A Novel Band Selection Approach for Hyperspectral Image Classification using the Kolmogorov Variational Distance

Mohammed LAHLIMI[1], Mounir Ait KERROUM[2] and Youssef FAKHRI[3]
Laboratory of Research in Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

***Abstract*—In this paper, we introduce a novel band selection approach based on the Kolmogorov Variational Distance (KoVD) for Hyperspectral image classification. The main reason we are taking interest in KoVD is its unique relation to the classification error. Our previous works on band selection using the Mutual Information (MI), the Divergence Distance (DD), or the Bhattacharyya Distance (BD) inspire this study; thus, we are particularly interested in finding out how KoVD performs against these distances in terms of the numbers of band retained and the classification accuracy. All the distances in this study are modeled with the Gaussian Mixture Model (GMM) using the Bayes Information Criterion (BIC) / Robust Expectation-Maximization (REM). The experiments are carried on four benchmark Hyperspectral images: Kennedy Space Center, Salinas, Botswana, and Indian Pines (92AV3C). The results show that band selection based on the Kolmogorov Variational Distance performs better than BD and DD, meanwhile against MI the results were too close.**

Keywords*—*Band selection; Bayes Information Criterion (BIC); Bhattacharyya Distance; divergence distance; hyperspectral imaging; Kolmogorov Variation Distance; Gaussian Mixture Model (GMM); Robust Expectation Maximization (REM); remote sensing

I. Introduction

In hyperspectral imaging, sensors record data from hundreds of contiguous bands of the electromagnetic spectrum. However, the Hughes phenomenon [1] [2] [3] and computational complexity [4] are two problems that appear during the classification process. Due to the small sample size problem [5] and the large number of bands acquired from the sensors, the classifier won't be properly trained [3]. Therefore, dimensionality reduction is needed.

Two approaches for dimensionality reduction can be found in literature, band selection [6] [3] [7] [8] and band extraction [9] [10] [11] [12]. The aim for band extraction is to create a new reduced dataset from the existing one using a linear/non-linear transformation [6]. The Principal Component Analysis, Projection Pursuit, Independent Component Analysis, Orthogonal Subspace Projection, Segmented Principal Component Analysis, and others [13] [14] have been used to reduce the data volume. However, due to the linear/non-linear transformation, the original data are replaced by a new set of variables with no actual physical meaning [6] which can be a disadvantage in some applications. Band selection, on the other hand, tries to find an optimal subset from the original pool by only selecting relevant bands with valuable information for the classifier, through maximizing a class separability criterion [6]. Between band extraction and band selection, in this study, the later is the preferred one since, with band selection, the data remains unchanged and the physical meaning is preserved [15].

Band selection techniques can be broadly classified into two categories: wrapper and filter techniques. The wrapper approach [6] take advantage of the classifier itself and use it as the criterion for band selection [16], the result is a subset with a high classification score, however, the drawback of this technique is that the bias toward the used classifier. Unlike the wrapper approach, the filter [6][16] deploy metrics and distances to evaluate the bands without involving the classifier. In theory, the best criterion to measure the pertinence of a band is the Bayes error. However, the calculation of the Bayes error is, in general, a very complex problem [17]. Therefore, some approach seeks an upper bound of the error probability such as the Chernoff and Bhattacharyya bounds.

A new band selection approach is introduced in this paper, based on the Kolmogorov Variational Distance for Hyperspectral image classification. This work is a sequel on our previous research on band selection with Mutual Information [18], Bhattacharyya Distance [8] and Divergence Distance [19]. The primary interest in KoVD is the fact that is uniquely related to the classification error [20] [6], which is often difficult to estimate [17]. KoVD has been used in other fields such as signal selection, communication and radar systems [20] [21] but not in the hyperspectral imaging context.

To model the Kolmogorov Variational Distance, the Gaussian Mixture Model is used with The Expectation-Maximization (EM) algorithm [9], however, with the EM algorithm we face two issues: The first one is the choice of the number of components K as it can affect the estimation of the covariance matrix [8] and the second issue is the sensitivity to the initial values choice [22]. With a bad choice of K, we can easily end up with the Curse of Dimensionality. As a solution two approaches are proposed; a GMM based on the Bayes Information Criterion (BIC) and a Robust Expectation-Maximization (REM) algorithm [22].

Our main contributions in this study is a novel band selection approach with the Kolmogorov Variational Distance modeled with GMM-REM and GMM-BIC. To assess the performances of KoVD two criteria are being used: the numbers of the retained band and the classification accuracy. The experiments are performed on four hyperspectral benchmark datasets: The scene Indian Pines (92AV3C), Botswana scene, Kennedy space center scene, and Salinas scene.

This paper is structured as follows: Sections II and III describe the fundamentals and the proposed band selection algorithm. Section IV discusses the experimental results. Finally the conclusion in Section V.

II. Background

A. Kolmogorov Variational Distance

The Kolmogorov Variational Distance (KoVD) is the integral of the absolute difference between two posterior probabilities. It expresses the distance between the densities [6]. The main advantage for KoVD is its direct relation to the classification error [20] [6]. KoVD is expressed as follows [6]:

$$J_{KoVD}(\omega_1, \omega_2) = \int |P(\omega_1|x) - P(\omega_2|x)| P(x)\mathrm{d}x \tag{1}$$

KoVD provides an indication of the amount of probability mass by which the two distributions differ. If the classes ω_1 and ω_2 are similar, $P(\omega_1|x) = P(\omega_2|x)$ then J_{KoVD} will equal zero, and if the classes ω_1 and ω_2 are disjoint $P(\omega_1|x) = 0$ and $P(\omega_1|x) \neq 0$, J_{KoVD} will attains its maximum value [6].

