectively. In general terms we have:

$$\mathcal{K}_S(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{F}\left(\mathcal{K}_M(\mathbf{x}_i, \mathbf{x}_j), \mathcal{K}_C(\mathbf{x}_i, \mathbf{x}_j)\right) \ , \tag{11}$$

where $\mathcal{F}$ is an operator defining the way to combine the two kernels. The simplest solution would be to consider a convex combination with possibly variable weights. In Section V we will discuss other possible solutions for the kernel mixing operator $\mathcal{F}$.

As for the shape of the two kernels, we propose to use an Epanechnikov smoothing kernel with variable kernel bandwidth:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j, d) = \begin{cases} \frac{3}{4}\left(1 - \frac{d(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_i \sigma_j}\right) & d(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma_i \sigma_j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where σ_i and σ_j are two variable scaling factors for the kernel (the bandwidths) depending on the two data points considered, and d is a quadratic distance function. We choose an Epanechnikov kernel because it has been proven to be the most efficient kernel for density estimation [48]. Moreover, contrary to the Gaussian kernel, it has *by definition* a bounded support. We believe that this is useful for hyperspectral data, where attributing a positive, even small, weight to a faraway data point from the one considered might be undesirable consequences. The two data points in fact correspond to distant spectral bands, possibly with incompatible spectral contents. As proposed in [49] the values of σ_i and σ_j are set by taking into account the distance to the T-th neighbor of the data point considered. In this way, the kernel bandwidths, rather than being fixed, will respect the local dispersion of the data. We therefore talk about *adaptive neighborhoods* [47]. In practice, we define $\sigma_i = d(\mathbf{x}_i, \mathbf{x}_i^T)^{1/2}$ and $\sigma_j = d(\mathbf{x}_j, \mathbf{x}_j^T)^{1/2}$. The value of T, the neighborhood size considered, is very important, as it determines the variable kernel

bandwidths. Experimentally we found that it has to depend on the ratio between the total number of data points (bands), n, and the number of clusters we want to form, k. In particular, we set:

$$T = \max\left(\min\left(\left\lceil \frac{n}{2k} \right\rceil, 9\right), 3\right) \ , \tag{13}$$

with T linearly depending on the ratio $\frac{n}{k}$ under the condition $3 \leq T \leq 9$. The expression (12) is general, with the distance used to be specified. By using the inter-band distances defined in Section IV-A, we have $\mathcal{K}_M(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j, d_M)$ and $\mathcal{K}_C(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j, d_C)$. We have then fully defined the two kernels that, combined as in (11), can be used in our conditional probability model (10). In the next section we will detail the optimization procedure to derive the parameters of the model $\{\boldsymbol{\alpha}_y\}$.

C. Optimization of model parameters

The proposed method belongs to the family of information-maximization clustering (see Section III), where posterior class probability is directly modeled and parameters of the model are derived by maximizing the information between the data vectors and the unknown class labels. To find the parameters of our kernel-based model, we use *Squared-loss Mutual Information* (SMI) (2) as an information metric. SMI has the advantage of leading to a quadratic form w.r.t. the posterior probability, thus making an analytical solution possible. Let first express the posterior class model (10) into the following, more concise, matrix form:

$$p(y|\mathbf{x}_j) = \boldsymbol{\alpha}_y^\mathsf{T} \mathbf{K}_{S,j} \ , \tag{14}$$

where $\mathbf{K}_S$ is an $n \times n$ matrix containing all the kernel distances between the n bands, taken two by two. $\mathbf{K}_{S,j}$ is the j-th column of that matrix. By plugging the expression above into (3), we then obtain:

$$\text{SMI} \simeq \frac{1}{2n} \sum_{y=1}^{k} \frac{1}{p(y)} \sum_{j=1}^{n} \boldsymbol{\alpha}_y^\mathsf{T} \mathbf{K}_{S,j} \mathbf{K}_{S,j}^\mathsf{T} \boldsymbol{\alpha}_y - \frac{1}{2} \ . \tag{15}$$

By imposing equi-probable clusters ($p(y) = \frac{1}{k} \forall y$), we finally obtain the following objective function to maximize:

$$\text{SMI} \propto \frac{k}{2n} \sum_{y=1}^{k} \boldsymbol{\alpha}_y^\mathsf{T} \mathbf{K}_S \mathbf{K}_S^\mathsf{T} \boldsymbol{\alpha}_y \ . \tag{16}$$

The expression in (16) is a sum of *Rayleigh quotients*. The set of parameter vectors that maximize it, $\{\hat{\boldsymbol{\alpha}}_y\}_{y=1}^k$, is therefore represented by the first k eigenvectors of the matrix $\mathbf{K}_S$. Moreover, a normalization condition on the solution has to be imposed, based on the hypothesis of equi-probable clusters. We can develop the cluster density probability as follows:

$$\begin{aligned} p(y) &= \int p(y|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &\simeq \frac{1}{n} \sum_{i=1}^{n} p(y|\mathbf{x}; \boldsymbol{\alpha}_y) \quad , \\ &= \frac{1}{n} \boldsymbol{\alpha}_y^\mathsf{T} \mathbf{K}_S \mathbf{1}_n \end{aligned} \tag{17}$$

where $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ indicates a column vector with n ones. By imposing $p(y) = \frac{1}{k} \ \ \forall y$, we then have the following normalization condition:

$$\mathbf{A}^\mathsf{T} \mathbf{K}_S \mathbf{1}_n = \frac{n}{k} \mathbf{1}_k \ , \tag{18}$$

where $\mathbf{A} \in \mathbb{R}^{n \times k}$ is a matrix obtained by gathering all parameter vectors $\{\boldsymbol{\alpha}_y\}_{y=1}^k$ as columns.

After normalizing the parameter vectors to satisfy the condition in (18), we are finally able to compute each class posterior probability, for each cluster given any band, as expressed in (10). Band selection is finally performed by selecting, for each cluster, the band leading to the highest posterior probability. This can be expressed as follows:

$$\mathbb{B} = \left\{ \arg\max_{\mathbf{x}_i \in \mathbb{R}^n} p(y = j|\mathbf{x}_i) \right\}_{j=1}^{k} \ . \tag{19}$$

V. Experimental results in classification

In this section we perform a quantitative assessment of our proposed approach for unsupervised band selection (UBS). All parameters and implementation choices are supported by experiments, and a comparison with other state-of-the-art UBS methods is carried out. The scheme adopted for the evaluation is reported in Fig. 4. Band selection is performed in a totally unsupervised manner, i.e. only the input image $\mathbf{H}$ is used to decide on the set of spectral bands to retain, $\mathcal{B}$. The spectral signatures of the hyperspectral image are reduced according to the indices of the selected bands. The evaluation of the benefits of this operation is conducted in a supervised setting, where a pixel map with ground-truth labels, $\mathbf{L}$, is available. Then, a classification algorithm takes as input the reduced spectral signatures in order to estimate each pixel label. It is important to emphasize that the selection of the bands is done upstream before the classification step, with no a priori knowledge on pixel labels. Given the fact the ratio between unlabeled and labeled pixels (typically, far below 50%), the decision taken from a unsupervised method for selection can benefit from having all available samples instead of few labeled samples. In order to select pertinent bands, the information delivered by unlabeled pixels is also important.

In the next subsection we provide details about the data sets used for the evaluation. Section V-B describes the protocol followed in the classification step. Section V-C presents experiments conducted for tuning parameters of our UBS method. Our proposed algorithm is finally compared with other UBS algorithms in the literature. For the latter, parameters have been set according to the instructions provided by the respective authors or carefully tuned, in the absence of details. The results of this comparison are presented in Section V-D.

