UNSUPERVISED CLASSIFICATION OF AVIRIS-NG HYPERSPECTRAL IMAGES

Kangning Cui

City University of Hong Kong
Department of Mathematics
83 Tat Chee Ave, Hong Kong

Robert J. Plemmons

Wake Forest University
Departments of Computer Science and Mathematics
Winston-Salem, NC 27109

ABSTRACT

In hyperspectral imaging for remote sensing, learning from unlabeled data by unsupervised methods is very challenging and it is the subject of considerable recent interest since the collection of large datasets by aircraft, UAVs and satellites has become ubiquitous. We experiment with unsupervised endmember extraction and classification of hyperspectral data collected over India by NASA's AVIRIS-NG airborne remote sensor. We have downloaded some of this data from the NASA-JPL portal in Pasadena, CA, for the purpose of studying land cover and land usage, and especially forests, in India. We report on results from our experiments with unsupervised endmember-based methods and clustering methods for classifying images from a mixed forest region that we selected from the Shoolpaneshwar Wildlife Sanctuary in Western India. Randomized numerical methods are used to speed up the large-scale computations.

Index Terms— Unsupervised hyperspectral classification, clustering, randomized computations, endmembers, unlabeled AVIRIS-NG data, forests, India.

1. INTRODUCTION

With the development of advanced remote sensing techniques, hyperspectral imagery (HSI) with high spatial and spectral resolution can be obtained by aircraft, drones and satellites. These newly collected high resolution HSI datasets enable us to classify the Land Use Land Cover (LULC) information of pixels more accurately and the classification results can be utilized in numerous applications.

The Indian Space Research Organisation (ISRO) and the United States National Aeronautics and Space Administration (NASA) launched the AVIRIS-NG (Next Generation) campaign in India beginning in 2015 [1]. This airborne HSI system measures a spectral range from 380nm to 2500nm, with a 5nm difference between neighboring spectral bands, while the spatial resolution of the system is 4.1m. Typical flight altitudes are 4-8 km. This study uses one of these datasets, collected in a flight path over a wildlife refuge in India. The data that we downloaded from JPL portal has already been ortho-corrected and atmospherically corrected, but unlabeled since the labeling process is expensive and time consuming. We also removed the bands influenced by water vapor absorption. The reduced data is then further modified by appropriate reduction methods, as part of our study.

Unsupervised classification does not require labels, and it is challenging to produce accurate result and analyze results quantitatively. In [2], several unsupervised clustering methods are tested on 4 publicly available datasets, but their work does not involve endmembers. In [3], spectral library is learned and can be used to unsupervised map mineral composition, involving remote exploration in astrobiology, such as is done by NASA's Mars Rover. In our work, we take the advantage of endmember-based and recent unsupervised methods.

Land-use change and the resultant changes in land cover impact a wide variety of ecological processes with environmental and economic implications. Clearly these implications are extremely important to India, the country of our focus in this study. As an important illustration, tree species can be identified and then applied to characterize species diversity using appropriate indices, and vegetation indices based on leaf surface reflectance can be used to approximate the fuel load of the forest [4] to estimate the possibility of destructive fires.

Our paper is organized as follows. In Section 2 the unlabeled data we are using are described, and the methods involved are introduced. Section 3 shows the experimental unsupervised classification results. Section 4 concerns the classification results and the additional work on unsupervised classification planned for the near future.

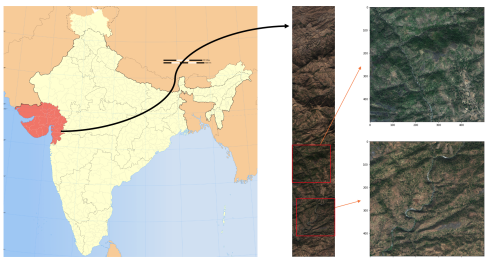


Fig. 1: Left: Location of study site in India, Middle: AVIRIS-NG dataset containing mixed forest, Right: Two regions covered by the forest. Top region contains the most trees.