In the case of multi-class problem, between each pairwise class (ω_i, ω_j), KoVD is computed as the average cost function.

$$J = \sum_i \sum_j P(\omega_i)P(\omega_j)J_{KoVD}(\omega_i, \omega_j) \tag{2}$$

B. Mutual Information

Given X and Y, two discrete random variables, the Mutual Information (MI) is the defined as [18]:

$$I(X;Y) = H(X) – H(X/Y) \tag{3}$$

$I(X;Y)$ expresses the information we gain by decreasing the uncertainty contained in the random variable X after knowing Y. With The entropy $H(X)$ of a random variable X and $H(X/Y)$ the conditional entropy of X given Y [18] [23].

C. Divergence Distance

The divergence distance (DD) [19] is a probabilistic distance that measure of the similarity between two classes ω_1 and ω_2 often used in information theory. DD is the sum of the two Kullback-Leibler divergences. Given $P(x|\omega_1)$ and $P(x|\omega_2)$, DD is defined as [6]:

$$J_{DD}(\omega_1, \omega_2) = \int [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} \mathrm{d}x \tag{4}$$

DD distance is interpreted as the amount of information necessary to change the prior probability distribution into posterior probability distribution [24]. In the case of multi-class problem, between each pairwise class (ω_i, ω_j), DD is computed as the average cost function according to equation (2).

D. Bhattacharyya Distance

The Bhattacharyya distance (BD) [8] is a similarity measurement of the scatter degree of two classes ω_1 and ω_2. The bhattacharyya distance is expressed as [6]:

$$J_{BD}(\omega_1, \omega_2) = -\log \int (p(x|\omega_1)p(x|\omega_2))^{\frac{1}{2}} \mathrm{d}x \tag{5}$$

In the case of multi-class problem, between each pairwise class (ω_i, ω_j), BD is computed as the average cost function according to equation (2).

E. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) models the density as the sum of one or more weighted Gaussian components [25] [8]. For a GMM, a probability density function is the sum of K Gaussian components:

$$p(x|\omega) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k) \tag{6}$$

where K the number of mixture component, π_k the mixing weight ($0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$) and $p(x|\mu_k, \Sigma_k)$ a d-dimensional gaussian distribution

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right] \tag{7}$$

With μ_k the mean and Σ_k the covariance matrix for the k^{th} component.
The parameters $\{\pi c, \mu_c, \Sigma_c\}$ are usually estimated by the EM algorithm [9].

III. Band Selection by Kolmogorov Variational Distance

Given a set of band $F = \{b_i\}_{i=1}^{d}$, the goal is to find an optimal subset $S = \{b_i'\}_{i=1}^{d'}, S \subset F, d' \leq d$ that only keeps the relevant bands that contribute to the classification task while discarding any redundancy. An exhaustive search for the optimal subset S can be impractical from a computational viewpoint, and the Sequential forward selection (SFS) is one of the simplest search strategy [26] [18]. With an empty set of bands S at the beginning, we start to add sequentially the bands that maximizes the KoVD cost function until the desired number of band is achieved, or no longer maximize the cost-function. SFS algorithm have a relatively low computational burden [27].

The algorithm (Fig. 1) is the same as [18] [28] [29] except the computation of the Mutual Information as a cost function between multiple variables. Instead, KoVD is used as a criterion between multiple bands to select the salient ones for hyperspectral image classification.

A. Bayes Error

In theory, the best criterion to measure the pertinence of a band is the Bayes error. The lower the error the better. However, the calculation of the Bayes error is, in general, a very complex problem [17] and it is often difficult to evaluate its probability.

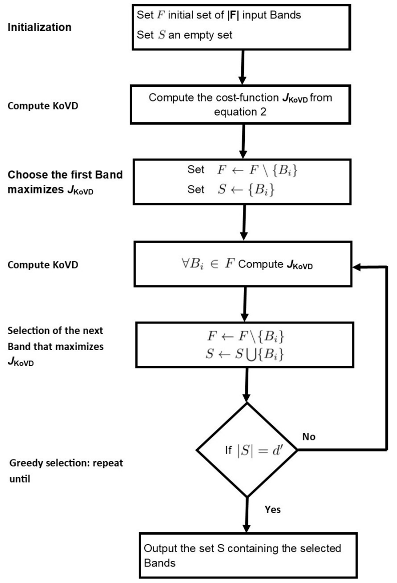

Initialization
Set F initial set of |F| input Bands
Set S an empty set
Compute KoVD
Compute the cost-function J_{KoVD} from equation 2
Choose the first Band maximizes J_{KoVD}
Set $F \leftarrow F \setminus \{B_i\}$
Set $S \leftarrow \{B_i\}$
Compute KoVD
$\forall B_i \in F$ Compute J_{KoVD}
Selection of the next Band that maximizes J_{KoVD}
$F \leftarrow F \setminus \{B_i\}$
$S \leftarrow S \bigcup \{B_i\}$
Greedy selection: repeat until
If $|S| = d'$
No
Yes
Output the set S containing the selected Bands

