Experimental Study of Hierarchical Clustering for Unmixing of Hyperspectral Images

José Prades
Institute of Telecommunications and Multimedia Applications
Universitat Politècnica de València
Valencia, Spain
jprades@dcom.upv.es

Addisson Salazar
Institute of Telecommunications and Multimedia Applications
Universitat Politècnica de València
Valencia, Spain
asalazar@dcom.upv.es

Gonzalo Safont
Institute of Telecommunications and Multimedia Applications
Universitat Politècnica de València
Valencia, Spain
gonsaar@dcom.upv.es

Luis Vergara
Institute of Telecommunications and Multimedia Applications
Universitat Politècnica de València
Valencia, Spain
lvergara@dcom.upv.es

***Abstract*— Estimation of the number of materials that are present in a hyperspectral image is a necessary step in many hyperspectral image processing algorithms, including classification and unmixing. Previously, we presented an algorithm that estimated the number of materials in the image using clustering principles. This algorithm is an iterative approach with two input parameters: the initial number of materials (*P_0*) and the number of materials added in each iteration (Δ). Since the choice of *P_0* and Δ can have a large impact on the estimation accuracy. In this paper, we made an experimental study of the effect of these parameters on the algorithm performance. Thus, we show that the choice of a large Δ can significantly reduce the estimation accuracy. These results can help to make an appropriate choice of these two parameters.**

Keywords*— *Hyperspectral images, endmembers, clustering, Independent Component Analysis, Principal Component Analysis.

I. Introduction

Some steps in hyperspectral imaging processing require to know how many materials are in an image. One of these steps is unmixing, which consists of the split of the components composing the pixel data, which forms a pixel signature obtained from hyperspectral images. Thus, from the output of unmixing, abundance maps describing the amount of every component in a given pixel can be built.

The calculation of the endmember number in hyperspectral images have been approached from different methods [2]-[13]. Recently, we proposed a method that implements a hierarchical clustering for this purpose [14]. From an initial number of clusters, a pyramidal structure is built, where the number of clusters decreases at each level of the pyramid until it reaches one cluster at the highest level. Thus, a set of partitions of the hyperspectral image is generated where each partition represents a number of materials. The optimal partition is determined applying a cluster validation index. A drawback of the algorithm is that the user must specify the maximum material number in the image (P), which is the number of clusters in the pyramid base. In [20], we extended this algorithm so that so that the input parameter P is not required. Thus, the estimation for several increasing values of P is performed iteratively until a condition is reached. This iterative algorithm has two input parameters: the number of materials considered in the first iteration (P_0) and the number of materials added in each iteration (Δ).

The choice of P_0 and Δ can have a large impact on the estimation accuracy. In this paper, we experimentally study the influence of these two parameters in the algorithm performance. The results of this study can help to make an appropriate choice of these two parameters.

In Section II, a review of previous works on estimation of the hyperspectral image number of materials is included. In Section III, we experimentally study how the choice of P in the algorithm of [14] affects the estimation performance. In Section IV an estimation algorithm is proposed. In Section V, the proposed algorithm using five hyperspectral images is experimentally evaluated. Finally, Section VI includes the conclusions and future lines of research.

II. Hierarchical Clustering

A hyperspectral image of N pixels and L spectral is defined as a $L \times N$ matrix, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$. Each column of $\mathbf{X}$ is an image pixel. Since the l-th component of a pixel is the measured energy at the l-th spectral band ($1 \leq l \leq L$), each pixel can be considered as a L-band spectrum. Each material of a hyperspectral image correspond to a characteristic L-band spectrum, called spectral signature or endmember. Commonly, the material number in the image, K, is much smaller than the band number, i.e., $K \ll L$. The material number in an image can be determined by estimating the number of endmembers that contains a hyperspectral image $\mathbf{X}$. Many algorithms that estimate K from $\mathbf{X}$ are founded on the *linear mixing model* (LMM), where each pixel is modeled as a random vector with the following expression [2],

$$\mathbf{x} = \sum_{i=1}^{K} c_i \, \mathbf{e}_i + \mathbf{n} \qquad (1)$$

where $\{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$ is the endmember set. The coefficients $\{c_1, \cdots, c_K\}$ are random variables that represent the fraction of each endmember (abundances) in $\mathbf{x}$; and the noise term, $\mathbf{n}$, is a random vector that accounts for any model or measurement error. Abundance variables satisfy $c_i \geq 0 \ (i = 1, \cdots, K)$ and$c_1 + \cdots + c_K = 1$ are named abundance constraints.

