

Unsupervised Learning and Feature Extraction in Hyperspectral Imagery

Eduardo K. Nakao(✉) and Alexandre L. M. Levada

Universidade Federal de São Carlos,
Rod. Washington Luís km 235, 310 São Carlos, Brazil
eduardokazuonakao@gmail.com

Abstract. Remotely sensed hyperspectral scenes are typically defined by large area coverage and hundreds of spectral bands. Those characteristics imply smooth transitions in the spectral-spatio domains. As consequence, subtle differences in the scene are evidenced, benefiting precision applications, but values in neighboring locations and wavelengths are highly correlated. Nondiagonal covariance matrices and wide autocorrelation functions can be observed this way, implying increased intraclass and decreased interclass variation, in both spectral and spatial domains. This leads to lower interpretation accuracies and makes it reasonable to investigate if hyperspectral imagery suffer from Curse of Dimensionality. Moreover, as this Curse can compromise linear method's Euclidean behavior assumption, it is relevant to compare linear and nonlinear dimensionality reduction performance. So, in this work we verify these two aspects empirically using multiple nonparametric statistical comparisons of Gaussian Mixture Model clustering performances in the cases of: absence, linear and nonlinear unsupervised feature extraction. Experimental results indicate Curse of Dimensionality presence and nonlinear adequacy.

Keywords: Hyperspectral imagery · Curse of dimensionality · Unsupervised feature extraction · Nonlinear dimensionality reduction · Unsupervised learning

1 Introduction

Hyperspectral imaging [12] captures the energy matter interaction in hundreds of channels that can precisely define the chemical composition of different materials by spectral similarity discrimination. A hyperspectral image (HI) is viewed as a cube of two-dimensional single band images stacked in the third dimension of sampled wavelengths. So if m spectral bands were collected, each pixel can be viewed as a m dimensional vector where its jth coordinate value is the reflectance

This study was financed in part by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 792–806, 2020.
https://doi.org/10.1007/978-3-030-58799-4_57

measured at the jth wavelength. This vector is the geometric pixel representation in the spectral domain.

Due to this presence of hundreds of channels, HIs are spectrally smooth and spatially piecewise smooth. Values in neighboring locations and wavelengths are highly correlated. This can be observed by extremely nondiagonal covariance matrices and wide autocorrelation functions [4]. So there are increased intraclass variation and decreased interclass variation, in both spectral and spatial domains [34], which leads to lower interpretation accuracies. Thus, the linear independence assumption in the basis vectors of an Euclidean space can be not suitable. If pixel vectors are seen as random variables, the high correlation between them suggests that a possible statistical independence assumption is not suitable as well.

Those facts raise the supposition that the pixel vector representation is embedded in a high-dimensional space, suffering from the Curse of Dimensionality. From Machine Learning perspective, the problem is described by the Hughes phenomena [14] which states that, when dimension grows from a certain point in relation to the number of samples, classification performance is degraded.

From a geometric point of view, in high dimensions it can also be proved [20] that the volume of a hypercube is concentrated on the corners [17,31] and the volume of a hypersphere or hyperellipsoid concentrates on an outside shell [17]. These facts imply the following consequences: space is mostly empty (so data can be projected to a lower dimension without losing significant class separability); local neighborhoods are almost empty (requiring large estimation bandwidth and losing detailed density estimation); normal distributions have their data concentration on the tails instead the means, and uniformly distributed data is more likely collected on the corners (making density estimation more difficult).

Other important high dimensionality characteristics are [20]: diagonal vectors become orthogonal to the basis vectors (the projection of a cluster in one of these "wrong directions" can ruin the information within the data [31]); data covariance matrix becomes ill-conditioned; Euclidean distance is not fully reliable [1]; low linear projections have the tendency to be Normal.

So, in Curse of Dimensionality presence, the clustering method Gaussian Mixture Model (GMM) can have its behavior compromised because it uses Euclidean distance in neighborhood search and have Gaussian hypothesis and statistical independence assumption. On the other hand, in the presence of few bands (i.e. low dimensionality), GMM is efficient to determine class distribution even with few samples in some classes (which is common in HIs due to their subtle substances discrimination derived from high spectral resolution). The model is efficient even if some classes are not normally distributed in fact [19] and it also takes advantage of the high dimensional space property of Normal distributions formation in the projection from high to low dimensionality.