Fig. 1. The Band Selection Algorithm by Kolmogorov Variational Distance

In m-class case, the Bayes Error Probability e is given as [30]:

$$e = \int [1 - max_i P(\omega_i|x)] P(x) \mathrm{d}x \tag{8}$$

where the posterior probability

$$\begin{aligned} P(\omega_i|x) &= \frac{P(x|\omega_i) \times P(\omega_i)}{P(x)} \\ P(x) &= \sum_{c=1}^{C} P(x|\omega_c) \times P(\omega_c) \end{aligned} \tag{9}$$

$P(x)$ is the mixture density function and $P(\omega_i)$ is the prior probability

A direct calculation of equation (8) in general is often impossible or impractical [17]. In two class case, the Error Probability can be expressed as:

$$e = \frac{1}{2}\left\{1 - \int |P(\omega_1|x) - P(\omega_2|x)| P(x) \mathrm{d}x\right\} \tag{10}$$

The Kolmogorov Variational Distance is the integral from the equation (10) . From equation (1) and (10), the Error Probability e can be expressed as:

$$e = \frac{1}{2}\left\{1 - J_{KoVD}(\omega_1, \omega_2)\right\} \tag{11}$$

From equation (11) we can notice that KoVD can be expressed in terms of classification error. It has a direct relation to Bayes Error Probability, which is its main advantage unlike other probabilistic distances that only provides a bound on the error. However, KoVD requires an estimate of a probability density function and its numerical integration, which can restricts its usefulness in many practical situations [6].

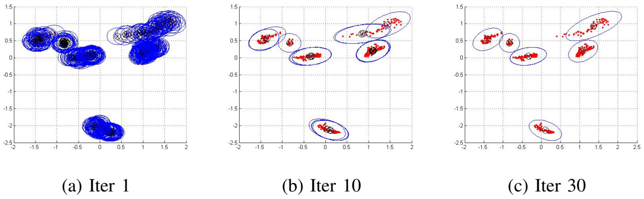

(a) Iter 1
(b) Iter 10
(c) Iter 30

Fig. 2. Robust Expectation Maximization Implementation on Synthetic Data (a) All Data Points Initialization of the REM Algorithm using -(b) Reducing the Number of Clusters -(c) Finding the Optimum Number of Clusters $k = 6$.

B. KoVD Based on Gaussian Mixture Model

The Kolmogorov Variational distance based on Gaussian Mixture Model can be expressed as follows:

$$J_{KoVD}(\omega_i, \omega_j) = \sum |P(\omega_i|x) - P(\omega_j|x)| P(x) \tag{12}$$

With the equation (6) and the equation (9), the KoVD can be expressed as:

$$\begin{aligned} &J_{KoVD}(\omega_i, \omega_j) = \sum |P(x|\omega_i) \times P(\omega_i) - P(x|\omega_j) \times P(\omega_j)| \\ &= \sum | \sum_{ki=1}^{Ki} \pi_{ki} p(x|\mu_{ki}, \Sigma_{ki}) \times P(\omega_i) \\ &- \sum_{kj=1}^{Kj} \pi_{kj} p(x|\mu_{kj}, \Sigma_{kj}) \times P(\omega_j)| \end{aligned} \tag{13}$$

To compute our cost-function $J_{KoVD}(\omega_1, \omega_2)$ from equation (13) we need to estimate the following parameters: the number of clusters K, the covariance matrix Σ, the mean μ and the mixing coefficient π.

With GMM the challenge is the estimation of the parameters π, μ, Σ, K, the first three the parameters can be estimated with the Expectation-Maximization (EM) algorithm [16]. And K the number of components, the fourth parameter, is user-defined and has to be given a priori. Choosing the right value for the number of components K is crucial since it has a direct effect on the estimation of the covariance matrix. With a bad choice of K, ill-conditioned covariance matrices can be formed and the Curse of Dimensionality then can't be avoided [5] [2].

To overcame this challenge, we pursuit two approaches in order to define the optimal value for the parameter K. The first one is based on the Bayes Information Criterion (BIC) [31]. BIC is a popular measure for comparing maximum likelihood models and the model with the smallest value is the preferred one [32] [33]. BIC was introduced by [34], and defined as:

$$BIC = -2 \times \ln(likelihood) + \ln(N) \times k \tag{14}$$

With k the number of parameters estimated and N the number of observations.

The second approach is the Robust Expectation-Maximization (REM) algorithm [22]. The main advantage of this algorithm is its ability to find an optimal number of clusters K automatically, thus the number of the component will no longer have to be defined a priori. REM also solves the issue of the initial value of the standard EM algorithm, the problem of choosing cluster centers. At first, the Robust Expectation-Maximization algorithm uses all data points as centers, and from there try to automatically reach an optimal number of clusters by discarding the clusters that do not meet the required criteria (see Fig. 2); For more detail about the algorithm, see [22].