Several algorithms for endmember estimation are based on eigenvalue computation from the image sample correlation matrix [3]-[10]. LMM considers that a single endmember represents each material of the image. However, in practice, the material endmember could vary spatially. The normal compositional model accounts for endmember variability by using random endmembers [15]. In [11], an algorithm based on

This research has been supported by Generalitat Valenciana, grant PROMETEO 2019/109.

the normal composition model was proposed. Other algorithms from different perspectives to estimate the endmember number have been proposed in [12], [13].

We presented an algorithm based on clustering approach in [14]. In this work, we consider that there is a principal endmember in a specified amount of pixels of a hyperspectral image. From this assumption, cluster centered on those pixels could be defined. Under this assumption, we can estimate K by determining the cluster number in the image. This method performs a single hierarchical clustering that generates hierarchical image partitions, selecting the optimal partition by means of a clustering quality measure. The estimated endmember number, $\widehat{K}$, is set to the cluster number contained in the chosen partition. Figure 1 shows an outline of the algorithm.

In Figure 1, the preprocessing of the image converts the columns of $\mathbf{X}$ into feature vectors. This preprocessing involves the following: centering, which provides zero-mean variables; dimensionality reduction (using principal component analysis, PCA, to reduce the number of variables from L to M); and normalization (to provide feature vectors with unit-variance variables). The preprocessing provides a matrix $\mathbf{Y} \in \mathbb{R}^{M \times N}$ whose columns are normalized feature vectors of the image. The clustering step is performed in two stages: K-means and hierarchical partition. The K-means stage split into P clusters the columns of $\mathbf{Y}$ and provides partition $\mathcal{C}$. Then, the hierarchical clustering stage takes $\mathcal{C}$ as input and provides a hierarchy of groups or partitions $\{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots, \mathcal{C}^{(P)}\}$, where $\mathcal{C}^{(1)} = \mathcal{C}$ and $\mathcal{C}^{(P)}$ is an individual partition containing all of the vectors of features. This process is based on ICA to model the cluster densities and the Kullback-Leibler distance (KLD) to measure the distance between clusters [16]-[20].

Firstly, the parameters of the ICA corresponding to each cluster in $\mathcal{C}$ are obtained. Then, the KLD between each pair of clusters in $\mathcal{C}$ is computed using their corresponding ICA parameters. Finally, the clusters of $\mathcal{C}$ are iteratively merged until there is only one cluster. In each merging iteration, the two closest clusters are joined (see [14] for details). We assume that one particular cluster $\{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots, \mathcal{C}^{(P)}\}$ will be close to the real constituent materials of $\mathbf{X}$. Thus, the cluster number in that partition will approach the endmember number (K).

One of the hierarchy partitions of the hierarchy is selected in the cluster validation step, so $\widehat{K}$ is set to the cluster number. At each iteration, r of the merging process, the centroid of each of the two merged clusters of feature vectors at $\mathcal{C}^{(r)}$, i.e., $\mathbf{m}_i$ and $\mathbf{m}_j$, is obtained; and v_{P-r+1} is set to

$$v_{P-r+1} = \sum_{l=0}^{M-1} \left(m_{i,l} - m_{j,l}\right)^2 \qquad (2)$$

where the l-th component of $\mathbf{m}_i$ is denoted as $m_{i,l}$. From the obtained sequence v_k ($2 \leq k \leq P$), $\widehat{K}$ is set to the index k where v_k reaches its maximum value. In addition, the proposed method provides the image segmented into $\widehat{K}$ clusters and obtain an estimation of the material endmembers, i.e., the centroid of the clusters.

