UNSUPERVISED DISCRIMINATIVE DIMENSION REDUCTION FOR HYPERSPECTRAL CHEMICAL PLUME SEGMENTATION

James M. Murphy [1], *Mauro Maggioni* [2]
[1]Tufts University, Department of Mathematics, Medford, MA 02155, USA; [2]Johns Hopkins University, Department of Mathematics, Mathematical Institute of Data Sciences, Baltimore, MD 21218, USA

ABSTRACT

We propose a novel algorithm for unsupervised segmentation of hyperspectral imagery (HSI). Representative cluster modes are learned through the diffusion geometry of the HSI, which is highly invariant to nonlinearities present in HSI clusters. Mode detection is followed by partial least squares regression to project the data onto a low-dimensional space that discriminates between the learned modes and to assign labels in the low-dimensional space. We evaluate this method for unsupervised chemical plume segmentation in HSI, showing it performs competitively versus benchmark and state-of-the-art unsupervised learning techniques.

Index Terms— Hyperspectral images, machine learning, unsupervised learning, image segmentation, diffusion geometry, spectral graph theory

1 INTRODUCTION

As the volume of data collected by remote sensors continues to grow, human capacity for generating labeled training sets is outpaced by the sheer volume of data. Many state-of-the-art machine learning methods for hyperspectral images (HSI) are based on support vector machines [1] or deep neural networks [2], both of which are supervised and require large training data sets. In contexts where the creation of training sets is impractical, new, unsupervised methods are required.

We propose a new method for unsupervised segmentation of HSI, consisting of two steps. First, modes—high-density representative elements of distinct classes—of the high-dimensional HSI are learned through *diffusion geometry*. Then, *partial least squares regression (PLSR)* is performed to identify a low-dimensional subspace that discriminates between the learned modes. Labels are assigned in this discriminative low-dimensional space. After reviewing some background in Sec.2, we describe the method in Sec.3 and evaluate it in Sec.4 before concluding in Sec. 5.

This research was partially supported by NSF-ATD-1737984, AFOSR FA9550-17-1-0280, and NSF-IIS-1546392.

2 BACKGROUND

The goal of unsupervised HSI segmentation is to provide an HSI dataset $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ with labels $\{y_i\}_{i=1}^n$, $y_i \in \{1,\dots,K\}$, without access to any labeled training pairs (x_i, y_i). The number of clusters K may be unknown a priori and may need to be estimated. A range of techniques for this task has been proposed, including centroid-based methods, density methods, and graph theoretic methods [3]. Classical approaches may fail to accurately cluster HSI due to high dimensionality, nonlinear geometry, and low signal-to-noise level. Fortunately, HSI typically consist of classes that are parametrized, often in a nonlinear fashion, by a small number latent variables. Methods which exploit this intrinsic geometry improve over methods based simply on Euclidean distances, which may be insufficient to capture underlying nonlinearities [4, 5].

Diffusion distances [6, 7] have been proposed as a metric for HSI that respects the underlying data geometry [4]. Consider the HSI data as a point cloud $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$, where n is the number of pixels and D the number of spectral bands. Let $\mathcal{G}$ be a graph with vertices corresponding to $\{x_i\}_{i=1}^n$ and edges stored in a (symmetric) weight matrix $W_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$ if x_i is among the k-nearest neighbors of x_j or vice versa and σ is a scaling parameter. Let $P = D^{-1}W$ be a corresponding Markov transition matrix on the data, with D the diagonal degree matrix where $D_{ii} = \sum_{j=1}^n W_{ij}$. Diffusion distances are derived from P as $d_t(x_i,x_j) = \sqrt{\sum_{\ell=1}^n (P(i,\ell) - P(j,\ell))^2}$. P admits an eigendecomposition $\{(\lambda_i, \phi_i)\}_{i=1}^n$ which may be used to compute $d_t(x_i,x_j) = \sqrt{\sum_{\ell=1}^n \lambda_\ell^{2t}(\phi_\ell(i) - \phi_\ell(j))^2}$. Diffusion distances capture the geometry as learned by the random

diffusion process P: $d_t(x_i,x_j)$ is small for points x_i, x_j well-connected according to P^t, and large otherwise.