Another problem is that, as GMM, linear Dimensionality Reduction (DR) methods also assume Euclidean data behavior. So those methods can be compromised too. Given a image, it is not possible to know beforehand theoretically if its pixels vectors will be better represented by a linear or a nonlinear

DR approach. Therefore, empirical computational experiments are necessary to investigate linear against nonlinear data representation adequacy for HIs.

Considering all those inherent properties and possible problems, the focus in this study is to analyze how unsupervised feature extraction methods affect clustering by GMM and if nonlinear DR is more adequate than linear for HIs. The common lack of class information in hyperspectral imagery justify the interest in unsupervised feature extraction and its impact in unsupervised learning. We compare Principal Component Analysis (PCA) - the most traditional linear DR method - against Isometric Feature Mapping (ISOMAP) and Locally Linear Embedding (LLE) - both pioneering nonlinear methods with global and local approaches respectively. All three methods yield an orthonormal basis in an Euclidean space where the vectors are written as linear combination of the others. Besides that, as vectors are seen as random variables by GMM, it is assured also that those variables are not correlated. So the goal of adjusting the HIs original data to a feasible condition for GMM is guaranteed.

The remaining of the paper is organized as follows: next section brings some related work. Section 3 presents the unsupervised feature extraction methods used in the work. Section 4 describes how experiments were conducted. Section 5 explains the experimental results. Finally, Sect. 6 presents conclusions and final remarks.

2 Related Work

Feature Selection for Classification of Hyperspectral Data by SVM [27]. The main conclusion of this study is that, in HIs, the classification accuracy of Support Vector Machine (SVM) technique is influenced by the number of features used. Thus, it is shown that SVM is affected by Hughes phenomenon. The results presented in this paper show for AVIRIS sensor dataset that feature addition led to statistically significant decline in accuracy, for all sizes of training sets evaluated. With DAIS dataset, such decline was also observed, both for small training sets (25 samples per class or less) and for a large training sample. Therefore, it has been shown that feature selection provides a reduced feature set which gives similar classification accuracy to that of larger feature sets. In addition, the analysis of four different selection methods showed a wide variation between the number of features selected by each method, which highlights the influence of method choice in the process.

As in our study, this reference showed the manifestation of Hughes phenomenon in HIs and the influence of the DR method on the estimation of intrinsic dimensionality. However, classification was used instead of clustering and feature selection was used in place of extraction.

Feature Mining for Hyperspectral Image Classification [16]. Arguing that Hughes problem can be avoided by the use of few discriminating features, this article provides an overview of fundamental and modern feature extraction and selection which are applicable in HI classification. The common unavailability of class

labeling in this type of data is discussed. It is emphasized that in supervised parametric approaches, class separation measures require large training amount, which may not be accessible or physically feasible in practice. Therefore, nonparametric and unsupervised methods are preferred. It is informed that, current area research focus is on nonlinear feature extraction.

So this article shows that DR bypasses the Hughes problem in HIs and complements our work with feature extraction and selection methods description. One more argument is provided about the relevance of nonparametric, unsupervised and nonlinear categories (which are used in our work). Finally, the current relevance of research in nonlinear feature extraction is highlighted (which is also investigated by us).

Subspace Feature Analysis of Local Manifold Learning for Hyperspectral Remote Sensing Images Classification [9]. This reference is able to show more specifically that, in HIs, the use of features obtained by nonlinear extraction provides greater classification accuracy than original features. This type of conclusion also reinforces the evidence shown in our work: not only the HI feature extraction but also the nonlinear benefits.

Unsupervised Clustering and Active Learning of Hyperspectral Images with Nonlinear Diffusion [24]. In this work it is also shown that, in HIs context, linear DR methods generate better results compared to DR absence in classification process. Additionally, it is also shown that nonlinear DR is superior to linear in this context. This indicates that HIs are mapped in structures whose intrinsic dimensionality is less than the original, and also that such structures are particularly nonlinear. The study also shows that, in HIs, clustering methods that use Euclidean distance have lower performance when compared to methods that use other strategies.

So this related work is another source that shows HIs DR benefit (and particularly nonlinear DR). Additionally, the work also brings information about Euclidean strategies failure in this type of image, reinforcing the validity of suspicions along the same lines that were proven in our case.