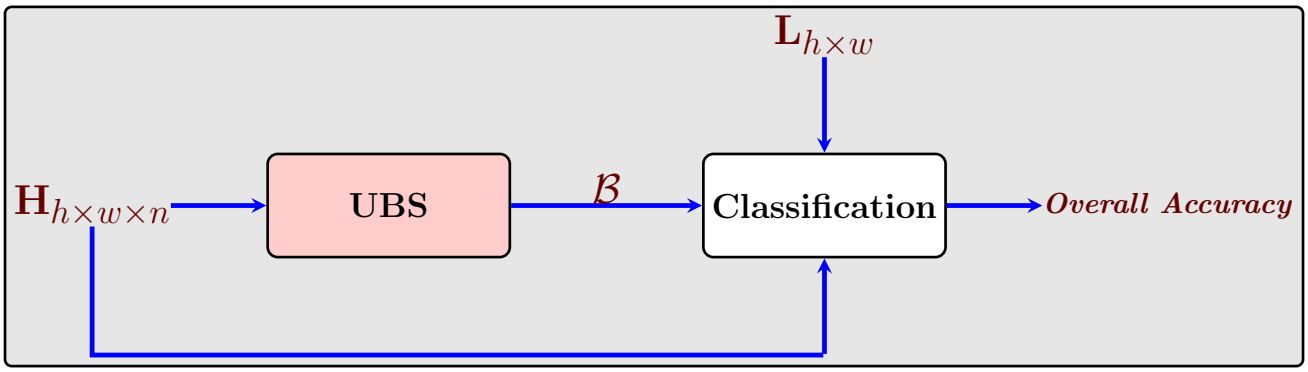


Fig. 4. Evaluation scheme for our proposed approach. Unsupervised band selection (UBS) is used to choose a set of spectral bands $\mathcal{B}$ from an input hyperpectral image $\mathbf{H}$. The same image, the set of chosen bands, and a pixel map with ground-truth labels $\mathbf{L}$, are the inputs of a classification step that allows us to compute an accuracy metric.

A. Data sets

To assess the performance of our proposed method we made use of some publicly available hyperspectral data sets [1]. All scenes arise from earth observation images taken from airborne vehicles or satellites. Table I resumes the main characteristics, from the analysis view point, of the data sets considered.

1) Kennedy Space Center (KSC): This data set was acquired by the NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) instrument over the Kennedy Space Center (KSC), Florida, in March 1996. The data consists of a natural wetland/upland environment in which the classes have a poorly defined spatial structure. The acquisition was made from an altitude of approximately 20 km and the resulting spatial resolution is 18 m. The AVIRIS sensor acquires data in 224 bands of 10 nm width with center wavelengths ranging from 400 to 2500 nm. The hyperspectral cube is reduced to 176 bands, after the removal of noisy and water absorption bands. For classification purposes, 13 classes representing the various land cover types were defined (varying from 2.0% of the labeled pixels, for the least populous class, to 17.8%, for the most populous one). Training data were selected using land cover maps derived from color infrared photography

2) Pavia University scene: This data set was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) during a flight campaign over Pavia, northern Italy. The number of sensed spectral bands is 103, for wavelengths ranging from 430 to 860 nm. The spatial resolution is 1.3 meters; this leads to images presenting a well-defined spatial texture. The image ground truth differentiates 9 classes of materials, varying in percentage from 2.2% to 43.6% of the total number of labeled pixels. Fig. 5 reports an example of a band extracted from this data set. Different colors have been overlaid to show the different classes.

[1] All data is downloadable at the following url: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

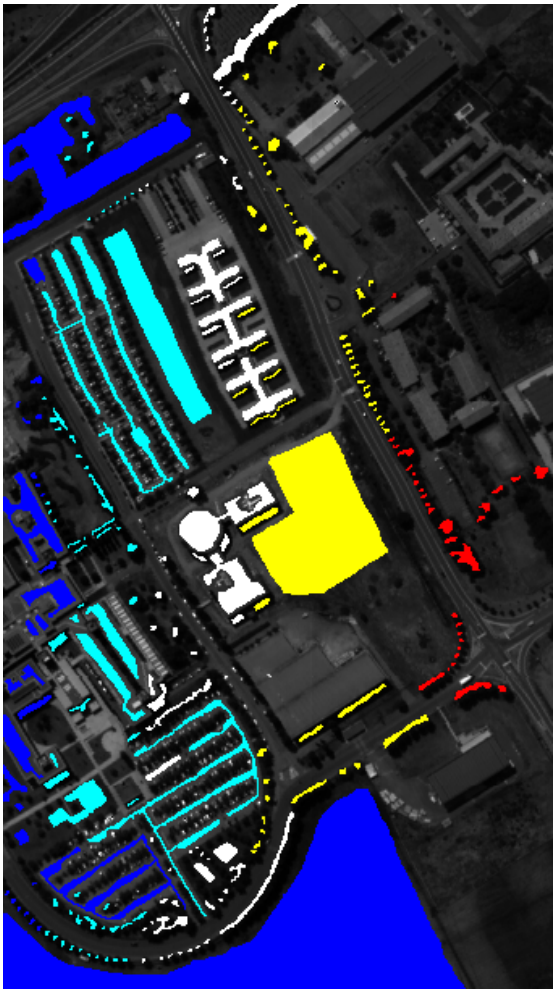


Fig. 5. Example of band from the *Pavia University* data set. The labeled samples are overlaid with different colors, according to the class.

3) Indian Pines scene: The data set was acquired by the AVIRIS sensor over the Indian Pines site in North-western Indiana in June 1992. The Indian Pines scene contains for two-thirds agricultural crops, and for one-third forest or other natural perennial vegetation. Originally, 220 spectral reflectance band in the wavelength range of 400-2500 nm were acquired, with a spectral resolution of 10 nm. Twenty water absorption bands (104-108, 150-163, and 220) were successively removed, thus yielding a 200-band image. The spatial resolution is 20 m. Most classes were crops located in fields with regular boundaries, resulting in a spatially structured data set (conversely, the detail definition is poor). The ground truth available is designated into sixteen classes.

TABLE I
DATA SETS CONSIDERED FOR THE EVALUATION AND THEIR MAIN CHARACTERISTICS.

Data set	Image res.	Spatial res.	# of classes
KSC	$512 \times 614 \times 176$	18 m	13
Pavia University	$610 \times 340 \times 103$	1.3 m	9
Indian Pines	$145 \times 145 \times 200$	20 m	16

B. Protocol for classification performance assessment and parameters of the classifiers

In the data sets described in the previous subsection a fraction of the pixels is provided with ground truth labels, the number of classes varying from 9 to 16 (see Table I). Typically, even the smallest classes count few dozens of samples. The UBS task has then been evaluated in a supervised setting. For multiple sample draws (100), the $\overline{m}$ labeled pixels (where $m = h \times w$ is instead the total number of pixels, according to the notation of Section II-A) are divided into m_{train} samples for training and m_{test} samples for testing, by respecting the original class distribution. Unless otherwise specified, the test-to-train ratio $\frac{m_{\text{test}}}{m_{\text{train}}}$

used by default is equal to 0.1. The training data set is used to learn a classification model, which is afterwards applied to classify the "unknown" pixels. The metric used to evaluate the classification performance is the *overall accuracy*:

$$\text{OA} = \frac{\sum_{i=1}^{m_{\text{test}}} 1_{l_i=\hat{l}_i}}{m_{\text{test}}} , \tag{20}$$

where l_i is the ground truth label for the i-th pixel, and $\hat{l}_i$ is the predicted class label.

The classification algorithms considered for evaluation are the K-nearest neighbor classifier (with K, the number of neighbors, set as 5), and the *C-SVM* classifier (Support Vector Machines with regularization parameter C) [50]. For the latter, we set C equal to 1 and we choose a Gaussian radial basis function as a kernel, with kernel scale optimized via cross-validation. As for the classification features, we used the *spectral signatures* of the hyperspectral image, i.e. the vectors obtained by gathering all spectral magnitude values of each image pixel across the whole spectrum. When UBS is performed, the spectral signatures are reduced to only the values corresponding to the selected wavelengths. This has the double benefit of (i) mitigating the data redundancy problem typical of hyperspectral data and (ii) reducing the computational time for classification, as the feature vectors to process are considerably shortened.

C. Choice of features and parameters for UBS

In this section we define the operational setting of our clustering-based method defined in Section IV, by justifying implementation choices and parameter tuning. For such a purpose, there are three main aspects that should be carefully addressed:

- the descriptors used to represent each band, which are the MLSD according to our final choice;
- the pixel selection procedure, i.e. how we choose the parameters s (block size) and γ (percentage of pixels to keep in each block) of Algorithm 1;
- the kernel parameters; namely how the kernel bandwidth are computed and how multiple kernels are mixed in (11).

We take up the three problems in the following.

1) Choice of the features: The MLSD are statistical descriptors (mean and covariance) computed over multiple-feature vectors. There are therefore several intermediate steps:

- multiple features are considered instead of the only spectral intensity;
- mean and covariance statistics are computed on the feature vectors;
- the two statistical descriptors are jointly used in a kernel model.