Diffusion Distances for HSI Analysis. Recently, diffusion distances have been combined with density estimation to label HSI with high empirical accuracy and provable performance guarantees [4, 8, 9, 10]. The *learning by unsupervised nonlinear diffusion (LUND)* algorithm first learns modes $\{x_i^*\}_{i=1}^K$ in data $X = \{x_i\}_{i=1}^n$ by finding high density points, as quantified by a kernel density estimator $p(x)$, that are far in diffusion distance from other high density points, as quantified by

$$\rho_t(x_i) = \begin{cases} \min\limits_{\{p(x_j)\geq p(x_i)\}} d_t(x_i,x_j), & x_i \neq \arg\max_\ell p(x_\ell), \\ \max_{x_j} d_t(x_i,x_j), & x_i = \arg\max_\ell p(x_\ell); \end{cases}$$

see Algorithm 1. LUND then orders HSI pixels by density and assigns them sequentially to their d_t-nearest labeled point of higher density: see [4] for details and a discussion on how K may be automatically learned.

Algorithm 1: Mode Detection Algorithm

Input: $X, K; t$.

1: Compute $\{p(x_i)\}_{i=1}^n$, $\{\rho_t(x_i)\}_{i=1}^n$.

2: Compute $\{x_i^*\}_{i=1}^K$, the K maximizers of $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$.

Output: $\{x_i^*\}_{i=1}^K, \{p(x_i)\}_{i=1}^n, \{\rho_t(x_i)\}_{i=1}^n$.

3 PROPOSED HSI SEGMENTATION ALGORITHM

Motivated by the success of PLSR for anomaly detection in hyperspectral movies [11], we propose to incorporate the learned diffusion modes into PLSR. PLSR is a *supervised* linear dimension reduction technique that reduces the dimension in such a way that not just variance (as in principal component analysis), but also discrimination between training data of different classes, is maximized in the low-dimensional embedding. To make use of PLSR in the unsupervised setting, we associate the modes learned with diffusion distances to a *core* C_i of points, consisting of the k_1-nearest neighbors of a mode x_i^* in diffusion distance[1]. Note that by construction C_i^* is far from C_j^*, $i \neq j$, since these are the points near the modes of separate classes. Each core is a set of points that, with high confidence, should have the same class label: we assign to each an arbitrary label, and then use these labels of the cores $\{C_i^*\}_{i=1}^K$ for PLSR. To map these cores to numerical labels needed by PLSR and in order to avoid imposing an artificial one-dimensional structure on the labels, we assign the label of C_i^* to the vertex of a K-simplex, and perform vector-valued regression with PLSR. In formulas: the points in C_i^* are labeled as $\mathbb{1}_i = (0,0,\ldots,0,1,0,\ldots 0)$, with 1 in the i^{th} coordinate. This generates a set of training pairs for the i^{th} class: $T_i = \{(x,y) \mid x \in C_i^*, y = \mathbb{1}_i\}$.

The data X is pre-processed so that the columns of X, the $n \times D$ matrix representing the data, have mean 0. This training data $T = \bigcup_{i=1}^K T_i$ is used in PLSR to acquire a regression coefficient $\hat{\beta}$. The regression coefficient $\hat{\beta}$ is a $D \times K$ matrix that is derived only from the learned cores. Multivariate regression on all of X is performed by computing $\hat{Y} = X\hat{\beta}$; $\hat{Y}$ is an $n \times K$ matrix with rows corresponding to points and columns corresponding to responses of each point. The columns of $\hat{Y}$ may be interpreted as affinity for each class. Class labels are learned as $\hat{y}_i = \arg\max_{k=1,\ldots,K} \hat{Y}_{i,k}$; see Algorithm 2.

Algorithm 2: HSI Segmentation with Diffusion Cores and PLSR

1 *Input*: $X, K; t$.

2 Learn modes $\{x_i^*\}_{i=1}^K$ of X by Algorithm 1.

3 Generate, from each x_i^*, a *core* C_i^* by computing the k_1-nearest neighbors of x_i^* in diffusion distance.

4 Learn PLSR coefficient $\hat{\beta}$ using $\{(C_i^*, \mathbb{1}_i)\}_{i=1}^K$ as training data (e.g., using the SIMPLS algorithm [12]).