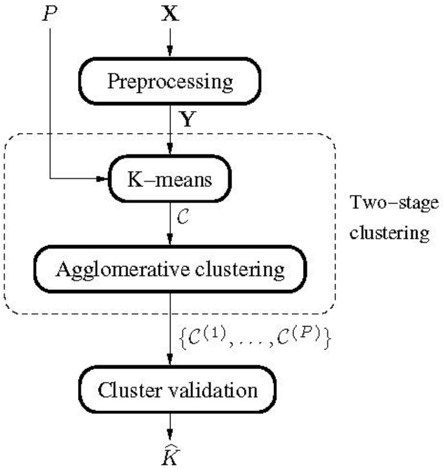

Fig. 1. Popeline of the proposed estimation method.

Figure 2 shows the hierarchy of partitions generated by the algorithm when the input is the hyperspectral image Jasper Ridge (the details of this image are provided in Section V) and P=10. This figure shows: the average of the image bands displayed in grayscale (a) and the partition of the image from 10 clusters (Fig. 2b), to 2 clusters (Fig. 2j). For this image, $\widehat{K} = 4$ is obtained, which, according to [21], is the true number of materials. The resulting image segmentation is shown in Figure 2(h). The clusters displayed in this image correspond to the following materials: Tree -green-, Soil -red-, Road -black-, and Water -blue-, and. According to [21], this segmentation is approximately correct. This allows to obtain an approximation of the endmember for each material by computing the centroid of the clusters [14].

III. Influence of P in the estimation accuracy

In the algorithm described in Section II, the user must set the value of input parameter P. In this section, the impact of the choice of P in the estimation accuracy is studied. To this end, we have obtained $\widehat{K}$ for some of P values and for the hyperspectral images Urban, WDM, and Cuprite (the features of these images are provided in Section V). Specifically, for each image and P, we run the algorithm ten times and computed the median of the ten $\widehat{K}$ values obtained. The results are shown in Figures 3, 4, and 5 for Urban, WDM, and Cuprite, respectively. Since the complexity of the estimation algorithm increases with P, we have focused the analysis in a set of small values of P (in Cuprite, the range of P values have been increased since this image has more materials than Urban and WDM).

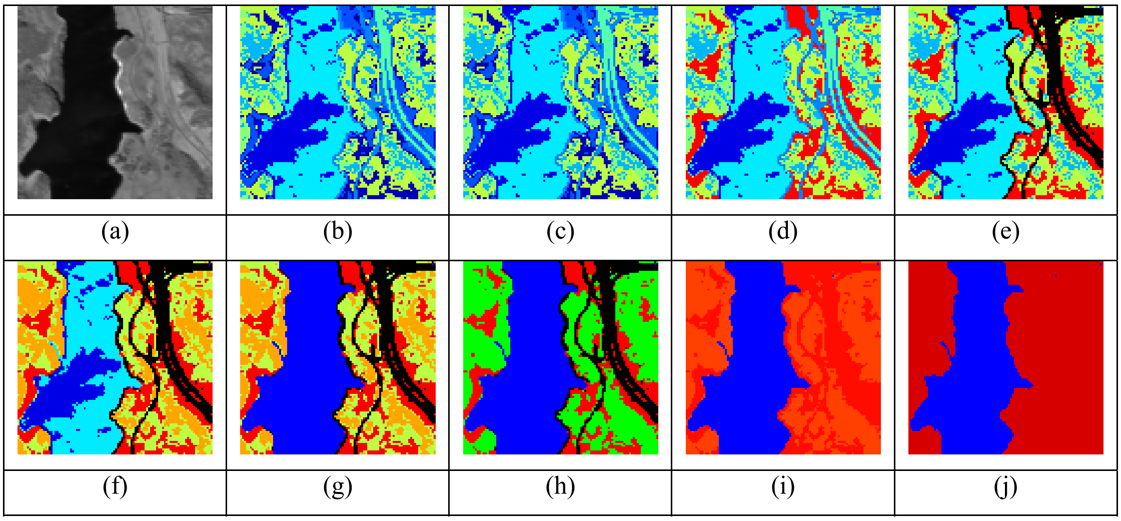

Fig. 2. Partitions of an image provided by the proposed algorithm for image Jasper Ridge when P=10. The figure shows the average of the image bands in grayscale (a), and the segmentation with 10, 9, 8, 7, 6, 5, 4, 3, and 2 clusters (in images (b), (c), (d), (e), (f), (g), (h), (i), and (j), respectively).