C. Regularization Problem

For the estimation of the covariance matrix in hyperspectral imaging, the ”Hughes phenomenon” and the singularity problems [25] are usually caused by the small sample size datasets. And by partitioning the already small dataset we can easily end up with an ill-conditioned mixture model [35]. For each component the sample size must not be less than the dimensionality of the data [25], since the covariance matrix should be invertible in order to compute equation (7). For Gaussian Mixture Model, the curse of dimentionality is primarily related to the estimation of the covariance matrix [36], and regularization techniques are one way around this problem:

1) Leave One Out Covariance (LOOC) : To avoid the singularity problem the LOOC estimator can be used to regulate the covariance matrix [37] [25] [3] [36]. Let S and $diag(S)$ be respectively the covariance matrix its diagonal version:

$$S_i^{looc}(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & \text{if } 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & \text{if } 1 \le \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & \text{if } 2 \le \alpha_i \le 3 \end{cases} \tag{15}$$

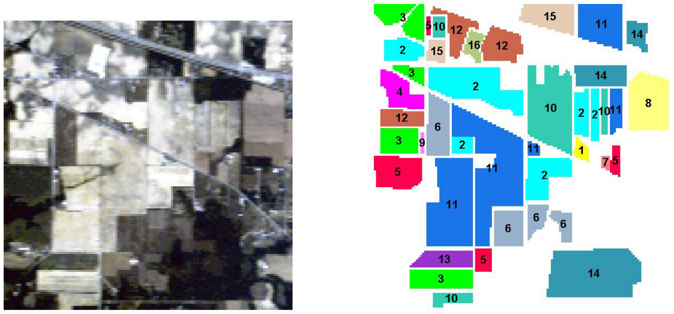
The LOOC estimator evaluate several values of α_i, and the value that maximize the average log likelihood of the Gaussian density is the optimal choice [37]. In our case, since we are using an iterative approach to select bands, using this regularization techniques as described by equation (15), did add to the complexity of the algorithm and to the computation time.

2) Maximum Entropy Covariance Selection (MECS): The MECS method deals directly with singular and unstable covariance matrices; rather than optimizing the group likelihood or the classification accuracy, MECS maximize the information under an incomplete and consequently uncertain context [38]. We are particularly interested in this method since according to [38], an optimization procedure isn’t required, whenever covariance matrices are ill-posed or poorly estimated, and finally it has a much lower computational cost while performing as well as any other method.

IV. Experimental Study

A. Dataset

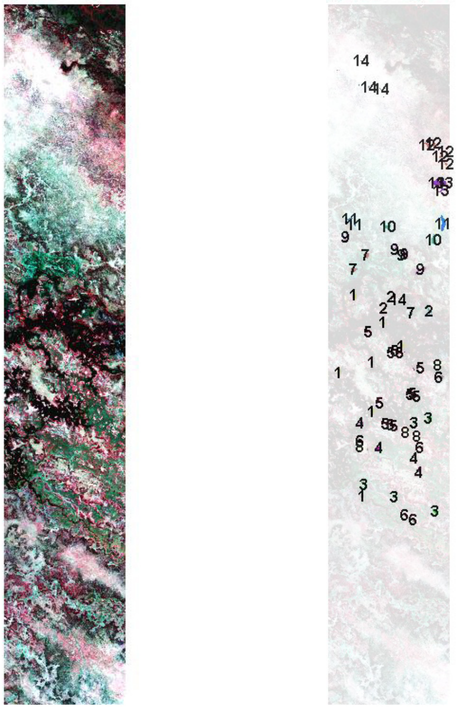
1) Indian Pines dataset: This scene was firstly used by David Landgrebe and his students [25] [39] [37] [2] and since become a benchmark dataset. Indian Pines dataset also known as 92AV3C dataset is a 145×145 pixels by 224 bands hyperspectral image scene captured over the Indian Pines test site in North-western Indiana on June 12, 1992, by AVIRIS sensor, with a spatial resolution of $18m$. Fig. 3 is a false-color composite of the Indian Pines scene and its ground truth map, and Table I describe the dataset.



(a) (b)

Fig. 3. (a) False Color Composite Image of Indian Pines Dataset and (b) Ground Truth.

2) Botswana dataset: In 2001-2004, over the Okavango Delta Botswana, the NASA EO-1 satellite with the Hyperion sensor gathered a sequence of data over a strip of $7.7km$. Fig. 4 is a false-color composite of the Botswana dataset and its ground truth map. The UT Center for Space Research Preprocessed the data, 14 classes were identified from the observations as described in Table II, and 145 bands were retained $[10-55, 82-97, 102-119, 134-164, 187-220]$ after removing the uncalibrated and noisy bands.



(a) (b)

Fig. 4. (a)False Color Composite Image of the Botswana Dataset and (b)Ground Truth Map.

3) Kennedy Space Center (KSC): This scene was acquired by the NASA AVIRIS, on March 23, 1996, over the Kennedy Space Center (KSC), Florida. The dataset got 176 bands after

TABLE I. Data description of the Indian Pines 16 class full scene

Class number	Class name	Total samples	Training samples	Test samples
1	Alfalfa	54	29	25
2	Corn-notill	1434	719	715
3	Corn-mintill	834	419	415
4	Corn	234	117	117
5	Grass-pasture	497	249	248
6	Grass-trees	747	374	373
7	Grass-pasture-mowed	26	13	13
8	Hay-windrowed	489	243	246
9	Oats	20	10	10
10	Soybean-notill	968	483	485
11	Soybean-mintill	2468	1234	1234
12	Soybean-clean	614	304	310
13	Wheat	212	108	104
14	Woods	1294	644	650
15	Buildings-Grass-Trees-Drives	380	190	190
16	Stone-Steel-Towers	95	46	49

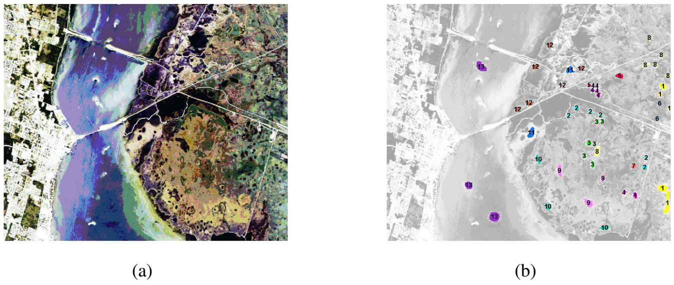
TABLE II. Data description of the Botswana dataset

