A LAYERED SPARSE ADAPTIVE POSSIBILISTIC APPROACH FOR HYPERSPECTRAL IMAGE CLUSTERING

Spyridoula D. Xenaki†*, Konstantinos D. Koutroumbas*†*, Athanasios A. Rontogiannis*†*, Olga A. Sykioti*†

†IAASARS, National Observatory of Athens, GR-152 36, Penteli, Greece
{ixenaki,koutroum,tronto,sykioti}@noa.gr

ABSTRACT

In this paper a new algorithm suitable for hyperspectral image clustering, called L-SAPCM, is proposed. The algorithm works in layers where at each layer, after suitable pre-processing, the SAPCM clustering algorithm ([1]) is applied. Preliminary results on real hyperspectral images show enhanced performance compared to other related methods.

Index Terms— layered clustering, sparsity, hyperspectral

1. INTRODUCTION

One of problems that have attracted considerable attention in hyperspectral image (HSI) processing is that of the classification of the various types of land cover/use. In addition, when ground truth information is not available, classification can only be unsupervised, which in this case is called clustering. Various clustering techniques such as the k-means and fuzzy c-means (FCM) have been applied to this problem.

In this paper, a new clustering technique called Layered-SAPCM (L-SAPCM) is proposed for HSI clustering. This performs a layered processing, where at each layer it uses as structural element a recently proposed clustering algorithm, called SAPCM [1], which belongs to the family of the possibilistic c-means (PCM) clustering algorithms. The two key features of SAPCM are that a) certain critical parameters are dynamically adapted, and b) sparsity is induced in the sense that each data point is forced to belong to only *a few* (or even *none*) of the clusters. These features make SAPCM flexible in tracking the underlying clustering structure.

Applying clustering to HSIs becomes much more challenging, due to a) their high dimensionality and b) the tendency of HSI pixels to form not clearly distinguishable clusters. To cope with the peculiarities of HSIs, the L-SAPCM algorithm is employed. Simulations on real HSI data show that the proposed algorithm outperforms other related state-of-the-art clustering techniques.

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: ARISTEIA - HSI-MARS - 1413.

2. THE SPARSE ADAPTIVE PCM (SAPCM)

Here, we briefly present an enhanced version of the SAPCM algorithm ([1]), which is "at the heart" of the proposed HSI clustering method, described in the next section. Let $X = \{\mathbf{x}_i \in \Re^\ell, i = 1, ..., N\}$ be a set of N, l-dimensional data vectors. Let also $\Theta = \{\boldsymbol{\theta}_j \in \Re^\ell, j = 1, ..., m\}$ be the set of m cluster representatives. In what follows, $\|.\|$ denotes the Euclidean norm. Let $U = [u_{ij}]$ be an $N \times m$ matrix whose (i, j) element stands for the, so called, *degree of compatibility* of $\mathbf{x}_i$ to the jth cluster, denoted by C_j and represented by the vector $\boldsymbol{\theta}_j$. Let also $\mathbf{u}_i{}^T = [u_{i1}, ..., u_{im}]$ be the vector containing the elements of the ith row of U with $u_{ij} \in [0, 1]$. Sparsity is imposed on $\mathbf{u}_i$'s via penalization of the PCM cost function with the l_p norm with $0 < p < 1$, i.e., $J(\Theta, U) = \sum_{i=1}^N [\sum_{j=1}^m u_{ij}\|\mathbf{x}_i - \boldsymbol{\theta}_j\|^2 + \sum_{j=1}^m \eta_j(u_{ij} \ln u_{ij} - u_{ij})] + \lambda \sum_{i=1}^N \sum_{j=1}^m u_{ij}^p$, where λ is a user-defined penalizing factor controlling sparsity and η_j, loosely speaking, is a measure of how much the *influence* of cluster C_j is spread around its respective representative $\boldsymbol{\theta}_j$. SAPCM described in **Alg. 1**, is derived via the minimization of $J(\Theta, U)$ w.r.t. $\boldsymbol{\theta}_j$'s and u_{ij}'s. Solving $\frac{\partial J}{\partial \boldsymbol{\theta}_j} = 0$ w.r.t. $\boldsymbol{\theta}_j$, we obtain the updating equation for $\boldsymbol{\theta}_j$ shown in **Alg. 1**. The updating for u_{ij}'s results from the solution of $\frac{\partial J}{\partial u_{ij}} = 0$. It turns out that this equation has at most two solutions lying in $[0, 1]$ and the solution that minimizes J is either 0 (implying sparsity) or the largest of the two solutions ([1]). Thus u_{ij} is set to that value.