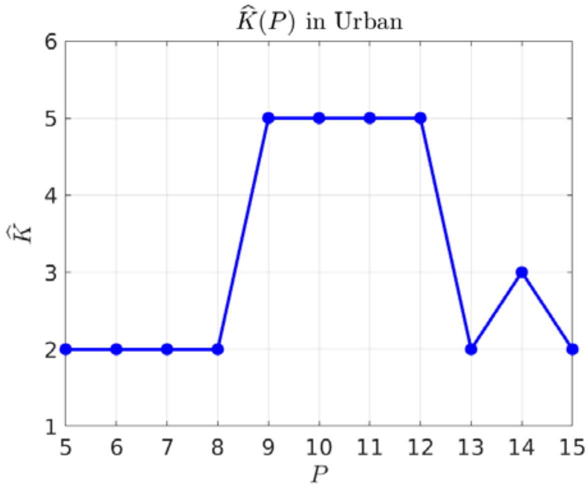

Fig. 3. Sequence $\widehat{K}(P)$ for the image Urban.

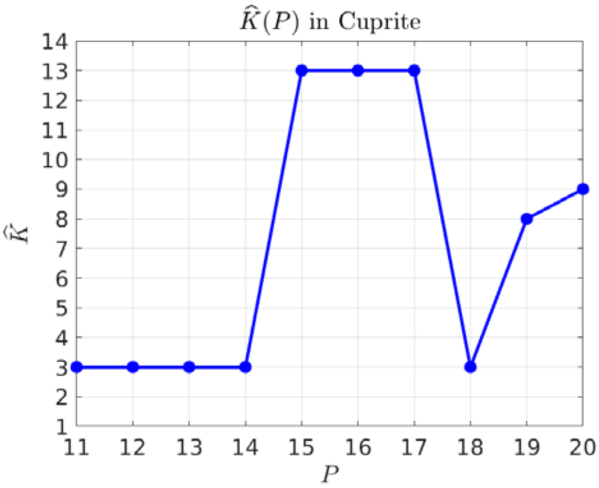

Fig. 5. Sequence $\widehat{K}(P)$ for the image Cuprite.

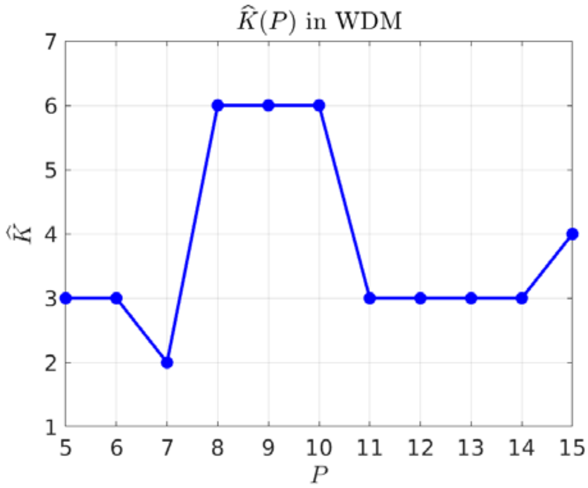

Fig. 4. Sequence $\widehat{K}(P)$ for the image WDM.

Note that Figures 3, 4, and 5 show similar patterns. Thus, $\widehat{K}$ remains constant in a plateau of values of P in which $\widehat{K}$ is close or equal to K ($\{9, 10, 11, 12\}$ in Urban, $\{8, 9, 10\}$ in WDM, and $\{15, 16, 17\}$ in Cuprite). Before each plateau, $\widehat{K}$ is smaller than K and is constant with P or has small fluctuations (in WDM). After each plateau, $\widehat{K}$ is also smaller than K and fluctuates (with a large variation in Cuprite). Also note that the plateau starts after the true value of K.