2. DATA AND METHODS USED FOR THE STUDY

2.1. Study Site

Our study area is located in the Shoolpaneshwar Wildlife Sanctuary (SWS) in northwest India. See the map in Figure 1. The SWS is a protected area in Gujarat state, and is $607.7km^2$ large. It encompasses mixed dry deciduous forest, southern tropical moist deciduous forest, pockets of moist teak forests, intrusive agricultural fields in border regions, small rivers and two water reservoirs. The unlabeled AVIRIS-NG dataset used in this study was collected in a flight path over a mixed dry deciduous forest region that we downloaded in 2020 from the NASA - JPL Portal at Pasadena. The dimensions of our selected HSI datacube at SWS is 11072×657 pixels, with 425 spectral bands [1]. The top region at the right-hand side of Figure 1 is chosen in this paper since it contains the most trees.

2.2. Dimensionality Reduction

We now consider an important preprocessing step, reduction of the number of bands in HSI datasets prior to classification. Fast randomized truncated SVDs are used here to modify and to speed up Principle Component Analysis (PCA), and we also use Minimum Noise Fraction (MNF), Independent Component Analysis (ICA), and the Random Projection (RP) method.

Dimensionality reduction techniques are unsupervised and widely used as the main preprocessing step when dealing with HSI data classification. This is because high spatial resolution HSI contains redundant information, and it's processing often leads to unnecessary high computational cost. It often makes the classification steps, generally involving numerical linear algebra on large datasets, computationally feasible.

Fortunantly, randomization methods can be used to greatly speed up linear algebra computations, including truncated SVDs, see [5]. Here, we use fast randomized truncated SVDs [6] to speed up PCA.

The PCA method finds an orthogonal basis of an optimal linear subspace, generally using a truncated SVD, to clean up the highly correlated spectral features of neighboring bands. It's properties are well known for use in HSI dimensionality reduction. The related MNF Transformation is also widely used. It incorporates noise reduction to improve the Signal-to-Noise (SNR) ratio for HSI datasets, see, e.g. [7]. MNF aims to determine the inherent dimensionality of HSI data in order to segregate noise in the data, and to subsequently reduce the computational requirements for subsequent processing, ee, e.g. [8]. The ICA method generates statistically independent components by estimating the eigenvectors of covariance matrix, and then computes the rotation matrix that maximize the independence of components. Finally, the Random Projection method projects the data to a lower dimension using a randomly generated matrix with the set of entries having a Gaussian distribution and is normalized.

2.3. Unsupervised Endmember Extraction Methods

Unsupervised classification methods require an accurate prior knowledge of the number of endmembers, see e.g. [7], and some methods require the number of clusters, e.g., [2, 9, 10]. HSI endmember extraction is an important step due to significantly improved spatial and spectral resolutions of the imaging sensors of the New Generation AVIRIS images. Since the endmember spectral signatures of our data are not known, the endmember extraction as well as the subsequent spectral unmixing and classification processes must be performed in an unsupervised way. In this study, we compare two convex-based endmember extraction methods: ATGP and N-Finder.

The estimation of the number of endmembers is crucial for unlabeled HSI datasets using convex-based endmember extraction methods, see [11]. The Hyperspectral Signal Subspace Identification by Minimum Error can perform well to estimate the number of endmembers [12]. We apply it on our dataset and decide to use 7 endmembers. Next, we describe the endmember extraction methods used in this study.

Briefly, the Automatic Target Generation Process (ATGP) makes iterative use of orthogonal subspace projections to identify the endmembers of significant interest, and stops once extract a pre-chosen p endmember targets [13]. Also, the N-finder method generates simplexes and finds pixels that maximize their volumes. We use a improved version [14] that takes the ATGP method to initialize the pixel samples.

