Diffusion Processes for Hyperspectral Data: Clustering and Active Learning

James M. Murphy (jm.murphy@tufts.edu)

Tufts University

Unsupervised and Active Learning In Remote Sensing

Hyperspectral imagery (HSI) is a significant data source in remote sensing. The large data size of HSI and their high dimensionality demand efficient machine learning algorithms to automatically process and glean insight from the deluge of hyperspectral data now available.

The process of labelling pixels "by hand" is costly and requires a human expert, motivating machine learning techniques that require little or no labelled training data. Methods of HSI **clustering**, or **unsupervised segmentation** label HSI with no training data. This is considerably more challenging than traditional classification, and is mathematically ill-posed without statistical or geometric assumptions on the data. **Active learning** is a supervised technique where a small, automatically but carefully chosen set of pixels is labelled, as opposed to the standard supervised learning setting in which the labels are random. Active learning can lead to high quality classification results with a very small number of labelled training samples.

Major challenges in machine learning for HSI include:

1. The **high dimensionality** of the data, with some HSI exceeding 200 spectral bands.
2. Spectral clusters in HSI are typically **nonlinear**, rendering linear methods ineffective.
3. There is often **significant noise and between-cluster overlap** among HSI classes, due to the materials being imaged and poor sensing conditions.
4. HSI images may be very **large in size**, requiring methods to scale quasilinearly in the number of pixels.

We propose to overcome these challenges by **combining density-based methods with geometric learning through diffusion geometry** [1; 2] in order to identify class modes. This information is then propagated to all data points through a nonlinear process that incorporates both spectral and spatial information. The use of data-dependent diffusion maps for mode detection offers significant empirical advantages and enjoys robust theoretical performance guarantees [3]. Diffusion distances exploit low-dimensional structures in the data, allowing the proposed method to handle data that is high-dimensional but intrinsically low-dimensional, even when nonlinear and noisy. Moreover, we propose a spectral-spatial labelling scheme which takes advantage of the geometric properties of the data to improve the empirical performance of clustering when compared to labelling based on spectral information alone [4; 5]. In addition, the proposed unsupervised method assigns to each data point a measure of confidence for the unsupervised label assignment. This leads naturally to an active learning algorithm in which points with low confidence scores are queried for training labels, which then propagate through the remaining data.

Diffusion Processes on Graphs

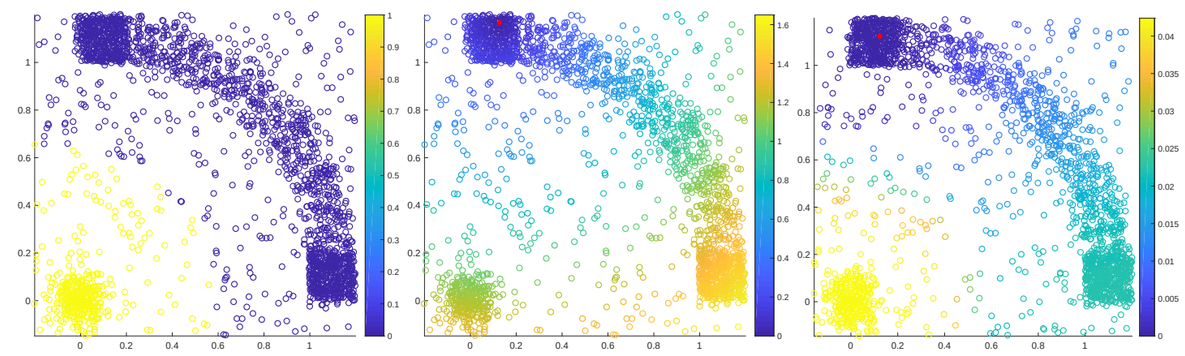


Figure 1: Data drawn from two distributions: μ_1 is a mixture of two isotropic Gaussians with means $(0,1)$ and $(1,0)$ connected by a parabolic shape; μ_2 is an isotropic Gaussian with mean $(0,0)$. Uniform background noise is added. *Left*: The data plotted and colored by cluster. *Middle*: The distances from the point $(0,1)$—colored red—in the Euclidean distance. *Right*: The distances from the point $(0,1)$ in diffusion distances. The parabolic segment bridges the two Gaussians and causes the high density regions near $(0,1)$ and $(1,0)$ to be closer in diffusion distance than they are in Euclidean distance, due to the introduction of many paths with short edges connecting the high density regions across this bridge.