In SAPCM the *initialization* of an overestimated number of $\boldsymbol{\theta}_j$'s and their corresponding η_j's comes after a proper normalization of X and it is based on a fast approximate variation of the Max-Min algorithm (see **Alg. 1**). This increases the probability to place each $\boldsymbol{\theta}_j$ close to a "dense in data" region. In *Cluster elimination* part of **Alg. 1**, the reduction of the number of clusters m down to its actual value is achieved by examining if there is even one data point that is most compatible with the cluster C_j. If this is the case, C_j is preserved. Otherwise, C_j is eliminated. In *Adaptation of η_j's* part of **Alg. 1**, the parameter η_j of C_j is estimated as the mean value of the Euclidean distances of the most compatible to C_j data vectors (of plurality n_j) from their mean vector $\boldsymbol{\mu}_j$. SAPCM returns the clusters C_j's of the underlying clustering structure.

Algorithm 1 $[C_1, ..., C_M] = SAPCM(X, \lambda, m)$

$t = 0$

Compute $\boldsymbol{\theta}'_j$, $j = 1, ..., m$: *via the* ***Max-Min Algorithm***
Compute η'_j: **Set:** $\eta'_j = \min_{\boldsymbol{\theta}'_s \neq \boldsymbol{\theta}'_j} \|\boldsymbol{\theta}'_j - \boldsymbol{\theta}'_s\|/2$, $j = 1, ..., m$
Initialization: **Normalize:** $X = X/\gamma$, $\boldsymbol{\theta}_j(0) \equiv \boldsymbol{\theta}_j = \boldsymbol{\theta}'_j/\gamma$, $\eta_j(0) \equiv \eta_j = \eta'_j/\gamma$, $j = 1, ..., m$, where $\gamma = \min_{k=1,...,m} \eta'_k$

Repeat:

Update U: As described in the text
$t = t + 1$
Update Θ: $\boldsymbol{\theta}_j(t) = \frac{\sum_{i=1}^{N} u_{ij}(t-1)\mathbf{x}_i}{\sum_{i=1}^{N} u_{ij}(t-1)}$, $j = 1, ..., m$
Cluster elimination:
Determine: $u_{ir} = \max_{j=1,...,m} u_{ij}$, $i = 1, ..., N$
If there is no $\mathbf{x}_i$ that is most compatible with C_j
then Remove C_j **end**
Data labeling: **If** $u_{ir} \neq 0$ **then Set:** $label(i) = r$
else Set: $label(i) = 0$ **end**
Adaptation of η_j*'s*:
$\eta_j(t) = \frac{1}{n_j(t)} \sum_{\mathbf{x}_i : u_{ij}(t) = \max_{r=1,...,m} u_{ir}(t)} \|\mathbf{x}_i - \boldsymbol{\mu}_j(t)\|$

Until: a termination criterion is met
Assign each unassigned point[1] to its closest among the m formed clusters $(C_1, ..., C_m)$
$C_j = \{\mathbf{x}_i \in X : label(i) = j, i = 1, ..., N\}$, $j = 1, ..., m$, where m is now the final number of clusters