5 Regress Y on X as $\hat{Y} = X\hat{\beta}$.

6 Assign labels to all points: $\hat{y}_i = \arg\max_{k=1,\ldots,K} \hat{Y}_{i,k}$.

7 *Output:* $\{\hat{y}_i\}_{i=1}^n$.

Computational Complexity. The total cost of Algorithm 1 is $O(C_d Dn \log n)$ [4], where d is the intrinsic dimension of the data (usually < 10 in HSI), and C_d is exponential in d. PLSR is implemented with the SIMPLS algorithm [12], which has computational complexity $O(Dnm + Dm^2 + m^3)$, where m is the number of partial least squares components used. In our algorithm, this is equal to the number of classes, either known or determined by studying the decay of the sorted $\mathcal{D}_t(x_i)$ curve [4] , and may be assumed to be $O(1)$ and indepen-

[1] On all the data sets we considered, the method is robust to the choice of k_1; in what follows we will use $k_1 = \lceil .02 \cdot n \rceil$.

dent of n, D. Thus, the overall complexity of the proposed algorithm is near linearly in n: $O(C_d D n \log n)$.

4 EXPERIMENTAL RESULTS

Experimental Data. An HSI capturing a faint chemical plume released into an otherwise homogeneous background, courtesy of the Johns Hopkins Applied Physics Lab (APL), appears in Fig. 1.

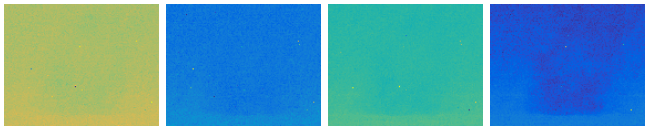

Fig. 1. A chemical plume HSI dataset of spatial dimensions 110×140. There are $n = 110 \cdot 140 = 15400$ pixels and $D = 129$ spectral bands. The bands shown (11,16, 18, 29) indicate the visibility of the chemical plume in certain bands, but not in others.

The goal is to efficiently segment the plume without supervision, and without using spatial information nor the fact that the camera, in this particular data set, is still (and so is the background). Several pixels are highly corrupted by noise. No labeled ground truth is available for this image, so we evaluate with only visual quality of the image segmentation. In order to determine the number of clusters, K, used by our clustering algorithms, we examine the plot of sorted $\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)$ values. We look for "kinks" in this plot, which correspond to a major drop in $\mathcal{D}_t(x_i)$ values (see Fig. 2). Under a flexible non-parametric data model, this heuristic for estimating the number of clusters is provably accurate [10].

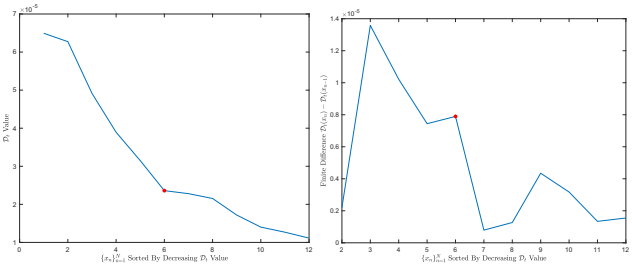

Fig. 2. Plot of sorted $\mathcal{D}_t(x_i)$ values for APL chemical plume HSI and a plot of the finite differences between successive values. The extrema of the difference curve were used to estimate the number of clusters $K = 6$ to use for plume segmentation.

Results. We consider a variety of benchmark and state-of-the-art methods of HSI clustering for comparison, see Fig.3. Each method is run on the data with $K = 6$ clusters; other parameters were chosen to optimize visual performance. Only the proposed method achieves reasonable plume segmentation (see Fig.3). PLSR with cores learned with Euclidean distances is unable to correctly segment horizontally, resulting in a plume that is spread too far. However, it correctly ascertains that the plume diffuses somewhat far in the vertical direction. PLSR with diffusion cores correctly segments in both the horizontal and vertical directions, and gives the best visual result. Several of the other methods were unable to detect anomalous pixels effectively, resulting in clusters of very small sizes. Spectral clustering and SMCE reasonably segment the bottom half of the plume, but fail completely for the top half.