Tables II and III show the results in terms of OA for the classification problem considered (by using the *KNN* and *SVM* algorithm, respectively). All the features obtained in the "intermediate steps" that lead to the MLSD are considered. The results presented in the tables II and III are for two data sets, *Kennedy Space Center*, and *Pavia University*) and two different numbers of selected bands ($k = 10$ and $k = 15$). The rest of the parameters, those regarding the pixel selection strategy and the kernels, are kept fixed, according to the choices that will be detailed in the following subsections. Note that the "Single features" approach basically corresponds to merely adapting the clustering algorithm in [39] to the data vectors resulting from collecting spectral magnitude values from each band image. The "Multi features" approach can be instead referred to the method presented in [31], where multi-feature vectors are considered. The use of the pixel selection and the new kernel strategies is nevertheless present in all approaches considered.

TABLE II
COMPARISON BETWEEN DIFFERENT POSSIBLE FEATURE SCHEMES FOR THE PROPOSED APPROACH (***KNN* CLASSIFIER**).

	Kennedy Space Center		***Pavia University***		***Indian Pines***		
	10 bands	15 bands	10 bands	15 bands	10 bands	15 bands	**Average**
Single features [39]	81.5	82.0	85.4	85.5	69.5	73.7	80.8
Multi features [31]	78.3	81.5	85.5	86.7	72.6	72.9	80.9
Single stat. (mean)	81.0	81.4	85.8	85.8	72.5	74.1	81.3
Single stat. (cov)	81.2	83.1	84.4	86.2	71.4	72.3	81.3
MLSD	81.3	83.6	86.2	86.6	72.5	73.7	82.0

As we can observe from the tables II and III, the MLSD leading to the best performance overall. The data sets that benefit the most from using the MLSD are *Pavia University* and *Kennedy Space Center*. In particular, the *Kennedy Space Center* set is characterized by a high spatial variability and benefits from the use of a covariance descriptor, especially for $k = 15$. In general, when aggregating the two statistical descriptors, we are able to outperform only-mean or only-covariance descriptors.

As for the multiple features considered in the experiments, we enriched spectral intensity with one type of spatial features. In practice, we considered a Laplacian of Gaussian (LoG) operator [51], where a Gaussian filter of size 5×5 and standard deviation $\sigma = 0.5$ is applied. Thus, the number of feature matrices as for the notation introduced in Section IV-A is $M = 2$.

TABLE III
COMPARISON BETWEEN DIFFERENT POSSIBLE FEATURE SCHEMES FOR THE PROPOSED APPROACH (***SVM* CLASSIFIER**).

	Kennedy Space Center		*Pavia University*		*Indian Pines*		
	10 bands	15 bands	10 bands	15 bands	10 bands	15 bands	**Average**
Single features [39]	87.4	87.6	86.9	88.0	72.0	75.8	84.4
Multi features [31]	84.0	86.7	87.2	89.5	74.4	74.7	84.4
Single stat. (mean)	86.7	88.0	88.4	89.0	75.0	76.5	85.4
Single stat. (cov)	85.7	88.2	86.0	89.3	72.6	74.9	84.3
MLSD	86.9	88.9	88.9	89.5	74.9	76.2	85.8

2) Pixel selection parameters: The pixel selection procedure, as detailed in Algorithm 1, mainly depends on two parameters: the block size (s) and the percentage of pixels that are selected in each block (γ). For each data set, we find the best couple of parameters $(\hat{s}, \hat{\gamma})$ by measuring the average OA on the pixels used for training for a range of k. In practice, we took k ranging from 5 to 12. The procedure is done via grid search, as the target values are not many (we investigate block size ranging from 3 to 19, and we look for γ with a resolution of 10%). Table IV reports the values finally found for each of the data sets considered. Fig. 6 shows instead multiple curves of average OA for the *Indian Pines* data set, for different block size and with γ varying. The couple of values leading to the highest values is the one reported in Table IV.

TABLE IV
BEST PARAMETERS FOUND FOR THE PIXEL SELECTION PROCEDURE (BLOCK SIZE AND PERCENTAGE OF SELECTED PIXELS).

Data set	**Block size**	**% Sel. Pixels**
Indian Pines	19	80%
Kennedy Space Center	7	30%
Pavia University	5	90%

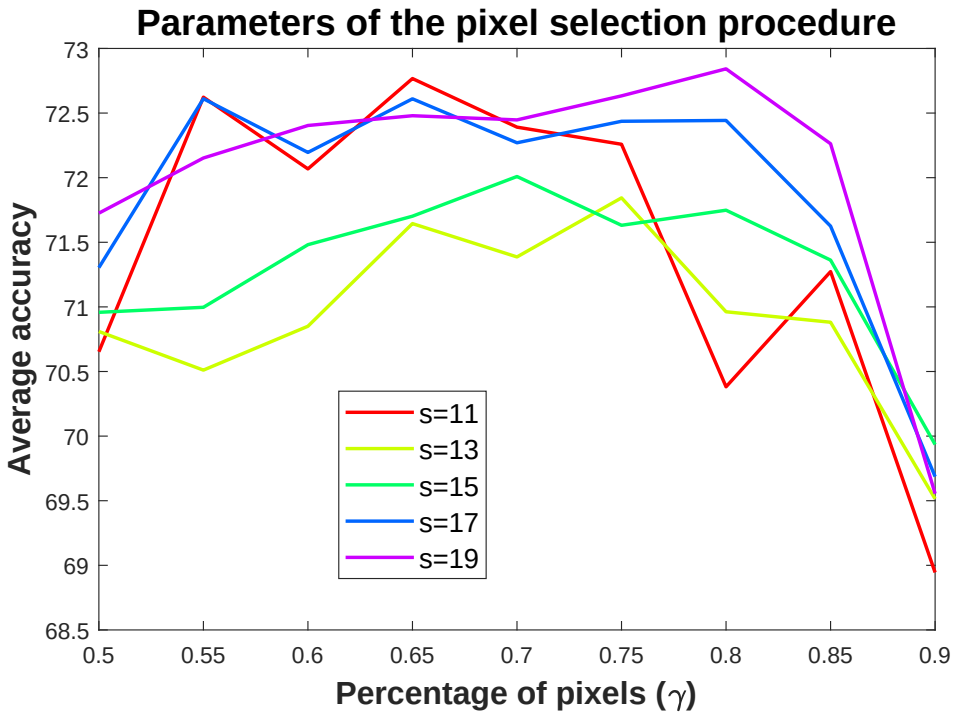


Fig. 6. Overall accuracy on training samples, while varying the parameters of the pixel selection procedure, i.e. block size (s) and percentage of selected pixels (γ). The data set here considered is the *Indian Pines* scene.

The best parameters of the pixel selection procedure can also tell something about the nature of the data sets. The *Pavia University* scene is for example rich in detail and with a low presence of noise. This can explain the fact that smaller blocks are the ideal ones and almost all pixels are pertinent. Conversely, the *Indian Pines* data set benefits from averaging over large blocks.

3) Kernel parameters: In Section IV-B we detailed the kernel probabilistic model chosen, which is based on Epanechnikov smoothing kernel with variable kernel bandwidths. To compute the bandwidth of the kernel we consider an adaptive neighbor-

hood, by looking at the distance between the actual data point and its T-th neighbor. We gave a formula for T in (13), which is a linear expression w.r.t. the ratio between the total number of bands and the number of clusters. In Fig. 7, from overall accuracy, we show the advantage of using such "adaptive T", rather than a fixed one, for different values of k (the number of clusters).