For each image and P value, the algorithm generally provides stable results. In fact, the ten values of $\widehat{K}$ are the same in many cases. We have only observed some remarkable fluctuation when P is large (e.g., for the image WDM and $P = 15$, and for the image Cuprite and $P = 17$ or $P = 20$).

Since we have shown that each sequence $\widehat{K}(P)$ has a plateau that contains good estimates of K, the estimator could identify a P value that belongs to the plateau and output the $\widehat{K}$ that this P value provides. This identification can be based on the monotonicity properties of the sequences $\widehat{K}(P)$. Thus, since before reaching the plateau, the sequences $\widehat{K}(P)$ are generally non-decreasing, we can iteratively estimate $\widehat{K}$ for several increasing values of $\widehat{K}$. Since the sequence values immediately after the plateau are smaller than those of the plateau, we stop after obtaining a $\widehat{K}$ smaller than the one in the previous iteration. Note that we use the estimate of the iteration previous to the last one (since the $\widehat{K}(P)$ provided by the last iteration does not belong to the plateau).

IV. Proposed algorithm

In this section, an algorithm that automatically determines a proper value of P for the estimation of K is proposed. This algorithm is a modification of the algorithm in [22] and it is based on the analysis of Section III. According to the results of this section, we can determine $\widehat{K}$ by iteratively estimating K for several increasing values of P. The iteration should be stopped when the estimated K decreases with respect to the value of the previous iteration, which provides the estimated valued of K. Based on this, we propose Algorithm 1 ('EstimateK').

Apart from the input image **X**, the Algorithm 1 has three input arguments: P_0 (the initial value of P), Δ (the increment of P in each iteration), and P_{max} (the maximum value of P considered). Arguments P_0 and Δ allow to control the set of P values that are considered by the algorithm. Argument P_{max} allow to stop when the sequence of $\widehat{K}$ values does not decrease.

Algorithm 1: Estimation of K.

```
Inputs: $\mathbf{X}$, $P_0$, $\Delta$, $P_{max}$
Output: $\widehat{K}$
$P = P_0$
$\widehat{K} = \text{EstimateK}(\mathbf{X}, P)$
Repeat
   $J = \widehat{K}$
   $P = P + \Delta$
   $\widehat{K} = \text{EstimateK}(\mathbf{X}, P)$
until $\widehat{K} < J$ or $P \geq P_{max}$
$\widehat{K} = J$
```

Note that large values of Δ speeds up the algorithm but could provide worse estimations since the algorithm could jump the plateau with accurate values of $\widehat{K}$ (see the figures of Section III). Also note that the value of P_0 must be small since otherwise the algorithm would start with a P value after the plateau.

V. Experimental results

We tested the accuracy of Algorithm 1 in five hyperspectral images. The images are: Samson, Jasper Ridge (JR), Urban, Washington DC Mall (WDM), and Cuprite. Table I shows the following for each of the images: spatial dimensions (#rows x #columns), number of spectral bands L_i, spectral band number used in estimation (L), and groundtruth endmember number K according to [21]. The table includes several groundtruth values of K for the Urban and Cuprite images since classifications with different number of materials have been reported.

TABLE I. Test images used in the experiments.

Image	Size	L_{i}	L	K
Samson	95 × 95	156	156	3
Jasper Ridge	100 ×100	224	198	4
Urban	307 ×307	221	162	4,5,6
Washington DC Mall	150 ×150	224	191	6
Cuprite	250 ×190	224	188	10,12,14

Table II shows the values of $\widehat{K}$ provided by the proposed algorithm for several values of P_0 (6, 8, and 10) and Δ (1, 2, 3, 4, and 5). In all the runs, P_{max} is set to 20 since the stability of the estimates decreases for $P > 20$. For each image, P_0, and Δ, the median of the K values obtained after running Algorithm 1 twelve times is shown.