3. LAYERED SAPCM FOR HSI CLUSTERING

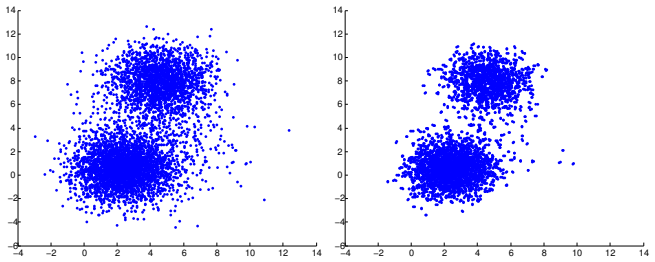
In HSIs, the number of image pixels, N, as well as the number of spectral bands, l, are usually very large. This increases dramatically both processing complexity and memory requirements. Taking into account, however, that contiguous HSI bands are usually highly correlated [2], computational complexity can be reduced by removing the redundancy introduced by the spectral information. To this end, we apply principal component analysis (PCA) as a first pre-processing step. As a result, the dimension l is dramatically reduced.

Algorithm 2 $[X\_cleared] = data\_purifying(X)$

Determine: $d_{\min}(i) = \min_{\mathbf{x}_s \in X - \{\mathbf{x}_i\}} \|\mathbf{x}_i - \mathbf{x}_s\|^2$, $i = 1, ..., N$
Compute: $\mu = \frac{1}{N} \sum_{i=1}^{N} d_{min}(i)$
Set: $X\_cleared = \{\mathbf{x}_i \in X : d_{min}(i) < \mu, i = 1, ..., N\}$

Another serious problem, frequently met in HSIs, is that the pixels are grouped to not very well distinguished "clouds". Thus, direct application of density-based clustering algorithms (such as SAPCM), could lead to poor clustering results. To face this problem, a pre-processing step, which removes the pixels that are not "too close" to the physical cluster centers, unravels the "cores" of the clusters, which are expected to be better distinguished. This can be achieved by first determining the mean of the distances of all pixels from their nearest neighbor and then removing those pixels whose distance from their nearest neighbor is larger than the mean (**Alg. 2**). As shown in Fig. 1, this pre-processing step allows clusters to be better distinguished, assisting density-based algorithms in unraveling the underlying clustering structure.

[1] A data vector $\mathbf{x}_i$ is considered unassigned if $u_{ij} = 0$, for $j = 1, ..., m$.



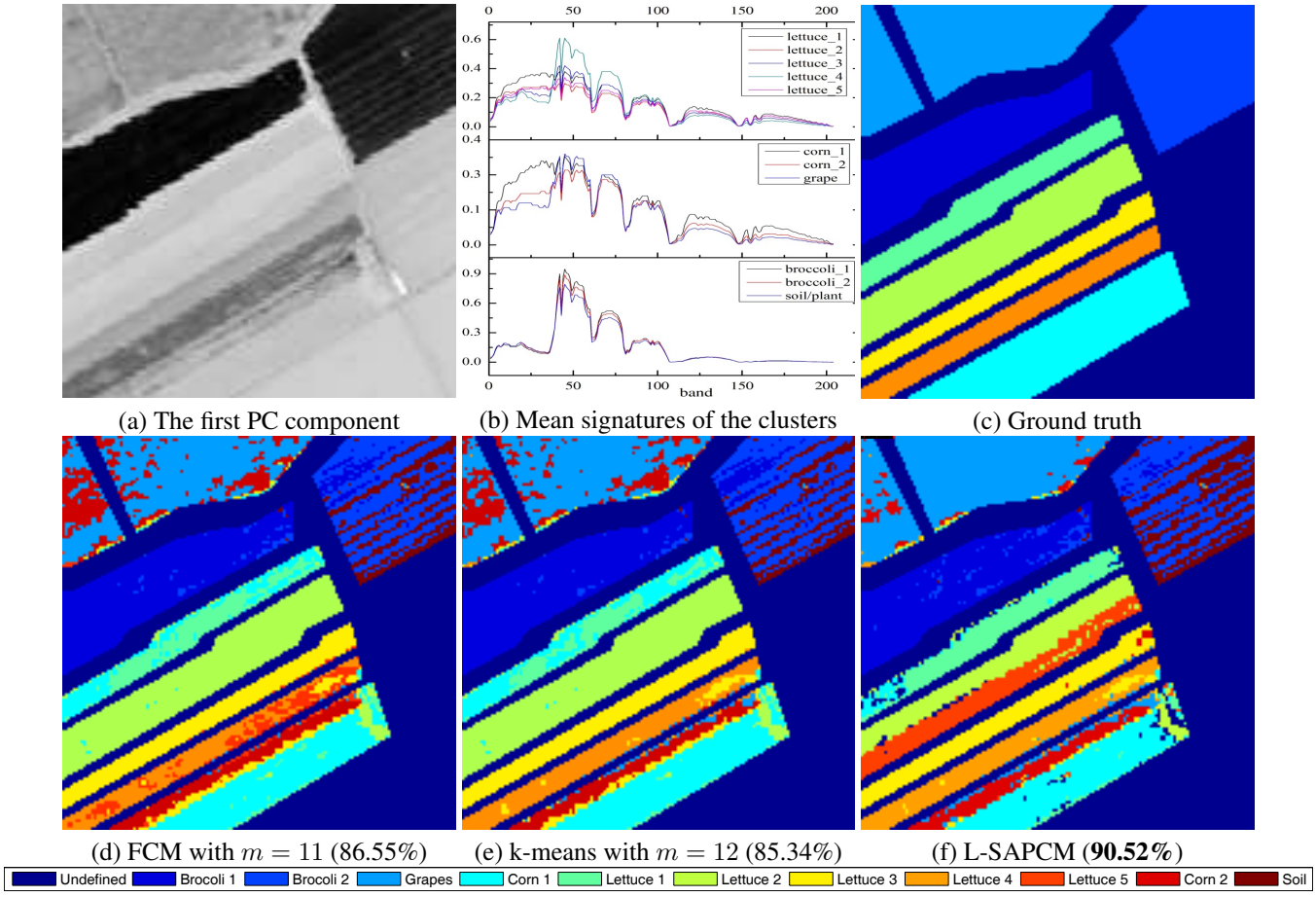
(a) (b)