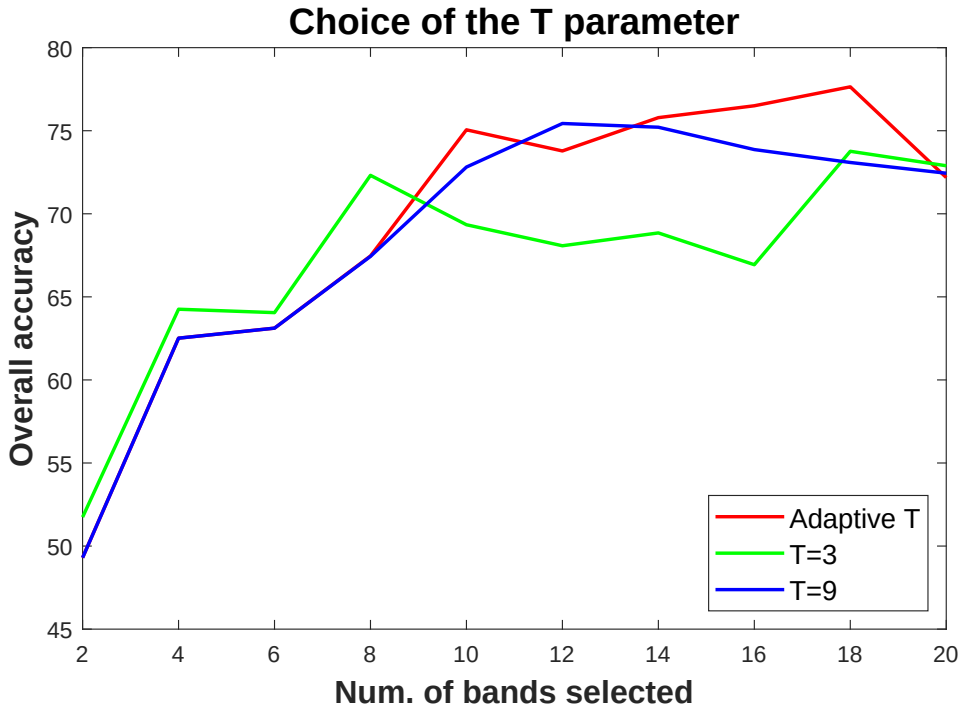


Fig. 7. Performance for different values of the T parameter (the size of the adaptive neighborhood in the kernel bandwidth). The data set here considered is the *Indian Pines* scene.

Eq. 11 is the general expression of the two kernels based on the mean and covariance descriptors combined to form a new kernel $\mathcal{K}_S$. The simplest solution is to consider a convex combination of the two kernels, i.e. $\mathcal{K}_S = \alpha_M \mathcal{K}_M + \alpha_C \mathcal{K}_C$. Via experiments, we found that a good compromise is to use quite uniform weights, with possibly a slightly higher importance given to the mean descriptors. In practice, we simply set $\alpha_M = 0.5$ and $\alpha_C = 0.5$ for all data sets. In [43], Li *et al.* propose a novel approach for kernel mixing called *generalized composite kernel* (GCK). Instead of considering a convex combination of kernels, GCK consists in simply concatenating the kernel outputs computed individually. As a consequence, the resulting kernel matrix $\mathbf{K}_S$ of Equations 14 and 15 have a double number of entries, i.e. is of size $2n \times n$. In our experiments, we also tested this solution, which provides equivalent performance. However, due to the slightly higher computational time, we have selected the convex combination solution.

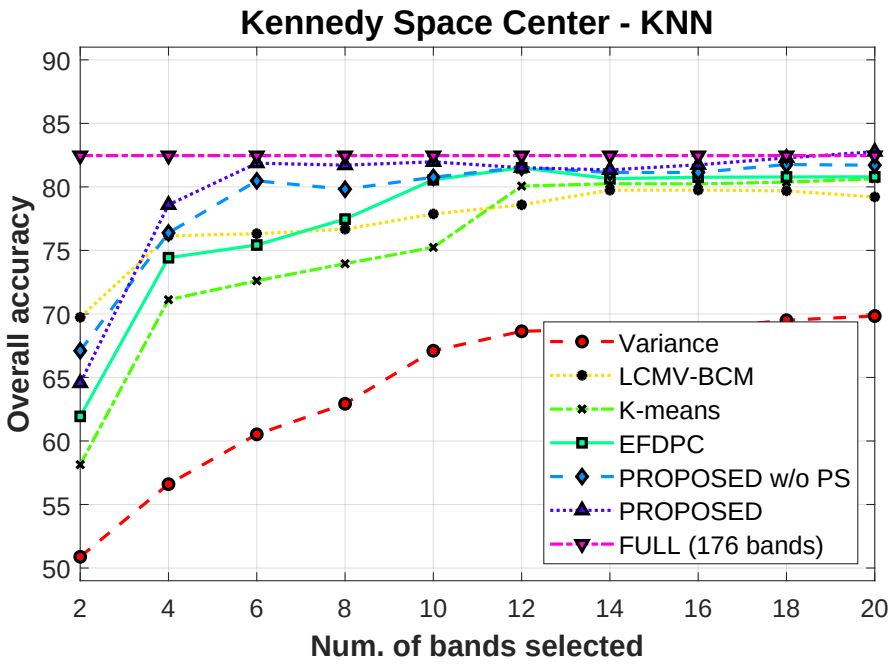
D. Comparison with state-of-the-art methods

In this section we assess the performance of our proposed method, as described in Section IV and with the settings discussed earlier, in comparison with other state-of-the-art UBS algorithms. For this evaluation, we consider the following algorithms to perform band selection:

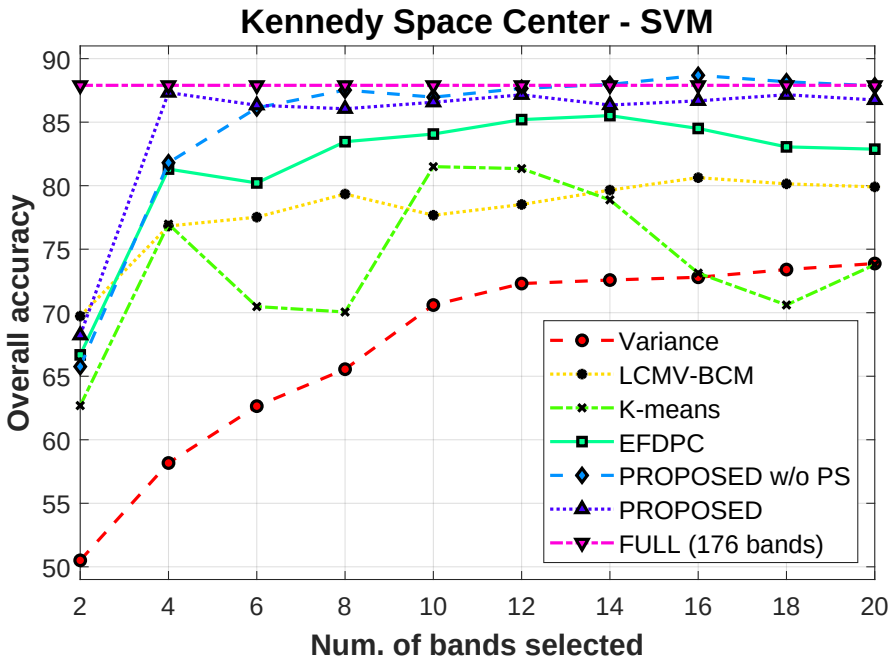
- Ranking-based band selection w.r.t the variance of the spectral intensity [17]: the k bands reporting the highest variance values are selected.
- The LCMV-BCM (linearly constrained minimum variance with band correlation minimization) method presented in [18]. In the family of ranking approaches, LCMV-BCM method considers band selection as a constrained energy minimization problem for which different criteria can be used. In the original paper, four different criteria have been considered; for our evaluation, we have used the band correlation minimization (BCM) criterion, which turned to be the best performing one in our experiments.
- The well-known k-Means clustering method [52], [23]. After clustering the bands in the image space, the cluster centroids are kept as bands to select.
- The recent *E-FDPC* (Enhanced Fast Density-Peak based Clustering) algorithm proposed in [30]. This method follows hybrid clustering-ranking approach. Under the hypothesis that the data is cluster-wise arranged, an indicator based on the likelihood is defined for each band. The main idea is that the exemplars should be data points with high local density, sufficiently "isolated"

The four considered state-of-the-art methods belong both to the family of ranking-based UBS algorithms [17], [18] and to the family of clustering-based UBS algorithms [23], [30]. All the methods, except for k-Means for which a standard implementation was used, have been re-implemented in Matlab, by carefully following the instructions described in the related papers. UBS algorithms are also compared with the case where band selection is not performed at all, i.e. the whole spectrum of bands is considered for classification. The test protocol is the one described in Section V-B. For our proposed method, we evaluate the performance with or without the pixel selection (w/o PS) procedure described in Section IV-A for pre-selecting a subset of pixels.

The resulting graphs showing the obtained overall accuracy (OA), function of the number of selected bands K, are reported in Fig. 8, 9 and 10, for the *KSC*, *Pavia University* and *Indian Pines* data sets, respectively. For each data set, we report the results obtained by using both KNN and SVM as classifiers.