Evaluation of Nonlinear Dimensionality Reduction Techniques for Classification of Hyperspectral Images [25]. Here it is also argued that, in many cases, combination of PCA with SVM provides good HIs classification results. However, hyperspectral scenes can bring complexity due to nonlinear effects, which can make PCA inappropriate. It is stated that for such scenes, therefore, it is necessary to choose carefully the output dimensionality and to consider the use of nonlinear DR techniques. It is also reported that the main disadvantage of nonlinear methods is characterized by high computational complexity in comparison with PCA.

Unsupervised Exploration of Hyperspectral and Multispectral Images [22]. Some of the most used unsupervised exploration methods (mainly DR and clustering techniques) are revised and discussed, focusing in hyper and multispectral

imagery. Use instructions are provided and vantages and disadvantages are highlighted in order to demystify common errors. It is argued that, when analysis focus is prediction, exploratory data analysis is frequently underrated. It is stated that, in fact, exploratory analysis provides wealth information itself and a first data insight that can be very useful, even if exploration is not the final goal. It follows that unsupervised methods provide valuable information and can be applied as a first approach for any hyper or multispectral analysis, given that such methods are hypothesis free.

3 Feature Extraction

The goal of Feature Extraction is to find the transformation function $f : R^m \rightarrow R^d$ that can map the point $\{x_i \in R^m\}_{i=1}^n$ to $y_i = f(x_i)$, where $\{y_i \in R^d\}_{i=1}^n$ and $d \leq m$, for all n points in the dataset, such that most data information is kept. The f term can be a linear or nonlinear transformation. In the next three subsections, we show the algorithms selected to work with.

3.1 Principal Component Analysis

Linear DR is widely used for analyzing high dimensional and noisy data [7]. Part of the success of linear methods comes from the fact that they have simple geometric interpretations and typically attractive computational properties [7]. Basically, in linear methods, we seek a projection matrix T that maps the samples in the original feature space (R^m) to a linear subspace in R^d, with $d < m$. Given n m-dimensional data points $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in R^{m \times n}$ and a choice of dimensionality $d < m$, the goal is to optimize some objective function $f_X(.)$ to produce a linear transformation $T \in R^{d \times m}$, and call $Y = TX \in R^{d \times n}$ the lower dimensional transformed data.

Principal Component Analysis is the most widely known linear method for data compression and Feature Extraction. It assumes that vectors are in an Euclidean space and from this assumption it searches for a new basis with d span vectors that maximizes the variance in the new representation. In this case, d is the intrinsic dimensionality and it is less than the original data dimension.

3.2 Isometric Feature Mapping

ISOMAP was one of the pioneering algorithms in nonlinear DR. It combines the major algorithmic advantages of PCA and Multidimensional Scaling (MDS) - computational efficiency, global optimality, and asymptotic convergence guarantees - with the flexibility to learn a broad class of nonlinear structures [33]. The basic idea of the ISOMAP algorithm is first to build a graph by joining the k-nearest neighbors (KNN) in the input space, then compute the shortest paths between each pair of vertices in the graph and, knowing the approximate distances between points, find a mapping to an Euclidean subspace of R^d that preserves those distances. The hypothesis of the ISOMAP algorithm is that the

shortest paths in the KNN graph are good approximations for the true distances in the original data structure. ISOMAP algorithm can be divided into three main steps:

1. From the input data $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in R^m$ build an undirected proximity graph using the KNN rule or the ϵ-neighborhood rule [21];
2. Compute the pairwise distance matrix D using n executions of the Dijkstra's algorithm or one execution of the Floyd-Warshall algorithm;
3. Estimate the new coordinates of the points in an Euclidean subspace of R^d by preserving the distances through the MDS method.