2.4. Unsupervised Classification Methods

Owing to the enhancement of the resolution of remote sensing spectrometers, the application of HSI classification has become much more significant in recent years, see details in [9]. Classification methods can be roughly divided into three major types: supervised, semi-supervised and unsupervised methods. Supervised methods are widely used for this task for cases where a sufficient number of pixels have known values, i.e., are labeled by field trips or other means. Semi-supervised systems with a relatively small number of labeled pixels are also used in HSI classification when techniques such as self-training methods are used, see [9, 15]. Several unsupervised HSI classification methods are recently tested and compared on different datasets in [2]. There, K-means, non-negative matrix factorization, and diffusion learning methods, which are clustering-based methods that produce reasonable results for selected HSI datasets. See the references [7, 10, 16].

In practice, however, the problem of unsupervised classification of hyperspectral data is quite challenging. This is due, in part, to such factors as high dimensionality, the presence of substantial noise, variability of the spectra, overlap of the classes, and inaccurate estimation of the number of endmembers. We use here unsupervised techniques to classify or to

cluster the pixel dataset. Our purpose is to consider the applications of this important airborne AVIRIS-NG unlabeled data for the study of land cover and land use and, especially, for forest health and diversity in India. Recall that the AVIRIS-NG data from India on the NASA-JPL portal is unlabeled. We first look at the closely related topic of methods for hyperspectral pixel clustering.

The K-means clustering splits sets of pixels into n groups with equal variance in order to minimize the within-cluster sum of the squares. One potential issue with K-means is that it can only produce spherical clusters. The Gaussian Mixed Model (GMM) is a weighted mixture of multivariate Gaussians which can represent skewed clusters [17]. Graph-based approaches can contribute to some unsupervised HSI classification tasks, however, computational costs can be higher than for methods such as K- means. The improved Anchor-based Graph Clustering (AGC-I) method in [10] uses a small portion of the dataset to optimize the affinity matrix by non-negative and orthogonal constraints, and then uses K-means to cluster the pixels. In this paper, we apply K-means and GMM on the indicator matrices that AGC-I produce.

Next we consider two commonly used classification methods, abbreviated as SID and SAM, e.g. [7]. The spectral information divergence (SID) algorithm measures the relative entropy between pixel vectors. It is capable of measuring both spectral similarity and discriminability. The Spectral Angle Mapper (SAM) is a method to compute a geometric angle in N-dimensional space between two spectra. The SAM algorithm maps spectral similarity between pixel vectors.

3. NUMERICAL EXPERIMENTS

In this section, we describe the results of dimensionality reduction methods on our AVIRIS-NG datasets and on endmembers extracted by multiple techniques. We then use the endmembers to perform classification. Code for this study can be found at: github.com/ckn3/UnsupAVIRISNG.

3.1. Dimensionality Reduction

Dimensionality reduction is a crucial step for large hyperspectral datasets. We keep 12 components for rSVD-based PCA, 83 components for MNF, and 30 components for ICA and RP. The reduced and denoised datasets reconstructed by PCA, MNF and ICA are further used to extract endmembers. The average reconstruction error for each pixel is defined as $err = \frac{(X-\hat{X}^2)}{N}$, where X is the original dataset, $\hat{X}$ is the reconstructed data, and N is the number of pixels. Here the average reconstruction error for PCA, randomized SVD, ICA, and MNF are: 0.0156, 0.0136, 0.0323, and 3.6402, respectively.

3.2. Endmember Extraction Results

The results derived by ATGP and N-FINDR are displayed in Figure 2, using spectral signatures of the endmembers. It shows that ATGP and N-FINDR get similar results, by comparing the two columns. By comparing our results with spectral libraries in the papers ([4, 7]) for SWS, we see that our extracted endmembers correspond nicely to the true LULC classes. Both ICA and rSVD-based PCA reconstructed data can produce endmembers that very similar to those of the original dataset. The MNF reconstructed data produces 2 endmembers (EM3, 5) that are noisy, but extracts 4 tree species as well. The NFINDR results are more robust with less randomness, so we use NFINDR for further processing.