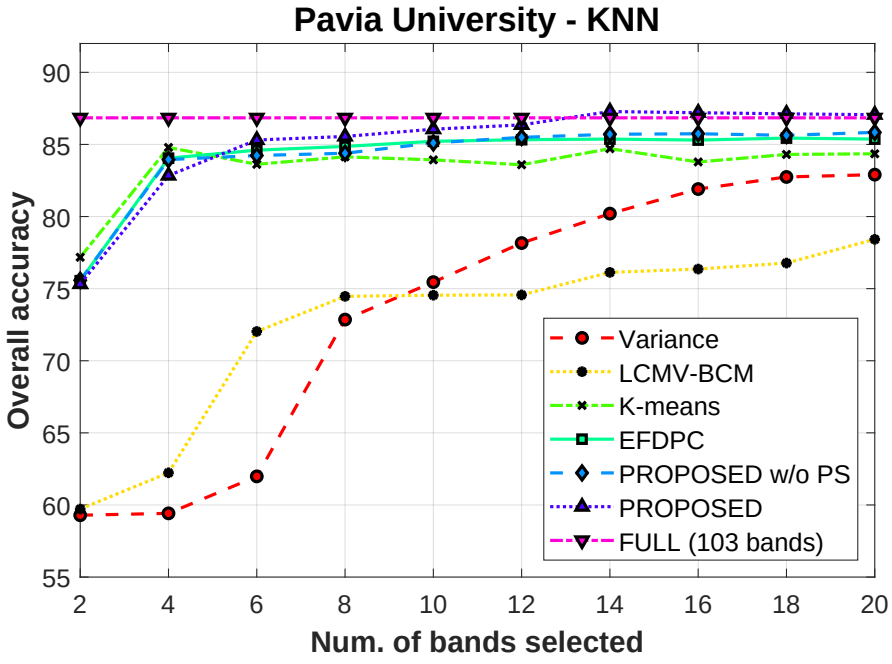


(a) KNN

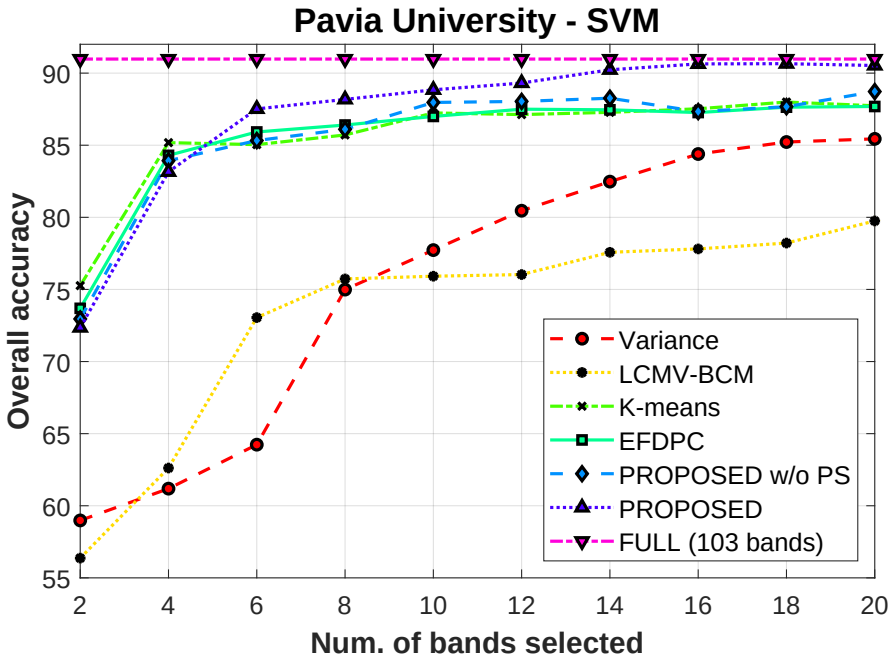


(b) SVM

Fig. 8. Performance (overall accuracy) on the *Kennedy Space Center* data set versus number of bands selected, for different band selection methods and two classifiers: KNN (a) and SVM (b).



(a) KNN

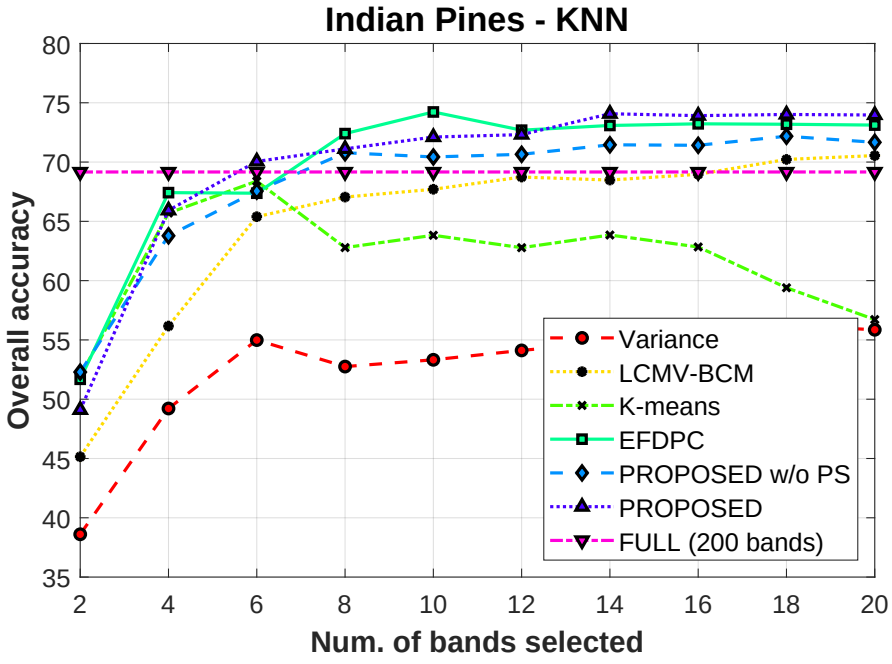


(b) SVM

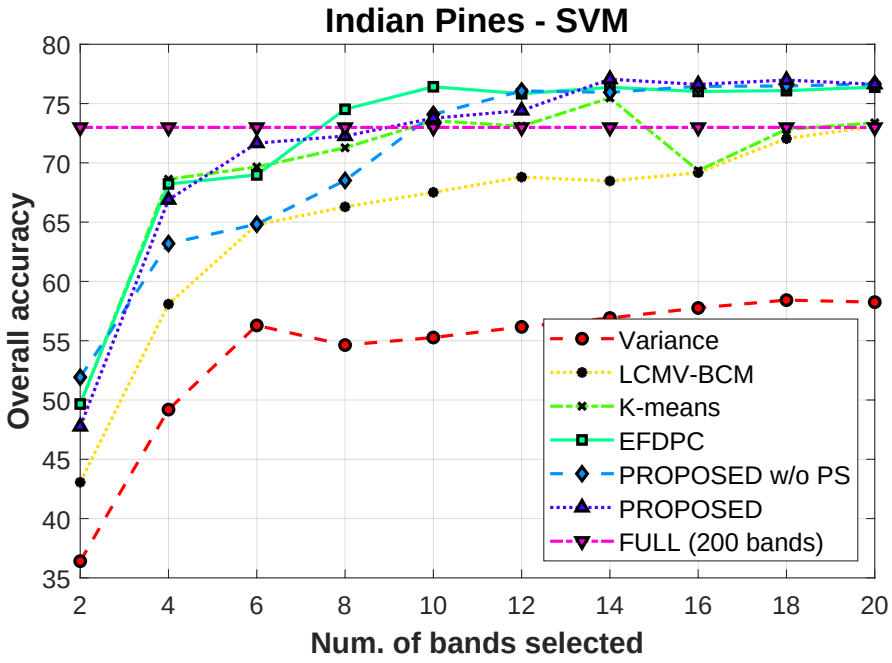
Fig. 9. Performance (overall accuracy) on the *Pavia University* data set versus number of bands selected, for different band selection methods and two classifiers: KNN (a) and SVM (b).

From the OA curves, we can conclude that globally our proposed UBS method is the one offering the best performance. For the *KSC* and *Pavia University* data sets we can observe an uplift on the computed OA value up to 3%, depending on the number of bands k chosen and the classification algorithm used. For the *Indian Pines* data set, our method is competitive with the E-FDPC algorithm. Thanks to the graphs, we can also observe the importance of the pixel selection (PS) strategy adopted as a first step of our method, before computing the MLSD. With the *Pavia University* data set the PS step enables a considerable gain; with the *Indian Pines* data set, it allows to reach the performance of the E-FDPC algorithm. It is also important to note that, with all data sets, band selection gives comparable, or sometimes even better, performance than using the whole spectrum.