3.3 Locally Linear Embedding

The ISOMAP algorithm is a global method in the sense that, to find the new coordinates of a given input vector $\boldsymbol{x}_i \in R^m$, it uses information from all the samples through the matrix B. On the other hand, LLE, as the name emphasizes, is a local method where the new coordinates of any $\boldsymbol{x}_i \in R^m$ depend only on the point neighborhood. The main hypothesis behind LLE is that for a sufficient density of samples, it is expected that a vector $\boldsymbol{x}_i$ and its neighbors define a linear patch (i.e. all belonging to an Euclidean subspace [29]). It is possible to characterize the local geometry by linear coefficients this way:

$$\hat{\boldsymbol{x}}_i \approx \sum_j w_{ij} \boldsymbol{x}_j, \qquad \text{for} \qquad \boldsymbol{x}_j \in N(\boldsymbol{x}_i), \tag{1}$$

that is, we can reconstruct a vector as a linear combination of its neighbors.

Basically, the LLE algorithm requires as inputs a $n \times m$ data matrix X, with rows $\boldsymbol{x}_i$, a desired number of dimensions $d < m$ and an integer $k > d + 1$ for finding local neighborhoods. The output is a $n \times d$ matrix Y, with rows $\boldsymbol{y}_i$. The LLE algorithm can be divided into three main steps [29,30]:

1. From each $\boldsymbol{x}_i \in R^m$ find its k nearest neighbors;
2. Find the weight matrix W which minimizes the reconstruction error for each data point $\boldsymbol{x}_i \in R^m$;

$$E(W) = \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \sum_j w_{ij} \boldsymbol{x}_j \right\|^2, \tag{2}$$

where $w_{ij} = 0$ unless $\boldsymbol{x}_j$ is one of $\boldsymbol{x}_i$'s k-nearest neighbors and for each i, $\sum_j w_{ij} = 1$.

3. Find the coordinates Y which minimize the reconstruction error using the optimum weights;

$$\Phi(Y) = \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \sum_j w_{ij} \boldsymbol{y}_j \right\|^2 , \tag{3}$$

subject to the constraints $\sum_i Y_{ij} = 0$ for each j, and $Y^T Y = I$.

Algorithm 1. Principal Component Analysis

function PCA(X)
  Compute the sample mean and the sample covariance matrix by:
  $\mu_x = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$
  $\Sigma_x = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \mu_x)(\boldsymbol{x}_i - \mu_x)^T$
  Compute the eigenvalues and eigenvectors of Σ_x
  Define the transformation matrix $T = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_d]$ with the d eigenvectors associated to the d largest eigenvalues.
  Project the data X into the PCA subspace:
  $\boldsymbol{y}_i = T\boldsymbol{x}_i$ for $i = 1, 2, ..., n$
  return Y
end function

Algorithm 2. Isometric Feature Mapping

function ISOMAP(X)
  From the input data $X_{m \times n}$ build a KNN graph.
  Compute the pairwise distances matrix $D_{n \times n}$.
  Compute $A = -\frac{1}{2} D$.
  Compute $H = I - \frac{1}{n} U$, where U is a $n \times n$ matrix of 1's.
  Compute $B = HAH$.
  Find the eigenvectors and eigenvalues of the matrix B.
  Select the top $d < m$ eigenvectors and eigenvalues of B and define:

$$\tilde{V} = \begin{bmatrix} | & | & ... & ... & | \\ | & | & ... & ... & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & ... & ... & \boldsymbol{v}_d \\ | & | & ... & ... & | \\ | & | & ... & ... & | \end{bmatrix}_{n \times d} \tag{4}$$

$$\tilde{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_d) \tag{5}$$

  Compute $\tilde{Y} = \tilde{\Lambda}^{1/2} \tilde{V}^T$
  return $\tilde{Y}$
end function

Algorithm 3. Locally Linear Embedding

function LLE(X, K, d)
  From the input data $X_{m\times n}$ build a KNN graph.
  for $\boldsymbol{x}_i \in X^T$ **do**
    Compute the $K \times K$ matrix C_i as:

$$C_i(j,k) = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_k)^T. \tag{6}$$

    Solve the linear system $C_i\boldsymbol{w}_i = \mathbf{1}$ to estimate the weights $\boldsymbol{w}_i \in R^K$.
    Normalize the weights in $\boldsymbol{w}_i$ so that $\sum_j \boldsymbol{w}_i(j) = 1$.
  end for
  Construct the $n \times n$ matrix W, whose lines are the estimated $\boldsymbol{w}_i$.
  Compute $M = (I - W)^T(I - W)$.
  Find the eigenvalues and eigenvectors of the matrix M.
  Select the bottom d nonzero eigenvectors of M and define:

$$Y = \begin{bmatrix} | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \dots & \boldsymbol{v}_d \\ | & | & \dots & \dots & | \\ | & | & \dots & \dots & | \end{bmatrix}_{n\times d}. \tag{7}$$

  return Y
end function

4 Methodology

Seven traditional HIs[1] were used for exploitation: Indian Pines, Salinas, SalinasA, Botswana, Kennedy Space Center, Pavia University, and Pavia Centre.