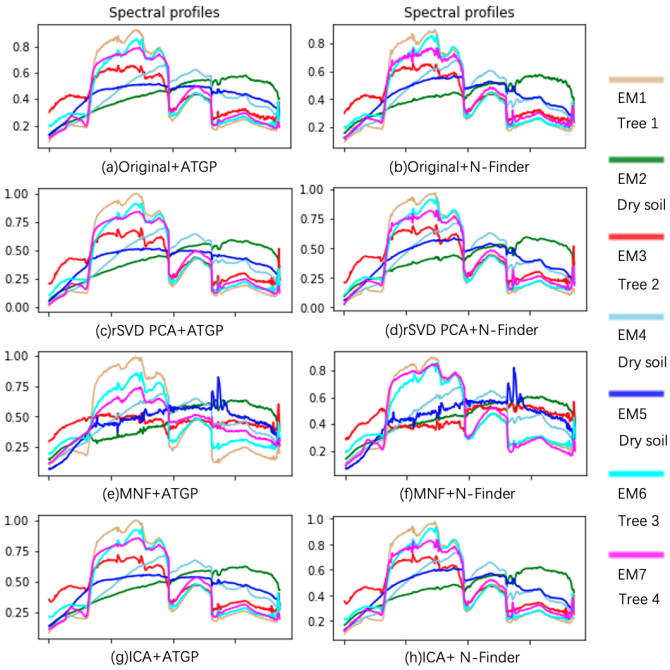


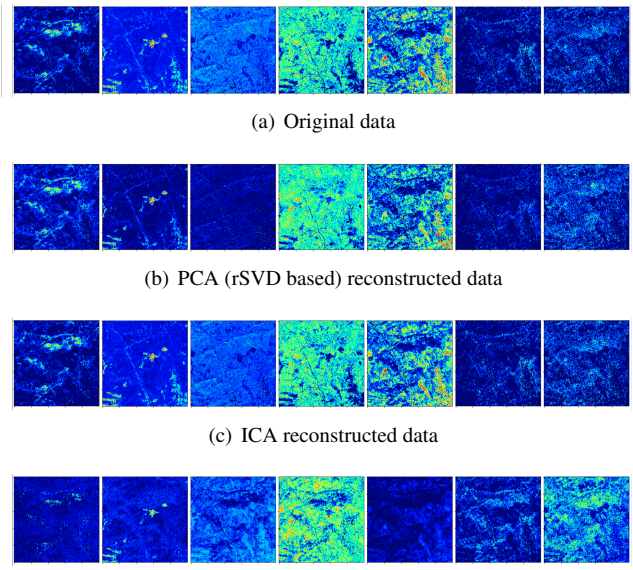
Fig. 2: Spectral Signatures of the Extracted Endmembers

The abundance maps shown in the next section are based on the endmembers obtained in this section, using the techniques we have discussed. The maps help us to verify and illustrate the classes associated with the extracted endmembers.

3.3. Abundance Maps

NNLS derived Spectral Abundance Maps are plotted to visualize the spectral abundance of each endmember. It is applied to each pixel x in the dataset using the endmember signatures Z [18], i.e. $\min \|x - Zd\|^2$, where d is the weight vector of endmembers. All the abundances are normalized, so the abundances of each pixel add up to 1.

Figure 3 shows the NNLS abundance maps using endmembers derived from the datasets we reconstruct and the



(a) Original data

(b) PCA (rSVD based) reconstructed data

(c) ICA reconstructed data

(d) MNF reconstructed data

Fig. 3: The NNLS Abundance Maps of the Endmembers: EM1 - EM7 from left to right

original dataset, using false colors. Together with these endmembers, the abundance maps can be used to align distinct classes. A brighter color means the corresponding endmember has a higher abundance. The rSVD-based PCA and ICA produce comparative results to those using the original dataset. The MNF reconstructed data gives nice results on identifying tree species, but mixes different dry soil types, which is indicated by the endmemebers in Figure 2.

3.4. Clustering & Classification Results

Here we apply some clustering methods, and classification methods based on spectral profiles of endmembers extracted in Section 3.2. All of these methods are unsupervised and labels are not available. In general, the clustering methods produce comparable results to those of endmember-based classification methods [9].