This is particularly true for the *Indian Pines* data set, for which band selection outperforms the full spectrum case in terms of classification accuracy. This is due to the fact that data suffers of noise. Moreover, given the high number of bands (200)



(a) KNN



(b) SVM

Fig. 10. Performance (overall accuracy) on the *Indian Pines* data set versus number of bands selected, for different band selection methods and two classifiers: KNN (a) and SVM (b).

and the low spatial resolution, several bands turn out to be similar in spectral content. The high correlation between features implies that a feature selection step can help improving the performance. The same phenomenon can be observed in [30, Fig. 6].

In general terms, we can state that, by using our UBS method, with $k = 20$ bands we have already enough information to resume the whole spectral content and get the best performance. This proves the goodness of the band selection strategy.

Fig. 8, 9 and 10 are obtained by fixing the test-to-train ration to 10%, as explained in Section V-B and letting the value k vary. In Fig. 11 we show the results if we act in the inverse way, i.e. we keep the number of bands to select fixed, and we vary the ratio between the samples used for testing and training. The results showed are for the *KSC* data set and for $k = 5$. As we can see, in this case too our proposed algorithm offers good performance, and this is extremely stable as the number

of samples used for training changes.

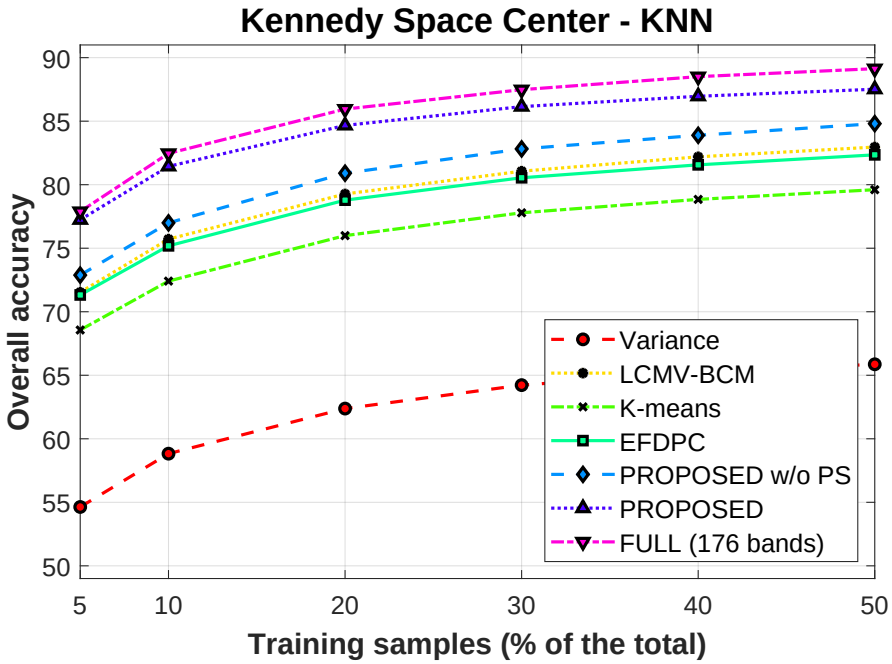


Fig. 11. Performance (overall accuracy) on the *Kennedy Space Center* data set versus percentage of training samples, the number of selected bands being fixed ($k = 5$).

As an example, in Fig. 12 we report the 5 selected bands that lead to the results showed in Fig. 11, for three methods:

a Band selection via k-Means clustering [23];

b The *E-FDPC* (Enhanced Fast Density-Peak based Clustering) algorithm [30];

c Our proposed approach featuring the pixel selection procedure.

Even though it is difficult to conclude by visually assessing the selected band images, our method seems to choose bands presenting the largest variety. Overall, all the 5 selected bands exhibit different spectral contents, whereas for k-Means and *E-FDPC* we can observe pairs of rather similar bands (bands # 59–92 and bands # 51–82, respectively).

Finally, in Table V we report the results in term of average running time, for the three data set and the UBS methods considered. All the simulations are executed on an Intel Core i7 3.7 GHz computer with 64GB RAM. The simulation environment is MATLAB (R2018a). Our method is unquestionably slower than the other methods taken for comparison. It in fact includes a first part of feature construction (including pixel selection and per-block statistic computation), which is absent in other algorithms. Once the kernel matrix is built, the optimization part described in Section IV-C is relatively fast. However, despite the higher average running time measured, we believe that such time is acceptable, considering that band selection is an operation to be performed offline and only once for a given data set. Moreover, code optimization can be certainly done. E.g., it has been showed that the computation of covariance matrices could be accelerated by the use of integral images [53].

TABLE V
MEASURED AVERAGE RUNNING TIME FOR DIFFERENT UBS METHOD.

	Time (sec.)		
Method	***Indian Pines***	***KSC***	***Pavia University***
Variance [17]	0.1	0.5	0.1
LCMV-BCM [18]	0.2	3.9	1.5
k-Means [23]	0.2	2.1	0.9
E-FDPC [30]	0.1	0.6	0.2
Proposed	1.8	19.9	9.2

VI. EXPERIMENTAL RESULTS IN ENDMEMBER EXTRACTION

In this section we evaluate our proposed method for unsupervised band selection (UBS) in a different context, other than the classification problem outlined in Section V: endmember extraction. In this context, we assess the effectiveness of the selected bands in helping extracting pertinent endmembers. i.e. which are the closest possible to the known ground-truth endmembers.

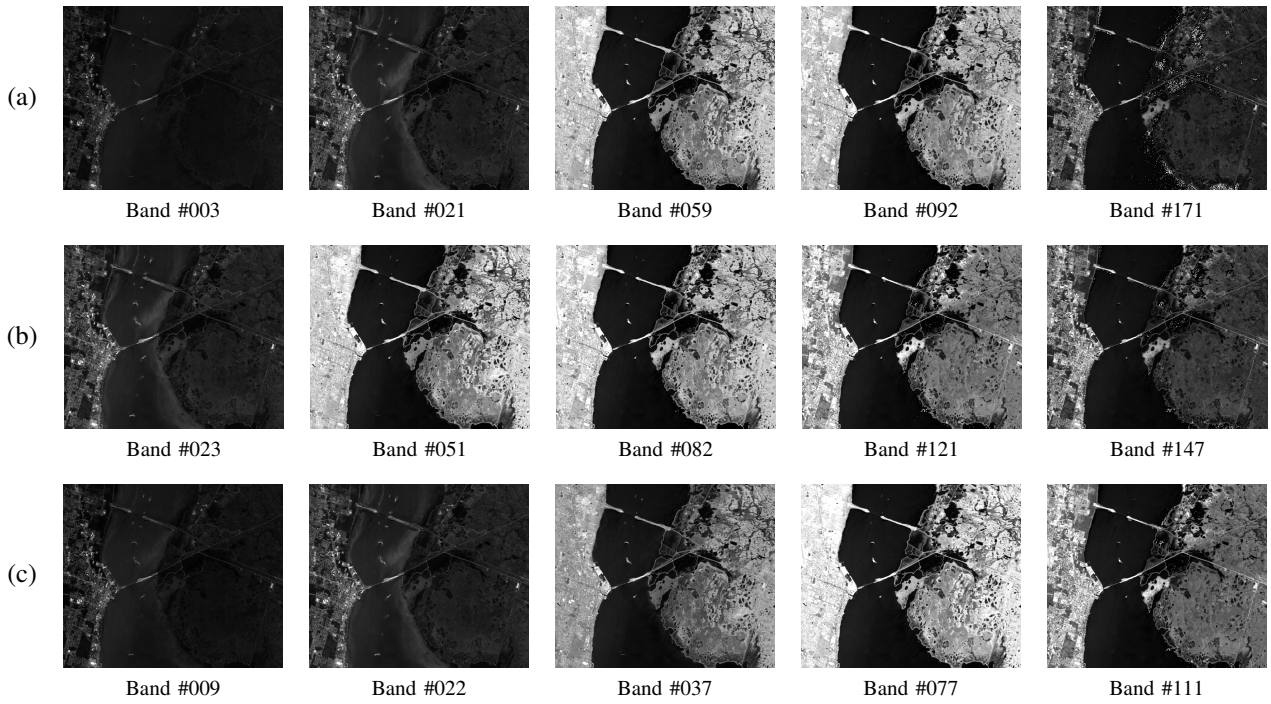


Fig. 12. Example of selected bands ($k = 5$) for the *Kennedy Space Center* data set. The results are reported for three methods: (a) band selection via k-Means clustering [23]; (b) the *E-FDPC* (Enhanced Fast Density-Peak based Clustering) algorithm [30]; (c) our proposed approach, including the pixel selection procedure.