Let $X = \{x_n\}_{n=1}^N \subset \mathbb{R}^D$ be discrete. The computation of the **diffusion distance** d_t [1; 2] constructs a weighted, undirected graph $\mathcal{G}$ with vertices corresponding to the points in X and weighted edges given by the $N \times N$ weight matrix $W(x,y) := e^{-\|x-y\|_2^2/\sigma^2}, x \in NN_k(y)$ and $W(x,y) = 0$ otherwise, for some σ, k, where $NN_k(x)$ is the set of k-nearest neighbors of y in X with respect to Euclidean distance. Let $P(x,y) = W(x,y)/\deg(x)$ be an $N \times N$ Markov transition matrix, where $\deg(x) := \sum_{y \in X} W(x,y)$ is the degree of x. For an initial distribution $\mu \in \mathbb{R}^N$ on X, the vector μP^t is the probability over states at time $t \geq 0$. As t increases, this diffusion process on X evolves according to the connections between the points encoded by P.

The diffusion distance at time t is $d_t^2(x,y) := \sum_{u \in X} (P^t(x,u) - P^t(y,u))^2 d\mu(u)/\pi(u)$, for $\pi P = \pi$ the stationary distribution. P has eigenvectors $\{\Phi_n\}_{n=1}^N$ and eigenvalues $1 = \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$, whence

$$d_t^2(x,y) = \sum_{n=1}^N \lambda_n^{2t} (\Phi_n(x) - \Phi_n(y))^2 \,. \tag{1}$$

Diffusion distances are parametrized by t, which measures how long the diffusion process on $\mathcal{G}$ has evolved. If $\mathcal{G}$ is connected, $|\lambda_n| < 1$ for $n > 1$. Hence, (1) may approximated by truncating at some suitable $2 \leq M \ll N$. The truncation simultaneously denoises and reduces computation by requiring only a few eigenvectors.

Algorithm Description: Spectral-Spatial Diffusion Learning (DLSS)

The first part of the proposed clustering [4; 5] and active learning algorithms is to **learn K modes of the data**, corresponding to unique clusters. This is detailed in Algorithm 1.

Algorithm 1: Geometric Mode Detection Algorithm

Input: X, K

1: Compute the empirical density $p(x_n)$ for each $x_n \in X$.
2: Compute $\{\rho_t(x_n)\}_{n=1}^N$, the diffusion distance from each point to its nearest neighbor in diffusion distance of higher empirical density, normalized.
3: Set the learned modes $\{x_i^*\}_{i=1}^K$ to be the K maximizers of $\mathcal{D}_t(x_n) := p(x_n)\rho_t(x_n)$.

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

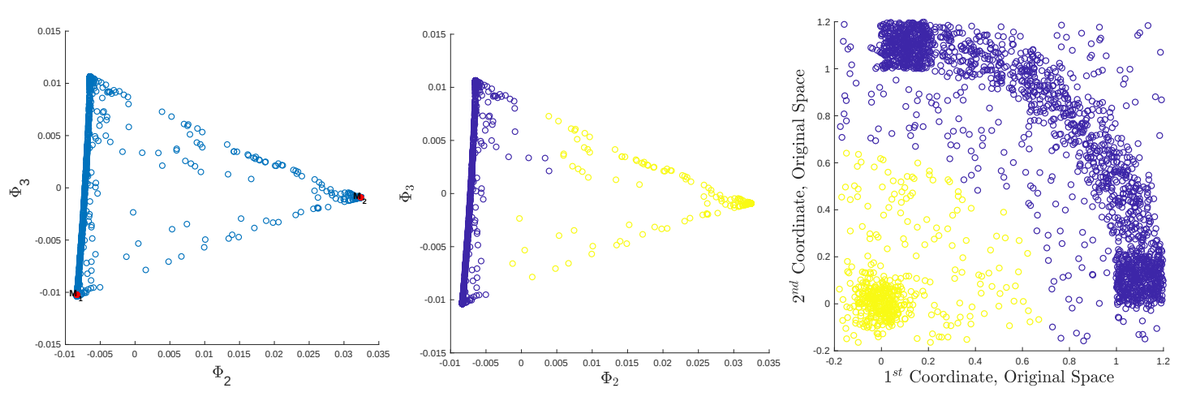


Figure 2: *Left*: Data from Figure 1 represented in the new coordinates given by the second and third eigenfunctions of P. In this coordinate system, the natural Euclidean distance is equal to the diffusion distance on the original image. The learned modes are computed in this low-dimensional embedding, as described in Algorithm 1. *Middle*: Clusters in the new coordinate system. *Right*: Points are labelled according to diffusion distances and the learned modes, as described in Algorithm 2.