We perform K-means and GMM on the original dataset as well as several datasets we derived and get comparable results, see Figure 4. For the AGC-I, 3.3% of the pixels are used as anchors to estimate the affinity matrices. The GMM highly improves the MNF+K-means result, since GMM can mitigate the effect of inaccurate approximation of noise in MNF.

Other unsupervised classification methods for HSI, SID and SAM, are tested based on the extracted endmembers as well. The ICA and PCA reconstructed data are used in these approaches, see Figure 5. In general, SAM outperforms SID, identifying more tree pixels which are correspond to the RGB scene. ICA reconstructed data distinguishes different tree species and dry soil better than the PCA reconstructed data, which has a very low rank. The SID and SAM results are

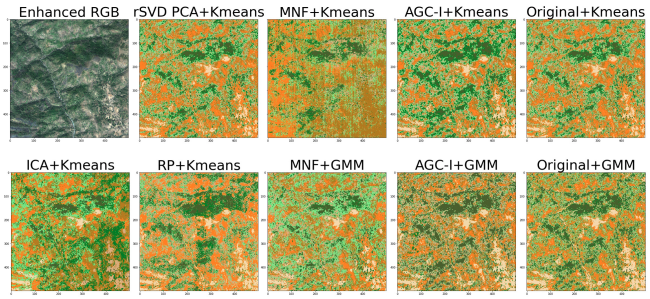


Fig. 4: Results of clustering based classification

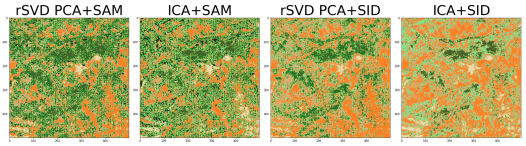


Fig. 5: Results of endmember based classification

similar to the clustering based classification results.

4. CONCLUSIONS AND FUTURE WORK

In this study, we have discussed and tested several methods for dimensionality reduction and endmember extraction, to be used in the challenging problems of unsupervised clustering and classification. Our purpose was to compare methods in terms of feasibility for use in land cover and land use (LCLU) studies in India using AVIRIS-NG hyperspectral data. We used test data from the Shoolpaneshwar Wildlife Sanctuary in Gujarat state, located in northwestern India. That site has been used for several recent LCLU studies, see e.g. [1, 4, 7].

In our tests we found that for the preprocessing step of dimensionality reduction, the PCA and ICA reconstructed datasets have the smallest reconstruction errors, and further produce very similar results as in processing the original dataset. The PCA method is implemented using a randomized truncated SVD method as well for numerical efficiency, which performs the fastest and preserves the fewest principal components, and has a small reconstruction error. Also, GMM can reduce the effect of inaccuracy of MNF noise estimation.

The endmember extraction methods yield several interesting results.

- The endmembers extracted from the PCA and ICA reconstructed datasets are similar to the endmembers extracted from Original dataset.
- The MNF reconstructed data produces a few noisy endmembers, but extracts tree species well.
- The endmembers derived from the PCA and ICA reconstructed data by N-finder were used further in our best classification results.

For major future work, we plan to test the diffusion machine learning methods [16], which take advantage of diffusion distance and avoid non-linear distribution of data which handicaps methods such as K-means. This work will be in collaboration with Prof. James Murphy and Sam Polk at Tufts University. More randomized dimensionality reduction methods will be considered in order to improve the computational efficiency, e.g. [19]. We will also improve the quality of extracted endmembers by normalizing and considering spectral variability the HSI data. This is often necessary since a single spectral profile is not always sufficient to represent a material [11].

5. ACKNOWLEDGMENTS

The authors would like to thank PI Carola Schoenlieb and members of the "Robust and Efficient Analysis Approaches of Remote Imagery for Assessing Population and Forest Health in India" project funded at the University of Cambridge. Interaction on this project motivated our work from the beginning. Jason Cui and Professor Plemmons are collaborators, and Prof. Plemmons is one of the advisors of the project. They would also like to thank Professor Raymond Chan, Vice President of City University of Hong Kong, for his advice and support.