Fig. 1. (a) Initial data, (b) Data after removing "noisy" points

Algorithm 3 $[clusters] = L - SAPCM(X)$

$X = PCA(X)$ and keep the l first components of PCA
$pending\_sets = \{X\}$
$clusters = \{\}$
While $pending\_sets \neq \emptyset$ **do**
Take an element C of $pending\_sets$
$[C] = data\_purifying(C)$
$\{C_1, ..., C_m\} = SAPCM(C, \lambda, m)$, where m is the final number of clusters that $SAPCM$ returns
If $m > 1$
$pending\_sets = (pending\_sets - \{C\}) \cup \{C_1, ..., C_m\}$
else if $m = 1$
$pending\_sets = pending\_sets - \{C\}$
$clusters = clusters \cup \{C\}$
End if
End
Assign each $\mathbf{x}_i \in X$ that has been removed from the $data\_purifying$ scheme to its closest among $clusters$

We describe now the proposed Layered SAPCM (L-SAPCM) algorithm suitable for HSI clustering (**Alg. 3**). The algorithm first performs PCA on the data set and then executes the SAPCM algorithm in a layered form. Before each execution of SAPCM, $data\_purifying$ (**Alg. 2**) is performed, as described above. Initially, SAPCM is applied on the whole data set producing some subsets (clusters) that constitute the first layer clustering. Then, for each subset of the first layer the SAPCM is recursively applied. The procedure continues for each one of the resulting data subsets and terminates when SAPCM returns a single subset (which means that the currently processed data subset does not possess a clustering structure). Such a data subset is considered as a



(a) The first PC component (b) Mean signatures of the clusters (c) Ground truth

(d) FCM with $m = 11$ (86.55%) (e) k-means with $m = 12$ (85.34%) (f) L-SAPCM (**90.52%**)

Fig. 2. Clustering results for **Salinas** image with the corresponding classification accuracy

cluster[2]. The whole procedure is given in **Alg. 3**.