Class number	Total samples	Training samples	Test samples
1	270	133	137
2	101	52	49
3	251	126	125
4	215	106	109
5	269	134	135
6	269	133	136
7	259	132	127
8	203	104	99
9	314	158	156
10	248	127	121
11	305	152	153
12	181	89	92
13	268	137	131
14	95	46	49

TABLE III. Data description of the Kennedy Space Center dataset

Class number	Total samples	Training samples	Test samples
1	761	373	388
2	243	122	121
3	256	128	128
4	252	127	125
5	161	79	82
6	229	116	113
7	105	51	54
8	431	214	217
9	520	259	261
10	404	203	201
11	419	211	208
12	503	250	253
13	927	465	462

removing low SNR bands with 18 m spatial resolution, and 13 classes of various types of land cover. Fig. 5 is a false-color composite of the Kennedy Space Center (KSC) scene and its ground truth map, and further details of the dataset are given in Table III


(a)

(b)

Fig. 5. The Kennedy Space Center Data Set. (a) False Color Composite Image and (b) Ground Truth.

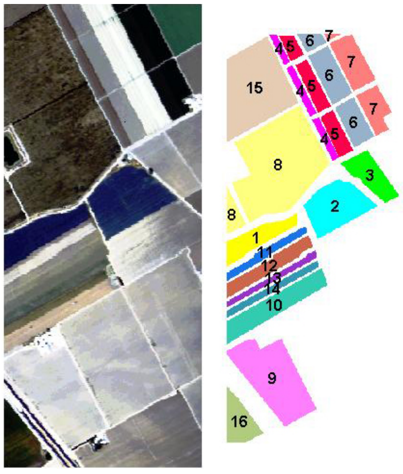
4) Salinas: Over Salinas Valley in California, the AVIRIS sensor collected 512×217 pixels by 224 bands with a spatial resolution as high as 3.7-meter per pixel. Including vineyard fields, bare soils, and vegetables, 16 classes were defined in the dataset and 20 water absorption bands were removed [108-112], [154-167], 224. Fig. 6 is a false-color composite of Salinas scene and its ground truth map.

B. Experimental Setup

The band selection approach using the Kolmogorov Variational Distance was tested using the following hardware setup: a 64-bit PC (i7-2.20GHz) with 6 GB RAM and Matlab (R2014a). The experiment was run on four benchmark hyperspectral images: the Indian Pine (92AV3C), Salinas, Kennedy Space Center, and Botswana datasets. For classification purposes, the dataset is split into two halves of training/testing. The selected bands are fed to classifiers in order de show their classification performances. The used classifier is SVM through the LIBSVM library with RBF as kernel function and the grid search technique to find the C and γ parameters [40].

C. Results and Discussions

To evaluate our proposed approach, tests were run on the benchmark dataset Indian Pine, as this scene has been often used in various studies such as [25] [39] [37] [2]. In the first



(a) (b)

Fig. 6. Salinas Data Set. (a) False Color Composite Image and (b) Ground Truth.

experiment, we measure each band independently from the rest and see how it ranks in terms of class separability according to our KoVD cost-function. The higher the value the more the classes are separable on that band. Fig. 7 we do notice that band region $170 \sim 190$ have the highest value.

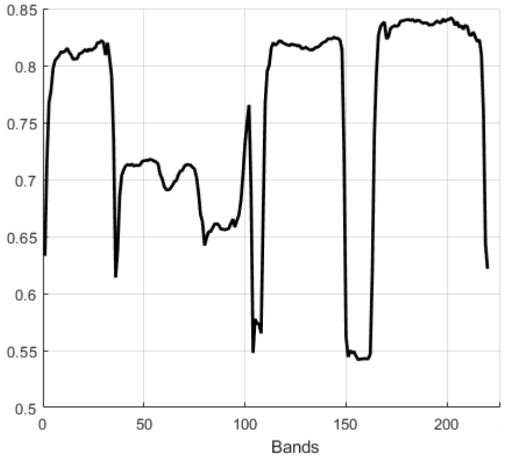
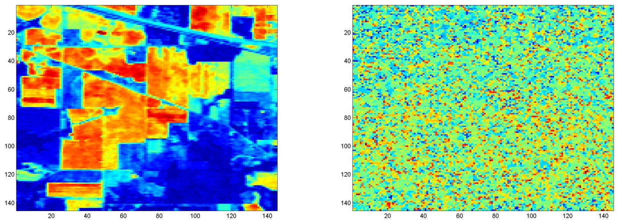


Fig. 7. KoVD Score for Each Band for 92AV3C Dataset.

In previous studies on the Indian Pine dataset [39] [41] [42] the bands $[104 - 108, 150 - 163, 220]$ were reported to be in the water absorption region with no useful information just noise as it can be seen in Fig. 7 and Fig. 8 as they got the lowest value. Hence, band selection with KoVD can successfully measure the pertinence of a band and discard those with no valuable information from the selection process.



(a) (b)

Fig. 8. (a) Scene at Band Number 168 in 92AV3C Dataset, (b) Scene at Band Number 153 in 92AV3C Dataset.