The data set considered is the *Cuprite* image scene, widely used to study endmember extraction, for which Matlab files are available online [2]. This scene was taken by the AVIRIS sensor over the Cuprite Mining District, Nevada, in 1997. It consists of 224 channels, ranging from 370 nm to 2480 nm. After removing noisy bands (1–2 and 221–224) and water absorption bands (104–113 and 148–167), 188 channels are remaining. A region of 250×190 pixels is considered. In this region 12 endmembers, corresponding to the 12 spectrally identified minerals, confirmed via X-ray crystallography, are taken into account. The most representative spatial location and the related spectral signatures are registered.

We perform UBS on the original hyperspectral cube and endmembers are then extracted on the reduced cube. To this end, the well-known *N-FINDR* algorithm by Winter [54] is used. The choice is arbitrary, any other endmember extraction algorithm could be used. Once a set of "supposed" endmembers is found, we compute the similarity between the latter and the known ground-truth endmembers (over the set of selected bands). The metric used to compute spectral similarity is, as in [18], Spectral Angle Mapper (SAM):

$$\mathrm{SAM}(\mathbf{t}, \mathbf{r}) = \cos^{-1}\left(\frac{\sum_{i=1}^{K} t_i r_i}{\sqrt{\sum_{i=1}^{K} t_i^2}\sqrt{\sum_{i=1}^{K} r_i^2}}\right), \tag{21}$$

where K is the number of bands, and $\mathbf{t}$ and $\mathbf{r}$ are the target and reference spectral signature vectors, respectively.

In this context, we compare our method to two other UBS algorithms: *LCMV-BCM* [18], and *E-FDPC* [30]. The number of selected bands is set to $K = 15$. For our approach, parameters and implementation choices are the same as the ones justified in Section V. Results for such comparison, represented as a matrix of correspondences between ground-truth and found endmembers reporting SAM values, are shown in Table VI.

From Table VI, we see that our method is the one that overall leads to the highest similarity between ground-truth and found endmembers. With the bands selected with *LCMV-BCM*, two endmembers are clearly missed. For *E-FDPC*, we have one endmember for which we cannot find a close correspondence. Table VII reports the average SAM value for all correspondences of endmembers.

Given the lowest average SAM value, we can conclude that our method allows to select more pertinent bands to ease the extraction of endmembers.

[2]http://lesun.weebly.com/hyperspectral-data-set.html

TABLE VI
SPECTRAL SIMILARITY MEASURED BY SAM BETWEEN GROUND-TRUTH ENDMEMBERS AND ENDMEMBERS FOUND AFTER AN UNSUPERVISED BAND SELECTION (UBS) STAGE. THE UBS ALGORITHMS CONSIDERED ARE: (A) *LCMV-BCM* [18], (B) *E-FDPC* [30], (C) THE PROPOSED METHOD.

EP	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12
# 1	**0.04**	0.22	0.10	0.11	0.05	0.12	0.12	0.13	1.27	0.11	0.15	1.27
# 2	0.17	**0.18**	0.11	0.24	0.19	0.10	0.11	0.07	1.34	0.08	0.14	1.33
# 3	0.06	0.19	**0.06**	0.14	0.08	0.07	0.08	0.07	1.30	0.06	0.11	1.29
# 4	0.12	0.32	0.22	**0.09**	0.11	0.23	0.24	0.24	1.20	0.23	0.25	1.20
# 5	0.04	0.22	0.10	0.11	**0.05**	0.12	0.12	0.13	1.27	0.12	0.15	1.27
# 6	0.06	0.19	0.06	0.14	0.07	**0.07**	0.08	0.08	1.30	0.07	0.12	1.29
# 7	0.11	0.15	0.06	0.19	0.13	0.06	**0.05**	0.08	1.33	0.08	0.10	1.33
# 8	0.16	0.18	0.10	0.23	0.17	0.09	0.11	**0.06**	1.34	0.06	0.13	1.33
# 9	0.26	0.24	0.20	0.32	0.28	0.19	0.20	0.15	**1.36**	0.16	0.22	1.36
# 10	0.12	0.19	0.08	0.19	0.13	0.07	0.09	0.04	1.31	**0.03**	0.13	1.30
# 11	0.04	0.21	0.09	0.11	0.06	0.10	0.11	0.10	1.28	0.09	**0.14**	1.27
# 12	0.05	0.24	0.12	0.10	0.05	0.14	0.14	0.15	1.26	0.14	0.17	**1.26**

(a)

EP	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12
# 1	**0.07**	0.20	0.18	0.12	0.30	0.23	0.15	0.24	0.24	0.34	0.09	0.20
# 2	0.28	**0.11**	0.16	0.21	0.11	0.14	0.18	0.10	0.11	0.10	0.24	0.12
# 3	0.20	0.14	**0.08**	0.18	0.21	0.19	0.14	0.15	0.17	0.21	0.17	0.13
# 4	0.15	0.15	0.12	**0.06**	0.20	0.12	0.10	0.16	0.15	0.28	0.12	0.19
# 5	0.32	0.19	0.20	0.21	**0.08**	0.12	0.20	0.13	0.14	0.18	0.28	0.22
# 6	0.19	0.10	0.11	0.10	0.12	**0.07**	0.09	0.08	0.09	0.20	0.16	0.14
# 7	0.17	0.13	0.14	0.12	0.21	0.18	**0.14**	0.16	0.18	0.24	0.15	0.12
# 8	0.22	0.06	0.09	0.14	0.11	0.09	0.11	**0.06**	0.08	0.15	0.18	0.10
# 9	0.27	0.12	0.16	0.19	0.10	0.11	0.17	0.10	**0.08**	0.16	0.23	0.17
# 10	0.35	0.16	0.21	0.29	0.17	0.21	0.25	0.16	0.17	**0.03**	0.31	0.17
# 11	0.41	0.22	0.27	0.35	0.20	0.26	0.30	0.21	0.22	0.07	**0.37**	0.22
# 12	0.14	0.11	0.11	0.11	0.21	0.16	0.11	0.15	0.16	0.23	0.11	**0.10**

(b)

EP	# 1	# 2	# 3	# 4	# 5	# 6	# 17	# 8	# 9	# 10	# 11	# 12
# 1	**0.10**	0.21	0.19	0.13	0.33	0.26	0.13	0.21	0.26	0.38	0.41	0.18
# 2	0.17	**0.07**	0.12	0.16	0.17	0.08	0.18	0.07	0.11	0.21	0.19	0.10
# 3	0.14	0.11	**0.08**	0.16	0.23	0.13	0.15	0.11	0.16	0.27	0.25	0.11
# 4	0.13	0.13	0.19	**0.07**	0.17	0.13	0.08	0.13	0.13	0.29	0.26	0.18

TABLE VII
AVERAGE SPECTRAL ANGLE MAPPER (SAM) MEASURING ENDMEMBER SIMILARITY FOR DIFFERENT UBS METHODS.

Method	**Average SAM**
LCMV-BCM	0,283
E-FDPC	0,103
Proposed	**0,092**

VII. DISCUSSION AND CONCLUSION

In this paper we presented a novel method to perform unsupervised band selection (UBS) with hyperspectral data, based on a probabilistic clustering framework. The contributions of the article are mainly three-fold.

1) New hyperspectral image descriptors have been proposed, consisting in local statistical measures over multiple-feature vectors. We therefore referred to this new image representation as multiple-feature local statistical descriptors (*MLSD*).
2) An ad-hoc pixel selection (PS) procedure has been introduced, in order to locally select the pixels subsequently used for the estimation of statistics.
3) A new kernel-based model for the posterior class probability has been defined on the base of the MLSD.

The main strength of the proposed statistical descriptors resides in their capability of providing a concise, yet rich, representation of a hypersprectral image, thus facilitating the process of selecting its most representative bands. The conciseness derives from the fact that the statistics are computed per block, according to a regular grid. Moreover, the process is made robust, thanks to the prior pixel selection step. The proposed posterior probability model is in the spirit of kernel density estimation (KDE). The different statistical descriptors are jointly considered by adopting a strategy that fuses different kernels. Unlike most of kernel-based methods in the literature that adopt Gaussian radial basis functions, we chose to employ an Epanechnikov kernel. The main reason behind this choice is that the Epanechnikov kernel, besides its theoretical optimality when doing KDE, has a naturally bounded support. This is particularly useful with hyperspectral data, where modeling relationships between distant bands (in terms of spectral content) with positive kernel weights might be meaningless. The proposed probabilistic model is flexible, in the sense that it offers a way to easily incorporate multiple image features, regardless of their nature and their quantity. The parameters of the model are optimized by maximizing an information measure, as done for those methods belonging to the family of information-maximization clustering. The information measure chosen, called Squared-loss Mutual Information (SMI), allows a closed-form solution that can be computed fast via an eigenvector problem. We then have a simple workflow, which, once a desired number of bands k is given as an input, returns the corresponding set of bands selected by the algorithm.