TABLE II. Estimated number of endmembers (JR=Jasper Ridge; WDM=Washington DC Mall)

P_0	Δ	Samson	JR	Urban	WDM	Cuprite
6	1	3	6	5	3	13
8	1	3	6	5	6	13
10	1	3	4	5	5	13
6	2	3	6	5	6	13
8	2	3	6	5	6	13
10	2	3	6	5	5	13
6	3	3	6	5	6	13
8	3	3	6	5	6	11
10	3	3	6	5	5	13
6	4	3	6	5	5	3
8	4	3	6	5	6	13
10	4	3	4	5	5	4
6	5	3	6	5	8	3
8	5	3	6	2	6	3
10	5	3	4	5	5	13

For Samson, Urban and WDM, the algorithm provided estimates that are equal or close to the groundtruth K in most cases. In Cuprite, good estimates are obtained when $\Delta < 4$. The number of materials was generally overestimated in Jasper Ridge since in this image the assumed monotonicity properties are not met as accurately as in the rest of images. Using a $\Delta > 5$ would speed up the process at the expense of a significant decrease of the estimation accuracy (as can be deduced from the figures of Section III).

VI. Conclusion

A method to estimate the number of materials in a hyperspectral image was developed. This method extends the algorithm in [22] and does not require the user to provide a proper value of parameter P. The estimation for several increasing values of P is performed iteratively until a condition is reached. The results showed the proposed method provided estimates that were equal or close to the groundtruth values in many cases.. In future works, advanced machine learning techniques including semi-supervised learning [23][24], oversampling [25]-[27], and graph signal processing [28][29] will be studied to improve endmember performance estimation.

References

[1] L. Wang and C. Zhao, Hyperspectral image processing. Springer: Berlin/Heildelberg, Germany, 2016.

[2] N. Keshava and J. F. Mustard, "Spectral unmixing," IEEE Signal Process. Mag., vol. 19, no. 1, pp. 47–57, 2002.

[3] J. Harsanyi, W. Farrand, and C.-I. Chang, "Determining the number and identity of spectral endmembers: An integrated approach using neyman-pearson eigenthresholding and iterative constrained rms error minimization," in Proc. 9th Them. Conf. Geol. Remote Sensing, 1993.

[4] C.-I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," IEEE Trans. Geosci. Remote Sens., vol. 42, no. 3, pp. 608–619, 2004.

[5] B. Luo, J. Chanussot, S. Dout´e, and L. Zhang, "Empirical automatic estimation of the number of endmembers in hyperspectral images," IEEE Trans. Geosci. Remote Sens., vol. 10, no. 1, pp. 24–28, 2013.

[6] K. Cawse-Nicholson, S. B. Damelin, A. Robin, and M. Sears, "Determining the intrinsic dimension of a hyperspectral image using random matrix theory," IEEE Trans. Image Process., vol. 22, no. 4, pp. 1301– 310, 2013.

[7] A. Halimi, P. Honeine, M. Kharouf, C. Richard, and J.-Y. Tourneret, "Estimating the intrinsic dimension of hyperspectral images using noise-whitened eigengap approach," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 7, pp. 3811–3820, 2016.

[8] M. Berman, "Improved estimation of the intrinsic dimension of a hyperspectral image using random matrix theory," Remote Sensing, vol. 11, no. 11, p. 1049, 2019.

[9] C. Andreou and V. Karathanassi, "Estimation of the number of endmembers using robust outlier detection method," IEEE J. Sel. Topics Appl. Earth Obs. Rem. Sens., vol. 7, no. 1, pp. 247–256, 2014.

[10] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," IEEE Trans. Geosci. Remote Sens., vol. 46, no. 8, pp. 2435–2445, 2008.

[11] O. Eches, N. Dobigeon, and J.-Y. Tourneret, "Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical bayesian algorithm," IEEE J. Sel. Top. Signal Process., vol. 4, no. 3, pp. 582–591, 2010.

[12] R. Heylen and P. Scheunders, "Hyperspectral intrinsic dimensionality estimation with nearest-neighbor distance ratios," IEEE J. Sel. Topics Appl. Earth Observations Rem. Sens., vol. 6, no. 2, pp. 570–579, 2013.