The purpose of this work is to verify if unsupervised Feature Extraction benefits HI clustering by GMM and if a nonlinear method is more efficient than a linear in this case. To conduct such verification, the performance of clusterings made on linear reduction (PCA), nonlinear reduction (ISOMAP and LLE) and nonDR (NDR) are compared. As Ground Truths were available, performances were evaluated by external measures. Measures used are: Rand [28], Jaccard [15], Kappa [5], Entropy and Purity [32]. Performances were then compared by Friedman [10] and Nemenyi [26] tests. All steps are shown in Fig. 1.

Regarding the application of nonlinear algorithms, most images were too large for our 24 GB RAM memory. This problem happened especially for ISOMAP algorithm second step. So, for those images, sets of representative pixels were chosen using Bagging [2] technique. Those sets were used by both linear and nonlinear DR methods in place of the original image. Clustering was made on those sets in the NDR case also.

[1] All information regarding those images (as number of classes, reflectance and radiance characteristics, number of bands, pre-processing, source equipment, date, and files) can be found at: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

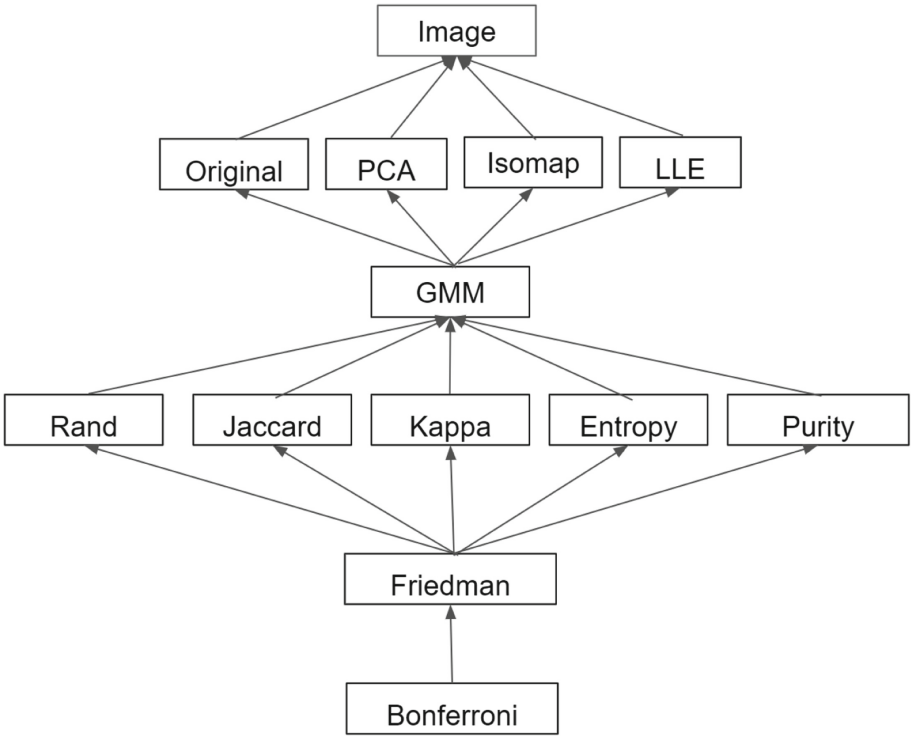

Fig. 1. Experiments execution flow pipeline

For each image, the three DR methods were applied. The choice of target dimensionality strongly affects the results and is usually defined by the intrinsic dimensionality. Such discovery can be made by several methods [3,6,13,23] and its strongly related to each image itself and its particular properties. The topic is currently under investigation in the area and, as it is not focus of this work, we have choosen a well-known basic empirical method. The method consists in the analysis of the sum of the first n largest eigenvalues over the sum of all eigenvalues of the transformation matrix of each DR method [6]. We seek the point where 95% of data variance is retained. The amount of eigenvalues that reach this point is understood as the intrinsic dimensionality.