The performance of the proposed UBS method has been evaluated in terms of classification performance. By varying k, we evaluated the effect of reducing the spectral signatures of the image pixels rather than using the whole spectrum, when spectral vectors are used as features for classification. When compared with other state-of-the-art UBS methods, our proposed approach showed in general higher overall accuracy, thus meaning that the selected bands are more informative. In particular, we remarked that the pixel selection step is essential to obtain an extra performance gain. Furthermore, we assessed the effectiveness of the proposed method w.r.t. another application: endmember extraction. Here, the similarity between known ground-truth endmembers and the ones extracted after the band selection stage has been measured. As a higher similarity was observed in comparison with other state-of-the-art methods, we can conclude that our method select more pertinent bands, thanks to which endmembers can be more easily identified.

The new image descriptors proposed, the MLSD, are based on multiple image features, including spatial features (besides "raw" spectral intensity values). At present, the evaluation procedure is based on classification performance of spectral classification algorithms. In future work, we plan to also evaluate spatial-spectral classification methods for hyperspectral images. We believe that in that case, as the MLSD have been designed by taking into account spatial information too, the classification performance can have an even bigger benefit from a prior band selection with our UBS method. Moreover, we plan to investigate the possibility of using the MLSD, which have been proven to be good image descriptors for band selection, directly as classification features. This would require a modification to the way the MLSD are built, and a careful evaluation of the inevitably increasing computational cost.

REFERENCES

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. 110–122, 2009.

[2] M. T. Eismann, *Hyperspectral Remote Sensing*, ser. SPIE monograph. SPIE, 2012.

[3] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of Biomedical Optics*, vol. 19, no. 1, pp. 010 901–010 901, 2014.

[4] A. A. Gowen, Y. Feng, E. Gaston, and V. Valdramidis, "Recent applications of hyperspectral imaging in microbiology," *Talanta*, vol. 137, pp. 43–54, 2015.

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

[6] S. Chatterjee and A. S. Hadi, *Regression analysis by example*, 4th ed. John Wiley & Sons, 2015.

[7] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, p. 32, 2000.

[8] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 2, pp. 192–195, 2005.

[9] W. Liao, A. Pizurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 184–198, 2013.

[10] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[11] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," in *Advances in Neural Information Processing Systems*, 2000, pp. 568–574.

[12] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 55–66, 2014.

[13] T. J. Paciencia and K. W. Bauer, "Hyperspectral anomaly detection using enhanced global factors," in *Automatic Target Recognition XXVI*, vol. 9844. International Society for Optics and Photonics, 2016, pp. 1–11.

[14] M. Riedmann and E. J. Milton, "Supervised Band Selection for Optimal Use of Data from Airborne Hyperspectral Sensors," in *IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IGARSS)*. IEEE, 2003, pp. 1770–1772.

[15] H. Yang, Q. Du, H. Su, and Y. Sheng, "An Efficient Method for Supervised Hyperspectral Band Selection," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 138–142, Jan. 2011.

[16] X. Cao, T. Xiong, and L. Jiao, "Supervised Band Selection Using Local Spatial Information for Hyperspectral Image," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 3, pp. 329–333, Mar. 2016.

[17] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. G. Althouse, "A Joint Band Prioritization and Band-Decorrelation Approach to Band Selection for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[18] C.-I. Chang and S. Wang, "Constrained Band Selection for Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[19] A. Martínez-Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-Based Hyperspectral Band Selection Using Information Measures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[20] C. Conese and F. Maselli, "Selection of optimum bands from TM scenes through mutual information analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 48, no. 3, pp. 2–11, Jun. 1993.

[21] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, "Band Selection for Hyperspectral Image Classification Using Mutual Information," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 522–526, Oct. 2006.

[22] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Computer Vision*, vol. 3, no. 4, pp. 213–222, Dec. 2009.

[23] M. Ahmad, D. I. U. Haq, Q. Mushtaq, and M. Sohaib, "A new statistical approach for band clustering and band selection using K-means clustering," *International Journal of Engineering and Technology*, vol. 3, no. 6, pp. 606–614, 2011.

[24] S. Jia, Z. Ji, Y. Qian, and L. Shen, "Unsupervised Band Selection for Hyperspectral Imagery Classification Without Manual Band Removal," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 531–543, Apr. 2012.

[25] Y. Yuan, J. Lin, and Q. Wang, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1431–1445, 2016.

[26] M. Zhang, J. Ma, and M. Gong, "Unsupervised Hyperspectral Band Selection by Fuzzy Clustering With Particle Swarm Optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 773–777, May 2017.

[27] G. Zhu, Y. Huang, J. Lei, Z. Bi, and F. Xu, "Unsupervised Hyperspectral Band Selection by Dominant Set Extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 227–239, Jan. 2016.

[28] Y. Yuan, X. Zheng, and X. Lu, "Discovering Diverse Subset for Unsupervised Hyperspectral Band Selection," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 51–64, 2017.

[29] A. Datta, S. Ghosh, and A. Ghosh, "Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2814–2823, 2015.

[30] S. Jia, G. Tang, J. Zhu, and Q. Li, "A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 88–102, Jan. 2016.

[31] M. Bevilacqua and Y. Berthoumieu, "Unsupervised Hyperspectral Band Selection via Multi-feature Information-Maximization Clustering," in *IEEE International Conference on Image Processing (ICIP)*. Beijing, China: IEEE, Sep. 2017, pp. 540–544.

[32] L. He, J. Li, C. Liu, and S. Li, "Recent Advances on Spectral–Spatial Hyperspectral Image Classification: An Overview and New Guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[33] K. Sun, X. Geng, and L. Ji, "Exemplar Component Analysis: A Fast Band Selection Method for Hyperspectral Imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 998–1002, May 2015.

[34] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2008.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[36] F. V. Agakov and D. Barber, "Kernelized Infomax Clustering," in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2006, vol. 18, pp. 17–24.

[37] C. E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *The Annals of Statistics*, pp. 1152–1174, 1974.

[38] A. Krause, P. Perona, and R. G. Gomes, "Discriminative Clustering by Regularized Information Maximization," in *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2010, vol. 23, pp. 775–783.

[39] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya, "On Informa-tion-Maximization Clustering: Tuning Parameter Selection and Analytic Solution," in *28th International Conference on Machine Learning*, 2011, pp. 65–72.

[40] M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya, "Information-Maximization Clustering Based on Squared-Loss Mutual Information," *Neural Computation*, vol. 26, no. 1, pp. 84–131, Jan. 2014.

[41] Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Multiple Spectral–Spatial Classification Approach for Hyperspectral Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4122–4132, 2010.

[42] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in Spectral–Spatial Classification of Hyperspectral Images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[43] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized Composite Kernel Framework for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[44] A. Cherian and S. Sra, "Positive Definite Matrices: Data Representation and Applications to Computer vision," *Algorithmic Advances in Riemannian Geometry and Applications: For Machine Learning, Computer Vision, Statistics, and Optimization*, p. 93, 2016.

[45] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2464–2477, 2015.

[46] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, "Fast Algorithms for Projected Clustering," in *ACM SIGMoD Record*, vol. 28, no. 2. ACM, Jun. 1999, pp. 61–72.

[47] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed., ser. Series in Statistics. New York: Springer, 2001, vol. 1.

[48] M. Samiuddin and G. El-Sayyad, "On nonparametric kernel density estimates," *Biometrika*, vol. 77, no. 4, pp. 865–874, 1990.
[49] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2004, vol. 17, pp. 1601–1608.
[50] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.
[51] A. P. Witkin, "Scale-space filtering," in *Readings in Computer Vision*. Elsevier, 1987, pp. 329–332.
[52] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[53] F. Porikli and O. Tuzel, "Fast Construction of Covariance Matrices for Arbitrary Size Image Windows," in *IEEE International Conference on Image Processing (ICIP)*, 2006, pp. 1581–1584.
[54] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end–member determination in hyperspectral data," in *Imaging Spectrometry V*, vol. 3753. International Society for Optics and Photonics, 1999, pp. 266–276.