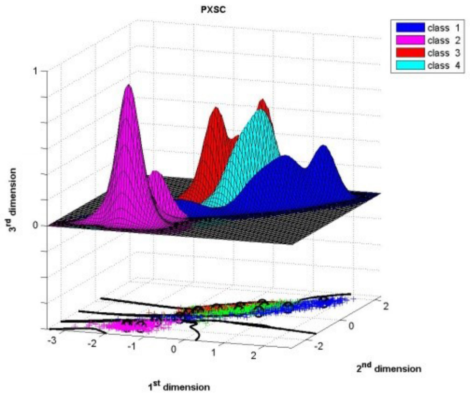
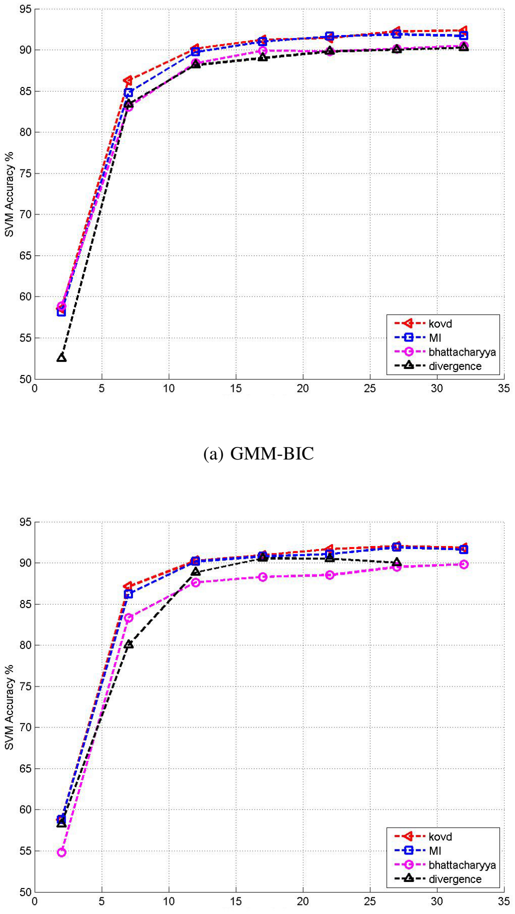


Fig. 9. Density Estimation by GMM and the Decesion Boundry that Separate between the 4 Classes.

For the second experiment, the goal is to measure the performance of the KoVD band selection algorithm with just the first two selected bands and to answer the question of whether KoVD modeled with GMM can separate classes successfully or not. For easier visual inspection the experiment was carried out on a portion of the benchmark dataset Indian Pine working with 4 classes instead of 16 similar to [39] [41]. The data as seen in Fig. 9 is highly correlated, nonetheless, we were able to separate one class from the rest with just 2 bands out of 220 with an SVM classification score of 81.92%. On the other hand, the other classes are still correlated thus the need to add more bands to achieve the desired result. For this sub-scene, a classification score accuracy of 93.81% with SVM is achieved with only five bands and a classification score of 97.04% at dimension thirty-six. Thus, the KoVD criterion modeled with GMM can be used as a class separability measurement for band selection.

In the next step, we are going to compare the performances of KoVD against its peers - the mutual information, the divergence, and Bhattacharyya distances - in terms of classification score and the number of retained bands. Due to the complexity of the dataset, all the probabilistic distances were computed through the Gaussian mixture model. The probability estimation is computed with GMM-BIC and the GMM-REM approach, meanwhile, the SVM is used as a classifier.

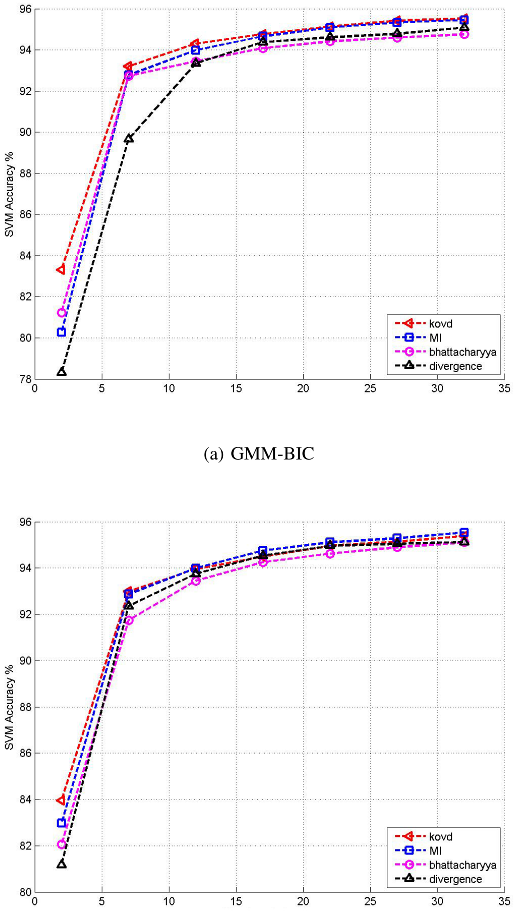
In Fig. 10, 11, 12 and 13 we do notice that, for the Indian Pine dataset, the SVM classification Score for the selected bands with KoVD performs better than the ones selected with the Bhattacharyya and Divergence distances. Meanwhile, compared to the mutual information, in terms of classification Accuracy, KoVD is slightly better, in fact, the curves almost overlap each other.