In the image pre-processing step[2], all mixed pixels were categorized to a background class which label is zero. So, after DR and before clustering, we adopted a simple procedure of setting those pixel vectors coordinates to (0,...,0) - as background class is not in the discrimination interest. The goal of this procedure is to avoid original geometry distortion in DR and also to help raising clustering performance considerably (as this class is a pixel mixture that is not in any other class defined by the expert). We notice that the procedure didn't

[2] This step was not done by us, image files used were already pre-processed.

produce much distortion in clustering (most class zero pixels in the Ground Truth were kept in the same group after clustering).

For each vector set in the reduced dimensional space, several executions of the clustering procedure were applied due to its known effect of random centroids initialization. As this particular image set is already labeled by specialists, external clustering validation is more robust than internal. So external measures were used. All metrics adopted assume higher values for better clusterings. Specifically for the Kappa measure, it must be ensured a reliable semantic correspondence between clustering and expert's labeling. We mapped this problem to one of finding the minimum pairing cost in a bipartite graph, where there is label association to minimize a cost function. The Munkres algorithm was used in this Optimum Allocation Problem [18].

All measures were applied to all clustering results and the representative value for each measure was given by the maximum value obtained. Next, for each image, Friedman test indicates if there was significant statistical difference between some pair of measure values (considering a pair of two lists, each corresponding to a DR method, whose elements are the external indices maximum values). Nemenyi test tells then which measure pairs were different. The results make possible to analyze if DR was benefical to clustering and in which images nonlinear overcame linear reduction. Tests results are shown in the next section.

The programming language and libraries used were: Python 3.6.5; NumPy 1.15.0; SciPy 1.0.0; Scikit-learn 0.20.0; Matplotlib 3.0.0, and STAC. We used the Scikit-learn implementations for all DR methods.

5 Experiments and Results

We can find in Tables 1, 2, 3, 4 and 5[3] the clustering validation measure maximum values by DR method, where each table corresponds to an image which Friedman test pointed significant difference between some pair of DR strategy. Usual p-value > 0.05 criteria to indicate equivalence was adopted. The last row presents the measure values mean given a DR method.

In Table 6 we present the Friedman p-values by image. Friedman test only tells us if some method pair differed. We need a post-hoc test to discover which pairs were different [8,11]. We show in Table 7 the Nemenyi p-values smaller than 0.05 and its occurrence pair by image.

We can see for Botswana that, for a trust level of 95% (i.e. significance level of 5%), there is evidence that ISOMAP produces better clustering results than DR absence. This imply that DR was benefical to this image. There is also evidence that ISOMAP got better attributes than PCA, which justifies the better adequacy of a nonlinear DR method for this data. Following the same reasoning, in KSC we can say that LLE is superior than PCA. In Pavia, LLE is superior than NDR and PCA. In PaviaU PCA had significant difference in comparison with NDR and was also superior than both nonlinear methods. In

[3] We truncated all values at the third decimal point or at the first decimal point possible.

Table 1. Botswana: clustering validation measure maximum values by DR method

	NDR	PCA	ISOMAP	LLE
Entropy	−0.177	−0.194	**−0.144**	−0.159
Jaccard	0.802	0.820	**0.836**	0.803
Kappa	0.802	0.797	**0.861**	0.804
Purity	0.952	0.948	**0.962**	0.955
Rand	0.995	0.995	**0.996**	0.994
Mean	0.674	0.673	**0.702**	0.679

Table 2. KSC: clustering validation measure maximum values by DR method

	NDR	PCA	ISOMAP	LLE
Entropy	−0.683	−0.705	−0.707	**−0.289**
Kappa	0.602	0.577	0.602	**0.792**
Jaccard	0.655	0.656	0.666	**0.698**
Purity	0.830	0.819	0.829	**0.929**
Rand	0.946	0.941	0.944	**0.989**
Mean	0.469	0.457	0.466	**0.623**