[13] R. Marrero, S. Lopez, G. M. Callic´o, M. A. Veganzones, A. Plaza, J. Chanussot, and R. Sarmiento, "A novel negative abundance-oriented hyperspectral unmixing algorithm," IEEE J. Sel. Topics Appl. Earth Observations Remote Sens., vol. 53, no. 7, pp. 3772–3790, 2015.

[14] J. Prades, G. Safont, A. Salazar, and L. Vergara, "Estimation of the number of endmembers in hyperspectral images using agglomerative clustering," Remote Sensing, vol. 12, no. 21, p. 3585, Nov. 2020.

[15] M. T. Eismann and D. Stein, "Stochastic mixture modeling," in Hyperspectral data exploitation: theory and application, C.-I. Chang, Ed. New York: Wiley, 2007, ch. 5, pp. 582–591.

[16] A. Salazar, L. Vergara, I. Igual, and J. Gosalbez, "Blind source separation for classification and detection of flaws in impact-echo testing," Mechanical Systems and Signal Processing, vol. 19, no. 6, pp. 1312-1325, 2005.

[17] A. Salazar, L. Vergara, R. Llinares, "Learning material defect patterns by separating mixtures of independent component analyzers from NDT sonic signals," Mechanical Systemsd and Signal Processing, vol. 24, no. 6, pp. 1870-1886, 2010.

[18] A Salazar, J Igual, G Safont, L Vergara, A Vidal, "Image applications of agglomerative clustering using mixtures of non-Gaussian distributions," in International Conference on Computational Science and Computational Intelligence, (CSCI), pp. 459-463, 2015.

[19] G.Safont, A. Salazar, L. Vergara, et al., "Nonlinear estimators from ICA mixture models," Signal Processing, vol. 155, pp. 281-286, 2019.

[20] G. Safont, A. Salazar, L. Vergara, E. Gomez, and V. Villanueva, "Multichannel dynamic modeling of non-Gaussian mixtures," Pattern Recognition, vol. 93, pp. 312-323, 2019.

[21] F. Zhu, "Hyperspectral unmixing: Ground truth labeling, datasets, benchmark performances and survey," arXiv, arXiv:1708.05125, 2017.

[22] J. Prades, A. Salazar, G. Safont, and L. Vergara, "Determining the number of endmembers of hyperspectral images using clustering," in 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1664-1668, 2020.

[23] A. Salazar, G. Safont, and L. Vergara, "Semi-supervised learning for imbalanced classification of credit card transaction," in 2018 International Joint Conference on Neural Networks, IJCNN 2018, art. no. 8489755, pp. 4976-4982, 2018.

[24] A. Salazar, G. Safont, L. Vergara, E. Vidal, "Pattern recognition techniques for provenance classification of archaeological ceramics using ultrasounds," Pattern Recog. Lettrs, vol. 135, pp. 441-450, 2020.

[25] A. Salazar, G. Safont, and L. Vergara, "Surrogate techniques for testing fraud detection algorithms in credit card operations," in Int. Carnahan Conf Sec. Tech, ICCST 2014, no. 6986987, pp. 124-129, 2014.

[26] J. Belda, L. Vergara, G. Safont, A. Salazar, Z. Parcheta, "A new surrogating algorithm by the complex graph Fourier transform (CGFT)," Entropy, 21, no. 8, art. no. 759, 2019.

[27] A. Salazar, L. Vergara, G. Safont, "Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets," Expert Systems with Applications 163, art. no. 113819, pp. 1-12, 2021.

[28] J. Belda, L. Vergara, A. Salazar, G. Safont, "Estimating the Laplacian matrix of Gaussian mixtures for signal processing on graphs," Signal Processing, vol. 148, pp. 241-249, 2018.

[29] J. Belda, L. Vergara, G. Safont, and A. Salazar, "Computing the partial correlation of ICA models for non-Gaussian graph signal processing," Entropy, vol. 21, no. 1, art. no. 22, 2019.