Once the modes are computed, remaining points are labelled in a **joint spectral-spatial iterative procedure**, as described in Algorithm 2. A crucial notion is that of consensus spatial label, which is essentially the most common label among the spatial nearest neighbors [5]. An example of the two-stage labelling procedure appears in Figure 3. Notice that not all points are labelled in the first stage, only those near the spectral modes. In the second stage, remaining points are labelled using spatial information.

Algorithm 2: Spectral-Spatial Labelling Algorithm:

Input: $\{x_i^*\}_{i=1}^K, \{p(x_n)\}_{n=1}^N, \{\rho_t(x_n)\}_{n=1}^N$

1: Assign each mode a unique label.
2: Iterating in order of decreasing $p(x)$ among unlabelled points, assign each point the label of its nearest spectral d_t-neighbor of higher density, unless the spatial consensus label exists and differs, in which case the point is left unlabelled.
3: Iterating in order of decreasing $p(x)$ among unlabelled points, assign each point the consensus spatial label, if it exists, otherwise the same label as its nearest spectral d_t-neighbor of higher density.

Output: Labels $\{y_n\}_{n=1}^N$.

Figure 3: *Left:* First principal component of Indian Pines subset. *Second from left:* Ground truth labels. *Third from left:* The partial labelling from the first stage (Algorithm 2, (2)). After mode identification, points are labelled with the same label as their nearest spectral neighbor of higher density, unless that label differs from the consensus label in the spatial domain, in which case a point is left unlabelled. This leads to points far from the centers of the classes staying unlabelled after the first stage. *Right:* Results from second, final stage clustering. In the second stage (Algorithm 2, (3)), unlabelled points are assigned labels by the same rule, unless there is a clear consensus in the spatial domain, in which case the unlabelled point is given the consensus spatial label.

Algorithm 2 is called **spectral-spatial diffusion learning (DLSS)**, while the variant without spatial information being incorporated is called **diffusion learning (DL)**.

Experimental Analysis and Conclusions

Numerical results appear in Figure 4; images for the results of DL and DLSS on Salinas A are in Figure 5.

Method	OA I.P.	AA I.P.	κ I.P.	OA P.	AA P.	κ P.	OA S.A.	AA S.A.	κ S.A.	OA K.S.C.	AA K.S.C.	κ K.S.C.
SMCE	0.52	0.45	0.22	0.83	0.77	0.79	0.47	0.42	0.30	0.36	0.26	0.01
HNMF	0.41	0.32	-0.02	0.72	0.74	0.66	0.63	0.66	0.53	0.36	0.25	0.00
FMS	0.57	0.50	0.27	0.77	0.64	0.69	0.70	0.81	0.65	0.74	0.70	0.65
FSFDPC	0.58	0.51	0.26	0.78	0.75	0.73	0.63	0.61	0.54	0.36	0.25	0.00
DL	0.67	0.62	0.44	0.85	0.78	0.81	0.83	0.88	0.79	0.81	0.72	0.74
DLSS	**0.85**	**0.82**	**0.75**	**0.94**	**0.83**	**0.93**	**0.85**	**0.90**	**0.81**	**0.83**	**0.73**	**0.76**

Figure 4: Overall accuracy (OA), average accuracy (AA), and Cohen's κ for real HSI clustering experiments on subsets of Indian Pines (I.P.), Pavia (P.), Salinas A (S.A.), and Kennedy Space Center (K.S.C.). Best results are in bold, second best are underlined. Experiments were performed for the proposed unsupervised methods (DL, DLSS) and compared to a range of benchmark and state-of-the-art clustering algorithms, including sparse manifold clustering and embedding (SMCE); hierarchical non-negative matrix factorization (HNMF); fast Mumford-Shah segmentation (FMS); and fast search and find of density peaks clustering (FSFDPC). Results against other methods are reported in the journal article for this research [5].