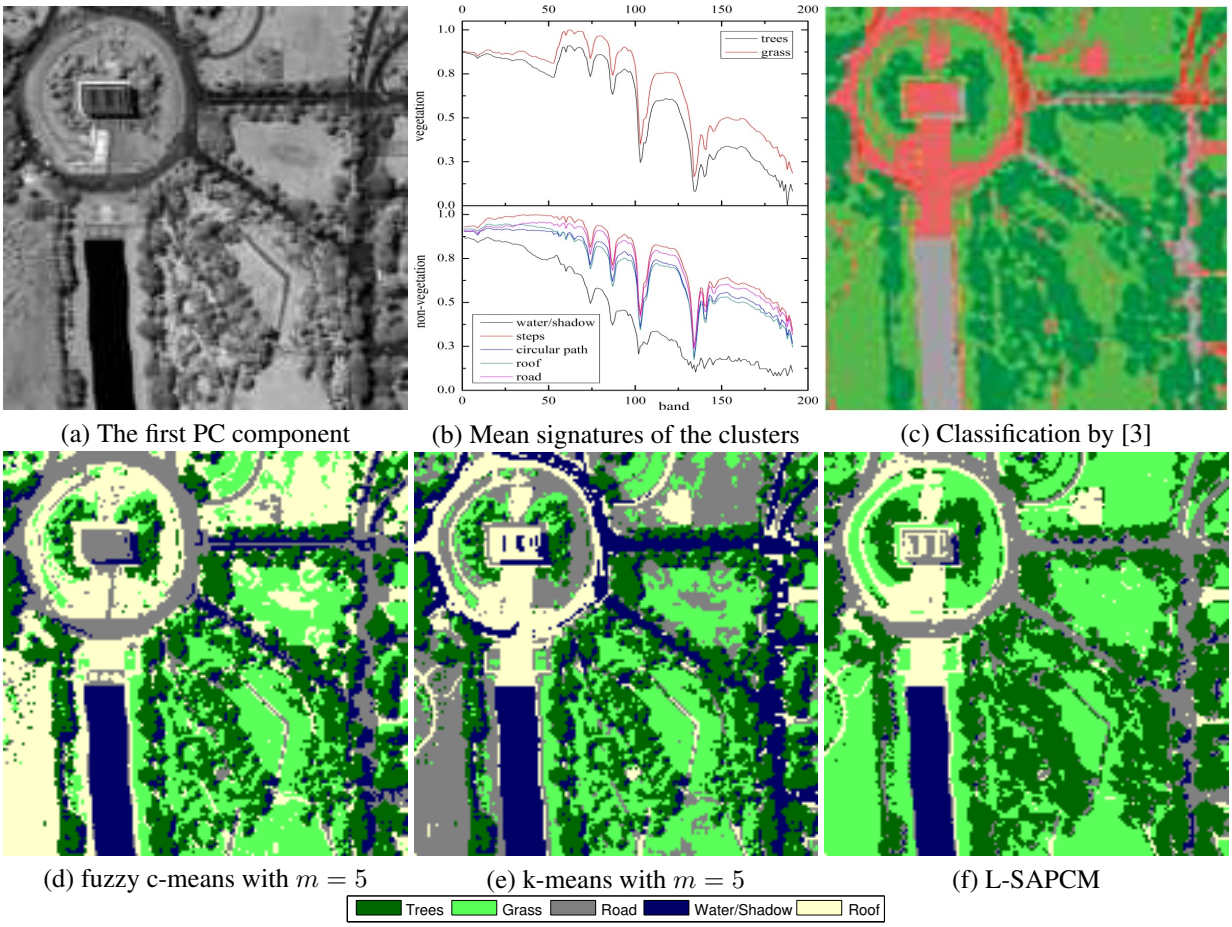
4. RESULTS AND CONCLUSIONS

Tests are performed to two specific datasets of 150x150 pixels from a rural and an urban area. In the sequel, the results of k-means, FCM and the proposed L-SAPCM are discussed.