Table 3. PaviaU: clustering validation measure maximum values by DR method

	NDR	PCA	ISOMAP	LLE
Entropy	−0.210	**−0.182**	−0.196	−0.284
Jaccard	0.783	**0.797**	0.792	0.753
Kappa	0.758	**0.826**	0.814	0.771
Purity	0.939	**0.949**	0.944	0.936
Rand	0.980	**0.984**	0.982	0.951
Mean	0.650	**0.674**	0.667	0.625

Table 4. Pavia: clustering validation measure maximum values by DR method

	NDR	PCA	ISOMAP	LLE
Entropy	−0.090	−0.084	−0.080	**−0.066**
Jaccard	0.747	0.882	0.889	**0.897**
Kappa	0.864	0.889	0.894	**0.932**
Purity	0.978	0.978	0.980	**0.986**
Rand	**0.994**	0.993	0.993	**0.994**
Mean	0.698	0.731	0.735	**0.748**

Table 5. SalinasA: clustering validation measure maximum values by DR method

	NDR	PCA	ISOMAP	LLE
Entropy	−0.358	**−0.228**	−0.266	−0.246
Jaccard	0.302	**0.559**	0.458	0.347
Purity	0.882	0.910	0.906	**0.931**
Kappa	0.857	**0.869**	0.859	0.843
Rand	0.836	**0.870**	0.861	0.852
Mean	0.503	**0.596**	0.563	0.545

Table 6. Friedman p-value by image

	p-value
Botswana	0.004
KSC	0.001
PaviaU	0.0000001
Pavia	0.003
Salinas	0.175
SalinasA	0.002
Indian Pines	0.322

Table 7. Nemenyi p-values

	Botswana	KSC	PaviaU	Pavia	SalinasA
NDR × PCA			0.007		0.001
NDR × ISOMAP	0.010				
NDR × LLE				0.004	
PCA × ISOMAP	0.004				
PCA × LLE		0.001	0.0006	0.010	
ISOMAP × LLE		0.019	0.027		

SalinasA we can only conclude that PCA yeld better results in comparisons to NDR. So, in four images DR brought benefit to clustering. Additionally, in three images we saw that nonlinear DR performed better than linear.

6 Conclusion

As one cannot state beforehand theoretically if HIs naturally suffer from Curse of Dimensionality, in this work we have shown empirical evidence that unsupervised Feature Extraction significantly increased GMM clustering performance in the majority of studied images. Significant statistical difference between

DR and NDR was found using nonparametric multiple comparisons hypothesis tests on several large real images. Unsupervised DR and unsupervised learning approaches were used to reach this conclusion and Ground Truth was used only to compute several external measures in clustering evaluation. Additionally, label matching by Munkres algorithm increased Kappa values and emphasized performance differences between DR methods in this matter.

The efficiency of nonlinear DR compared to linear in HI context cannot be shown beforehand theoretically also. Our experiments have shown that nonlinear DR is more appropriate for the majority of HIs in which DR brought significant contribution to clustering. We gave theoretical justification to why nonlinear DR can be more adequate by presenting the PCA, ISOMAP and LLE strategies and algorithms.

Future works may include: other Feature Extraction techniques, specially modern nonlinear algorithms; more sophisticated intrinsic dimensionality discovery strategies; other HI images; more clustering methods; use of internal clustering validity indexes to test a complete unsupervised procedure; use of supervised classifiers ensemble to obtain a supervised perspective. We believe that such expansions can bring more evidence for the Curse of Dimensionality in HIs and build a stronger basic study for unsupervised Feature Extraction in HIs, compiling the main interest points in the area and serving as guide for first contact.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_27
2. Breiman, L.: Bagging predictors (1994)
3. Camastra, F.: Data dimensionality estimation methods: a survey. Pattern Recogn. **36**(12), 2945–2954 (2003). https://doi.org/10.1016/S0031-3203(03)00176-6. http://www.sciencedirect.com/science/article/pii/S0031320303001766
4. Camps-Valls, G., Tuia, D., Gómez-Chova, L., Jiménez, S., Malo, J.: Remote sensing image processing. In: Remote Sensing Image Processing (2011)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960). https://doi.org/10.1177/001316446002000104
6. Cox, T.F., Cox, M.A.A.: Multidimensional scaling. In: Monographs on Statistics and Applied Probability, vol. 88. Chapman & Hall, New York (2001)
7. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: survey, insights, and generalizations. J. Mach. Learn. Res. **16**, 2859–2900 (2015)
8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. Ding, L., Tang, P., Li, H.: Subspace feature analysis of local manifold learning for hyperspectral remote sensing images classification. Appl. Math. Inf. Sci. **8**(4), 1987 (2014)
10. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**(200), 675–701 (1937). https://doi.org/10.1080/01621459.1937.10503522

11. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf. Sci. **180**, 2044–2064 (2010). https://doi.org/10.1016/j.ins.2009.12.010
12. Goetz, A.F., Vane, G., Solomon, J.E., Rock, B.N.: Imaging spectrometry for earth remote sensing. Science **228**(4704), 1147–1153 (1985). https://doi.org/10.1126/science.228.4704.1147. http://science.sciencemag.org/content/228/4704/1147
13. He, J., Ding, L., Jiang, L., Li, Z., Hu, Q.: Intrinsic dimensionality estimation based on manifold assumption. J. Vis. Commun. Image Representation **25**(5), 740 – 747 (2014). https://doi.org/10.1016/j.jvcir.2014.01.006. http://www.sciencedirect.com/science/article/pii/S1047320314000078
14. Hughes, G.: On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory **14**(1), 55–63 (1968). https://doi.org/10.1109/TIT.1968.1054102
15. Jaccard, P.: The distribution of the flora in the alpine zone 1. New Phytologist. **11**(2), 37–50 (1912). https://doi.org/10.1111/j.1469-8137.1912.tb05611.x. https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x
16. Jia, X., Kuo, B., Crawford, M.M.: Feature mining for hyperspectral image classification. Proc. IEEE **101**(3), 676–697 (2013). https://doi.org/10.1109/JPROC.2012.2229082
17. Kendall, M.: A course in the geometry of N dimensions. In: Dover Books on Mathematics. Dover Publications (2004). https://books.google.com.br/books?id=_dFJ6pSzRLkC
18. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logistics Q. **2**(1–2), 83–97 (1955). https://doi.org/10.1002/nav.3800020109. https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109
19. Landgrebe, D.: Multispectral data analysis: a signal theory perspective, January 1998
20. Landgrebe, D.A.: Information extraction principles and methods for multispectral and hyperspectral image data (1998)
21. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
22. Marini, F., Amigo, J.M.: Chapter 2.4 - Unsupervised exploration of hyperspectral and multispectral images. In: Amigo, J.M. (ed.) Hyperspectral Imaging, Data Handling in Science and Technology, vol. 32, pp. 93–114. Elsevier, Amsterdam (2020). https://doi.org/10.1016/B978-0-444-63977-6.00006-7. http://www.sciencedirect.com/science/article/pii/B9780444639776000067
23. Miranda, G.F., Thomaz, C.E., Giraldi, G.A.: Geometric data analysis based on manifold learning with applications for image understanding. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), pp. 42–62, October 2017. https://doi.org/10.1109/SIBGRAPI-T.2017.9
24. Murphy, J.M., Maggioni, M.: Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. IEEE Trans. Geosci. Remote Sens. **57**(3), 1829–1845 (2018)
25. Myasnikov, E.: Evaluation of nonlinear dimensionality reduction techniques for classification of hyperspectral images, July 2018
26. Nemenyi, P.: Distribution-Free Multiple Comparisons. Princeton University (1963). https://books.google.com.br/books?id=nhDMtgAACAAJ
27. Pal, M., Foody, G.M.: Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Remote Sens. **48**(5), 2297–2307 (2010)

28. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971). https://doi.org/10.1080/01621459.1971.10482356. https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356
29. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)
30. Saul, L., Roweis, S.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. J. Mach. Learn. Res. **4**, 119–155 (2003)
31. Scott, D.: Multivariate Density Estimation: Theory, Practice, and Visualization. A Wiley-Interscience Publication, Wiley (1992). https://books.google.com.br/books?id=7crCUS_F2ocC
32. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. https://ieeexplore.ieee.org/document/6773024/
33. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
34. Zhang, L., Zhang, L., Tao, D., Huang, X.: On combining multiple features for hyperspectral remote sensing image classification. IEEE Trans. Geosci. Remote Sens. **50**(3), 879–893 (2012). https://doi.org/10.1109/TGRS.2011.2162339