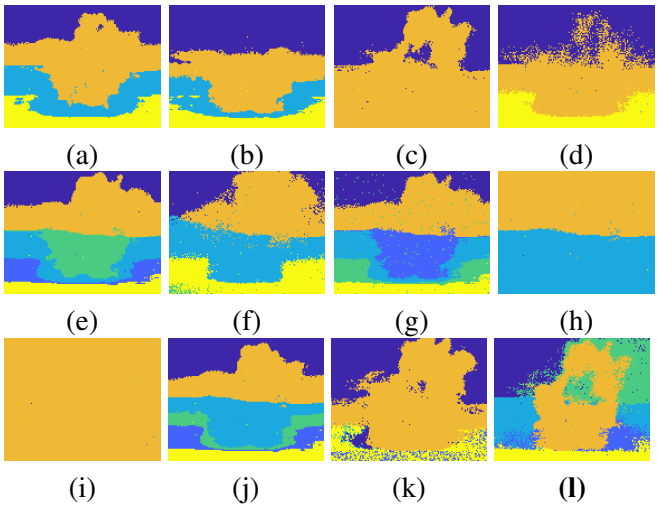

Fig. 3. Results for APL chemical plume dataset. Comparison method are: (a) K-means; (b) K-means combined with principal component analysis (PCA); (c) K-means combined with independent component analysis (ICA) [13]; (d) K-means combined with Gaussian random projections [14]; (e) spectral clustering [15]; (f) Gaussian mixture models (GMM) [16]; (g) sparse manifold clustering and embedding (SMCE) [17]; (h) hierarchical clustering with non-negative matrix factorization (HNMF) [18]; (i) fast search and find of density peaks clustering (FSFDPC) algorithm [19]; (j) LUND [4]; (k) proposed method with modes learned with Euclidean distance; (l) proposed method.

5 CONCLUSIONS & FUTURE WORK

Using diffusion geometry yields robustness to different data geometries [10], but the subsequent PLSR *linear* projection places statistical assumptions on the data that may not always hold. In order to improve robustness to more complicated data geometries, developing a fully *nonlinear partial least squares regression method* is of interest. This approach has some similarities with the Schrödinger eigenmaps approach [20], though Schrödinger eigenmaps separates only a prior-

itized class (often thought of as a background class), while we propose to develop a method that can handle many classes, possibly of very different sizes.

6 References

[1] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[2] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *J. Sel. Topics Appl. Earth Observ.*, vol. 7, no. 6, pp. 2094–2107, 2014.

[3] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[4] J.M. Murphy and M. Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1829–1845, 2019.

[5] J.M. Murphy and M. Maggioni, "Spectral-spatial diffusion geometry for hyperspectral image clustering," *arXiv preprint arXiv:1902.05402*, 2019.

[6] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 21, pp. 7426–7431, 2005.

[7] R.R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[8] J.M. Murphy and M. Maggioni, "Iterative active learning with diffusion geometry for hyperspectral images," in *9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2018, To Appear.

[9] J.M. Murphy and M. Maggioni, "Diffusion geometric methods for fusion of remotely sensed data," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIV*. International Society for Optics and Photonics, 2018, vol. 10644, p. 106440I.

[10] M. Maggioni and J.M. Murphy, "Learning by unsupervised nonlinear diffusion," *arXiv preprint arXiv:1810.06702*, 2018.

[11] Y. Wang, G. Chen, and M. Maggioni, "High-dimensional data modeling techniques for detection of chemical plumes and anomalies in hyperspectral images and movies," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 9, no. 9, pp. 4316–4324, 2016.

[12] S. De Jong, "SIMPLS: an alternative approach to partial least squares regression," *Chemom. Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993.

[13] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.

[14] S. Dasgupta, "Experiments with random projection," in *UAI*. Morgan Kaufmann Publishers Inc., 2000, pp. 143–151.

[15] A.Y. Ng, M.I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *NIPS*, 2001, vol. 14, pp. 849–856.

[16] N. Acito, G. Corsini, and M. Diani, "An unsupervised algorithm for hyperspectral image segmentation based on the gaussian mixture model," in *IEEE IGARSS*, 2003, vol. 6, pp. 3745–3747.

[17] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[18] N. Gillis, D. Kuang, and H. Park, "Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2066–2078, 2015.

[19] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[20] W. Czaja and M. Ehler, "Schroedinger eigenmaps for the analysis of biomedical data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1274–1280, 2013.