6. REFERENCES

[1] B.K. Bhattacharya, R.O. Green, S. Rao, M. Saxena, S. Sharma, K.A. Kumar, P. Srinivasulu, S. Sharma, D. Dhar, S. Bandyopadhyay, et al., "An overview of AVIRIS-NG airborne hyperspectral science campaign over india," *Current Science*, vol. 116, pp. 1082–1088, 2019.

[2] H. Yadav, A. Candela, and D. Wettergreen, "A study of unsupervised classification techniques for hyperspectral datasets," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 2993–2996.

[3] A. Candela, D. Thompson, S. Kogdule, S. Vijayarangan, K. Edelson, E. Noe Dobrea, and D. Wettergreen, "Estimation of Surface Reflectance and Mineral Composition by Combining In Situ and Remote Spectroscopic Measurements," in *AGU Fall Meeting Abstracts*, Dec. 2019, vol. 2019, pp. GC51E–1117.

[4] C.S. Jha, J. Singhal, C.S. Reddy, G. Rajashekar, S. Maity, C. Patnaik, A. Das, A. Misra, C.P. Singh, J. Mohapatra, et al., "Characterization of species diversity and forest health using aviris-ng hyperspectral remote sensing data.," *Current Science (00113891)*, vol. 116, no. 7, 2019.

[5] P. Martinsson and J. Tropp, "Randomized numerical linear algebra: Foundations and algorithms," *Acta Numerica*, vol. 29, pp. 403–572, 2020.

[6] J. Zhang, J. Erway, X. Hu, Q. Zhang, and R. Plemmons, "Randomized svd methods in hyperspectral imaging," *Journal of Electrical and Computer Engineering*, vol. 2012, 2012.

[7] L.K. Sharma and R.K. Verma, "Aviris-ng hyperspectral data analysis for pre and post mnf transformation using per-pixel classification algorithms," *Geocarto International. Taylor & Frances*, 2020.

[8] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 26, no. 1, pp. 65–74, 1988.

[9] W. Lv and X. Wang, "Overview of hyperspectral image classification," *Journal of Sensors*, vol. 2020, 2020.

[10] Y. Zhao, Y. Yuan, and Q. Wang, "Fast spectral clustering for unsupervised hyperspectral image classification," *Remote Sensing*, vol. 11, no. 4, pp. 399, 2019.

[11] L. Drumetz, J. Chanussot, C. Jutten, W. Ma, and A. Iwasaki, "Spectral variability aware blind hyperspectral image unmixing based on convex geometry," *IEEE Transactions on Image Processing*, vol. 29, pp. 4568–4582, 2020.

[12] J.M. Bioucas-Dias and J.MP. Nascimento, "Hyperspectral subspace identification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2435–2445, 2008.

[13] H. Ren and C. Chang, "Automatic spectral target recognition in hyperspectral imagery," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1232–1249, 2003.

[14] X. Zhang, X. Tong, and M. Liu, "An improved n-findr algorithm for endmember extraction in hyperspectral imagery," in *2009 Joint Urban Remote Sensing Event*. IEEE, 2009, pp. 1–5.

[15] P. Sellars, A.I. Aviles-Rivero, and C.B. Schoenlieb, "Superpixel contracted graph-based learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4180–4193, Jun 2020.

[16] M. Maggioni and J.M. Murphy, "Learning by unsupervised nonlinear diffusion.," *Journal of Machine Learning Research*, vol. 20, no. 160, pp. 1–56, 2019.

[17] D.A. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, vol. 741, 2009.

[18] R. Bro and S. Jong, "A fast non-negativity-constrained least squares algorithm," *J. Chemometrics*, vol. 11, no. 5, 1997.

[19] C. Jayaprakash, B.B. Damodaran, S. Viswanathan, and K.P. Soman, "Randomized independent component analysis and linear discriminant analysis dimensionality reduction methods for hyperspectral image classification," *Journal of Applied Remote Sensing*, vol. 14, no. 3, pp. 036507, 2020.