(a) GMM-BIC

(b) GMM-REM

Fig. 10. Overall Classification Accuracy of the Selected Bands for Dataset 92AV3C using (a) GMM-BIC, (b) GMM-REM



(a) GMM-BIC

(b) GMM-REM

Fig. 11. Overall classification Accuracy of the Selected Bands for Dataset Salinas using (a) GMM-BIC, (b) GMM-REM.

Depending on the number of the selected bands, on how well the GMM was estimated, on how well the classifier parameter was chosen and on the data set itself how correlated it is and how its post-treatment was to deal with the outliers, we do notice that KoVD performs the best at times and others times the MI. According to Fig. 10, 11, 12 and 13, the results are close to each other and the margin between the classification curves of the selected bands with both distances is not wide enough to concur on the superiority of one on the others. Therefore, it is hard to decide which one of the distances is the best. Thus, we can conclude that in our setup, the KoVD performs as well as the MI and both of them perform better than the Divergence and Bhattacharyya Distances.

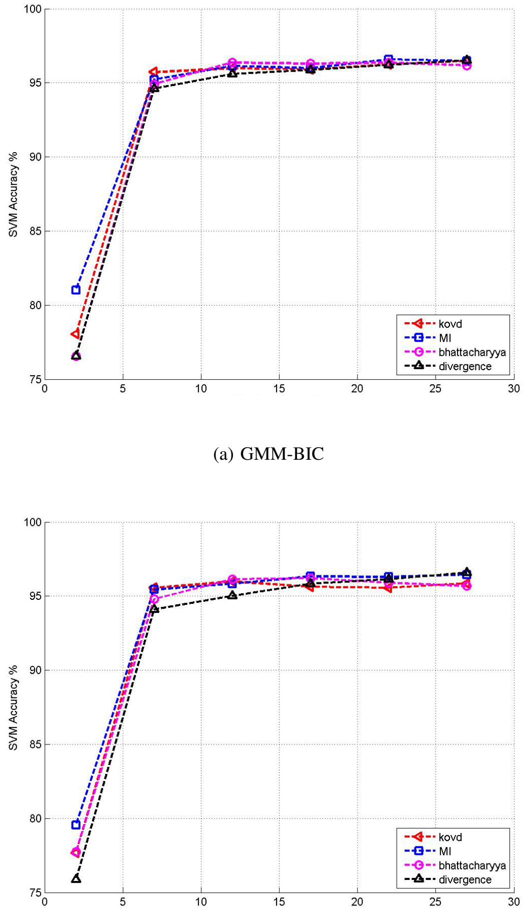
V. Conclusion

In this paper, a novel band selection approach based on the Kolmogorov Variational Distance for Hyperspectral image classification was introduced. The first experiment performed on the Indian Pine dataset have proved the efficiency and reliability of the KoVD criterion as a similarity measure. KoVD can measure the pertinence of a band, thus given a hyperspectral image dataset we can cluster the optimal bands while discarding those with no relevant information. This study was inspired by our previous work on the MI, BD, and DD. Thus, we were particularly interested in finding out how KoVD performs against these distances in terms of the numbers of bands retained and the classification accuracy. The experimental study showed that KoVD performs better than BD and DD, meanwhile against MI the results were too close; therefore, in the current setup, it is hard to decide which one is the best. Thus we can conclude that the KoVD performs as well as the MI.

References

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

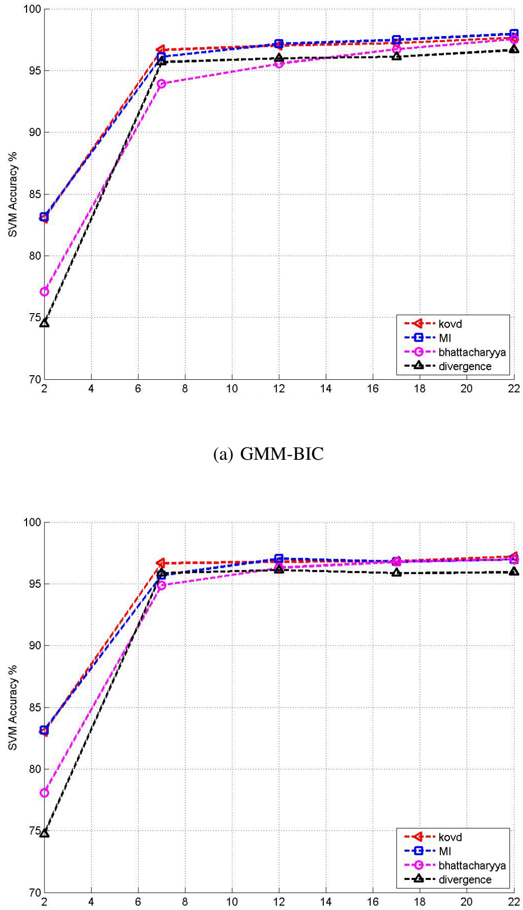
[2] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 1, pp. 39–54, Feb 1998.



(a) GMM-BIC

(b) GMM-REM

Fig. 12. Overall Classification Accuracy of the Selected Bands for Dataset KSC using (a) GMM-BIC, (b) GMM-REM



(a) GMM-BIC

(b) GMM-REM

Fig. 13. Overall Classification Accuracy of the Selected Bands for Dataset Botswana using (a) GMM-BIC, (b) GMM-REM

[3] J. Richards, *Remote Sensing Digital Image Analysis: An Introduction*. Springer Berlin Heidelberg, 2012. [Online]. Available: https://books.google.co.ma/books?id=ETfwQnBMP4UC

[4] A. Datta, S. Ghosh, and A. Ghosh, "Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination," *International Journal of Remote Sensing*, vol. 35, no. 2, pp. 554–577, 2014. [Online]. Available: https://doi.org/10.1080/01431161.2013.871392

[5] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, Sep 1994.