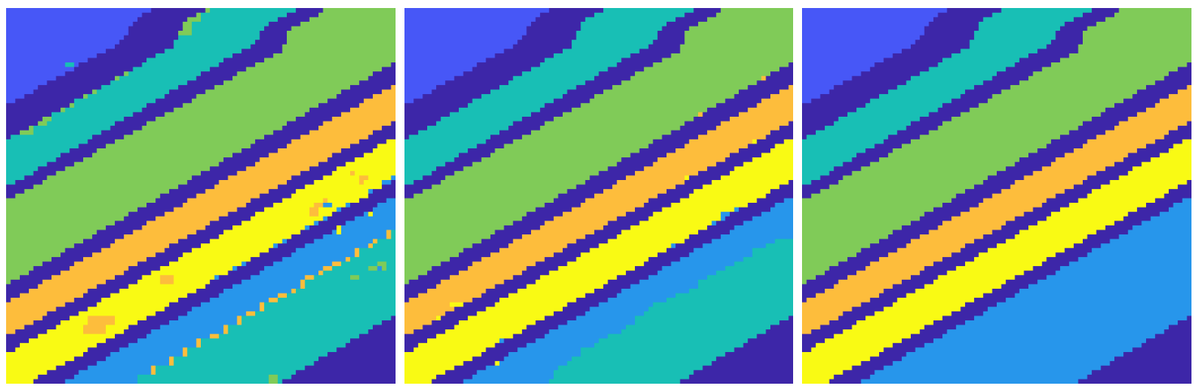


Figure 5: Clustering results for Salinas A. *Left*: DL results; *Middle:* DLSS results; *Right:* ground truth.

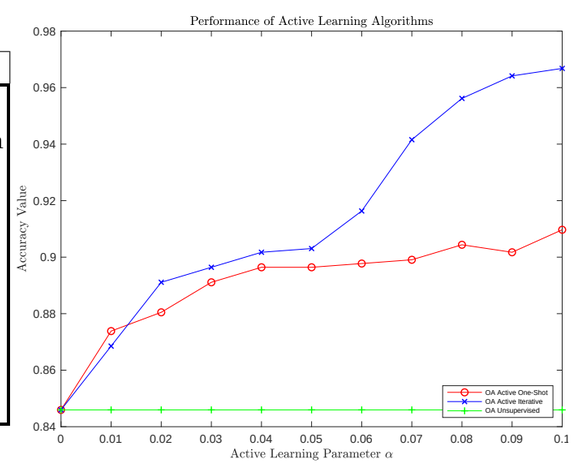
Active learning incorporates L labelled pixels into the proposed method, computed as the points whose d_t-nearest modes are most ambiguous. More precisely, we fix a time t and for each pixel x_n, let $x_{n_1}^*, x_{n_2}^*$ be the two modes d_t-nearest to x_n. We compute the quantity $F_t(x_n) = |d_t(x_n, x_{n_1}^*) - d_t(x_n, x_{n_2}^*)|$ [6]. See Algorithm 3 for details. The proposed active method is compared to a one-shot, non-iterative variant of the method as well as the fully unsupervised DLSS algorithm for the Indian Pines dataset below. For the Indian Pines dataset, the iterative method outperforms the one-shot method as α increases, with the gap increasing substantially as $\alpha = L/N$ approaches 10^{-1}. Both outperform purely unsupervised learning.

Algorithm 3: Iterative Active Diffusion Learning

Input: X, K, L.

1: Compute the modes of the data using Algorithm 1 and assign each a unique label.
2: **while** $L > 0$ **do**
3: Compute, for all x_n, $F_t(x_n)$.
4: Label the minimizer of F_t with ground truth.
5: $L = L - 1$.
6: **end while**
7: Proceed as in (2), (3) of Algorithm 2

Output: Labels $\{y_n\}_{n=1}^N$



In terms of **computational complexity**, the DL and DLSS methods scale as $O(C_d DN \log N + k_1 DN) \approx O(N \log(N))$ with C_d a constant that depends exponentially on the intrinsic dimension d of the data. All code is available on the authors' website (`https://jmurphy.math.tufts.edu/Code/`).

References

[1] 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.*, 102(21):7426–7431, 2005.
[2] R.R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.
[3] M. Maggioni and J.M. Murphy. Learning by unsupervised nonlinear diffusion. *arXiv:1810.06702*, 2018.
[4] J.M. Murphy and M. Maggioni. Diffusion geometric methods for fusion of remotely sensed data. In *SPIE Defense+Security*, volume 10644, page 106440I, 2018.
[5] J. Murphy and M. Maggioni. Nonlinear unsupervised clustering and active learning of hyperspectral images. *IEEE Trans. Geosci. Remote Sens.*, 2018. DOI: 10.1109/TGRS.2018.2869723.
[6] J.M. Murphy and M. Maggioni. Iterative active learning with diffusion geometry for hyperspectral images. In *IEEE WHISPERS*, 2018. To Appear.