1. *Salinas, California:* The dataset used is a subscene of the flightline acquired by the AVIRIS sensor over Salinas Valley, California (Fig. 2(a)). Salinas groundtruth contains eight vegetation classes: "corn", two types of "broccoli", four types of "lettuce" and "grapes" (Fig. 2(c)). The results of the three clustering methods are shown in Fig. 2(d-f). L-SAPCM distinguishes eleven vegetation clusters after a 3-layered data processing with suitable values for λ at each layer. These correspond to "grapes", two subclasses of "corn" ("corn 1" and "corn 2"), five subclasses of "lettuce" ("lettuce" 1,2,3,4, and 5), two "broccoli" classes ("broccoli 1" and "broccoli 2") and a "soil/plant" class, (Fig. 2(f)). "Corn" is distinguished into two subclasses, mainly due to their clear spectral differentiation within the first forty bands of the dataset. However, from band 40 and onward, the two subclasses present similar spectra. The "grapes" spectral signature shows similar characteristics with the previous "corn" classes, however the "red edge" pattern in the near infrared ($680 - 750nm$), is different (lower overall reflectance in the visible and higher overall reflectance in the NIR) as shown in Fig. 2(b) (bands 30-40). All algorithms distinguish an additional class named "soil/plant" appearing in linear stripes within the "broccoli 2" area. This class seems to correspond to mixed vegetation/soil pixels with higher spectral participation of soil in the overall spectral signature of the pixel. L-SAPCM distinguishes two subclusters for the class "lettuce 2", namely "lettuce 2" and "lettuce 5". Their spectral pattern is different between bands 25-40, but, they present the same diagnostic absorption features in position and depth at longer wavelengths, indicating the same species over the same soil. Finally, compared to k-means and FCM, L-SAPCM differentiates much better "corn 1" from "lettuce 1", detects an additional "lettuce" subcluster and shows less "corn 2" classified pixels in the "grapes" area.

2. *Washington DC Mall:* The dataset used is a section of the HSI flightline acquired by HYDICE over Washington

[2]L-SAPCM can also be viewed as an *in-depth* processing algorithm.



(a) The first PC component (b) Mean signatures of the clusters (c) Classification by [3]

(d) fuzzy c-means with $m = 5$ (e) k-means with $m = 5$ (f) L-SAPCM

Fig. 3. Clustering results for **Washington DC Mall** image

DC Mall that includes the Lincoln Memorial (Fig. 3(a)). The results of four analysis methods, namely the classification method in [3], k-means, FCM and L-SAPCM are shown in Fig. 3(c-f). L-SAPCM distinguishes five clusters: "water/shade", "grass", "tree", "shaded roof/road" and "bright surface" (Fig. 3(f)) after a "single-layer" processing. This slightly differs from the results in [3], where five main classes are considered: "water/shade", "grass", "roof", "tree" and "road". L-SAPCM distinguishes two new categories, "shaded roof/road" and "bright surface". Although, they show spectral similarities (Fig. 3(b)), they are differentiated by the high reflectance values of the latter. A visual interpretation of the initial image reveals that the "bright surface" cluster includes parts of the Memorial roof (the other part is categorized as "shaded roof/road"), marble surfaces, narrow paved paths and bare soil. The "water/shade" class is well distinguished due to its characteristic low reflectance spectral signature with characteristic water absorptions (Fig. 3(b)). Vegetation, "grass" and "tree", clusters are also very well distinguished due to their characteristic "red edge" patterns. The different "tree" and "grass" red edge positions distinguish the two categories ("tree" is slightly shifted to the left). Compared to k-means and FCM, L-SAPCM delineates with high precision the "grass" and "tree" clusters (i.e., left part of the image, vegetated area at the right of the Memorial, upper right part of the image), roads from shades (especially when compared to k-means) and the Memorial's roof consecutive inclined planes (a detail not mapped by FCM).

Concluding, L-SAPCM outperforms both k-means and FCM in the accurate detection of objects in HSIs. Improvements will include structure detection and spatial information.

5. REFERENCES

[1] S. D. Xenaki, K. D. Koutroumbas, and A. A. Rontogiannis, "Sparse adaptive possibilistic clustering," *IEEE ICASSP*, 2014.

[2] C. Rodarmel and J. Shan, "Principal component analysis for hyperspectral image classification," *Surveying and Land Information Systems*, vol. 62, pp. 115–000, 2002.

[3] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. & Rem. Sens.*, vol. 43, pp. 480–491, 2005.