[6] A. R. Webb, *Statistical pattern recognition*, 2nd ed. John Wiley & Sons, 2003.

[7] J. Wang, X. Wang, K. Zhang, K. Madani, and C. Sabourin, "Morphological band selection for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1259–1263, Aug 2018.

[8] M. Lahlimi, M. Ait Kerroum, and Y. Fakhri, "Band selection with bhattacharyya distance based on the gaussian mixture model for hyperspectral image classification," in *Recent Advances in Electrical and Information Technologies for Sustainable Development*, S. El Hani and M. Essaaidi, Eds. Cham: Springer International Publishing, 2019, pp. 87–94.

[9] W. L. Martinez and A. R. Martinez, *Computational statistics handbook with MATLAB*. CRC press, 2007, vol. 22.

[10] M. Imani and H. Ghassemian, "Two dimensional linear discriminant analyses for hyperspectral data," *Photogrammetric Engineering & Remote Sensing*, vol. 81, no. 10, pp. 777–786, 2015.

[11] A. Datta, S. Ghosh, and A. Ghosh, "Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis," *International Journal of Remote Sensing*, vol. 38, no. 3, pp. 850–873, 2017. [Online]. Available: https://doi.org/10.1080/01431161.2016.1271470

[12] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "Feature extraction for hyperspectral image classification," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec 2017, pp. 379–382.

[13] P. K. Varshney and M. K. Arora, *Advanced image processing techniques for remotely sensed hyperspectral data*. Springer Science & Business Media, 2004.

[14] K. Burgers, Y. Fessehatsion, S. Rahmani, J. Seo, and T. Wittman, "A comparative analysis of dimension reduction algorithms on hyperspectral data," *LAMDA Research Group*, pp. 1–23, 2009.

[15] C. Lee, D. Landgrebe *et al.*, "Feature extraction based on decision boundaries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 4, pp. 388–400, 1993.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, ser. Computer science and scientific computing. Elsevier Science, 2013. [On-

line]. Available: https://books.google.co.ma/books?id=BIJZTGjTxBgC

[18] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on gaussian mixture model for multispectral image classification," *Pattern Recogn. Lett.*, vol. 31, no. 10, pp. 1168–1174, Jul. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2009.11.010

[19] M. LAHLIMI, "Band selection by divergence distance based on gaussian mixture model for hyperspectral image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, pp. 2330–2338, 10 2019.

[20] C. Chen, "Theoretical comparison of a class of feature selection criteria in pattern recognition," *IEEE Transactions on Computers*, no. 9, pp. 1054–1056, 1971.

[21] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.

[22] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust em clustering algorithm for gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, 2012.

[23] A. Elmaizi, H. Nhaila, E. Sarhrouni, A. Hammouch, and N. Chafik, "A novel approach for dimensionality reduction and classification of hyperspectral images based on normalized synergy," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100831

[24] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[25] B.-C. Kuo and D. A. Landgrebe, "A robust classification procedure based on mixture classifiers and nonparametric weighted feature extraction," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 11, pp. 2486–2494, 2002.

[26] C. Chang, *Hyperspectral Data Exploitation: Theory and Applications*. Wiley, 2007. [Online]. Available: https://books.google.co.ma/books?id=NwVncgNwtI4C

[27] L. Burrell, O. Smart, G. K. Georgoulas, E. Marsh, and G. J. Vachtsevanos, "Evaluation of feature selection techniques for analysis of functional mri and eeg." in *DMIN*, 2007, pp. 256–262.

[28] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[29] N. Kwak and Chong-Ho Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, Jan 2002.

[30] U. Maulik, S. Bandyopadhyay, and J. Wang, *Computational Intelligence and Pattern Analysis in Biology Informatics*, ser. Wiley Series in Bioinformatics. Wiley, 2011. [Online]. Available: https://books.google.co.ma/books?id=9CBeyg2AQ4EC

[31] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using gaussian mixture models and markov random fields," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 153–157, 2014.

[32] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[33] H. D.-G. Acquah, "Comparison of akaike information criterion (aic) and bayesian information criterion (bic) in selection of an asymmetric price relationship," *Journal of Development and Agricultural Economics*, vol. 2, no. 1, pp. 001–006, 2010.

[34] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[35] M. M. Dundar and D. Landgrebe, "A model-based mixture-supervised classification approach in hyperspectral data analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 12, pp. 2692–2699, 2002.

[36] M. Fauvel, C. Dechesne, A. Zullo, and F. Ferraty, "Fast forward feature selection for the nonlinear classification of hyperspectral images," *arXiv preprint arXiv:1501.00857*, 2015.

[37] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 2113–2118, July 1999.

[38] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa, "A new covariance estimate for bayesian classifiers in biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 2, pp. 214–223, 2004.

[39] S. Tadjudin and D. A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 439–445, 2000.

[40] S. M and G. Sadashivappa, "Hyperspectral image classification using support vector machine with guided image filter," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0101038

[41] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*. John Wiley & Sons, 2009.

[42] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, Oct 2016.