# MULTISCALE CLUSTERING OF HYPERSPECTRAL IMAGES THROUGH SPECTRAL-SPATIAL DIFFUSION GEOMETRY

*Sam L. Polk and James M. Murphy*

Department of Mathematics, Tufts University, 503 Boston Ave, Medford MA, 02155, USA.

## ABSTRACT

Clustering algorithms partition a dataset into groups of similar points. The primary contribution of this article is the *Multiscale Spatially-Regularized Diffusion Learning (M-SRDL)* clustering algorithm, which uses spatially-regularized diffusion distances to efficiently and accurately learn multiple scales of latent structure in hyperspectral images (HSI). The M-SRDL clustering algorithm extracts clusterings at many scales from an HSI and outputs these clusterings' variation of information-barycenter as an exemplar for all underlying cluster structure. We show that incorporating spatial regularization into a multiscale clustering framework corresponds to smoother and more coherent clusters when applied to HSI data and leads to more accurate clustering labels.

*Index Terms*— Clustering, Diffusion Geometry, Hierarchical Clustering, Hyperspectral Imagery, Unsupervised Machine Learning

## 1. INTRODUCTION

Hyperspectral images (HSIs) are datasets storing reflectance at many electromagnetic bands. HSIs provide a rich characterization of a scene, enabling the precise discrimination of materials based on variations within pixels' spectral signatures. Indeed, success at material discrimination using HSIs has led to hyperspectral imagery's emergence as an important data source in remote sensing [1]. However, HSIs are generated in large quantities, making manual analysis is infeasible. Moreover, HSIs are very high-dimensional and typically encode multiple scales of latent structure, ranging from coarse to fine in scale [2, 3, 4, 5]. Thus, efficient algorithms are needed to automatically learn multiscale structure from HSIs.

The main contribution of this article is the Multiscale Spatially-Regularized Diffusion Learning (M-SRDL) algorithm, which is a multiscale extension of the Spatially-Regularized Diffusion Learning (SRDL) algorithm [6]. SRDL uses spatially-regularized diffusion distances to efficiently extract a fixed scale of cluster structure from an HSI [6]. To learn multiscale structure from an HSI, M-SRDL varies a diffusion time parameter in SRDL [2]. M-SRDL then finds the variation of information (VI)-barycenter of the extracted clusterings: the partition of the HSI that best represents all scales of learned cluster structure [2, 7]. In this sense, M-SRDL suggests multiple clusterings at different scales in addition to the one that best represents all underlying multiscale structure.

We organize this article in the following way. Background on unsupervised learning, diffusion geometry, and spatial regularization is provided in Section 2. M-SRDL is introduced in Section 3. Numerical experiments are presented in Section 4, and conclusions are given in Section 5.

## 2. BACKGROUND

A clustering algorithm partitions an HSI $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ (interpreted as a point cloud, where $n$ is the number of pixels and $D$ is the number of spectral bands) into *clusters* of data points, denoted $X_1, X_2, \ldots, X_K$ [8]. The partition $\mathcal{C} = \{X_k\}_{k=1}^K$ is called a *clustering* of $X$, and each pixel may be assigned a unique label corresponding to its $X_k$. In a good clustering of $X$, data from the same cluster are "similar," while data from different clusters are "dissimilar." The specific notion of similarity between points varies across the many clustering algorithms in the literature [3, 4, 5, 8, 9, 10].

### 2.1. Background on Diffusion Geometry

Data-dependent, diffusion-based mappings are highly effective at capturing an HSI's intrinsic low-dimensionality [11, 12]. These methods treat data points as nodes in an undirected graph. Edges between points are encoded in a weight matrix: $W_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$, where $\sigma$ is a scale parameter reflecting the interaction radius between points. Define the Markov *transition matrix* for a random walk on $X$ by $P = D^{-1}W$, where $D_{ii} = \sum_{j=1}^n W_{ij}$. Assuming $P \in \mathbb{R}^{n \times n}$ is reversible, aperiodic, and irreducible, there exists a unique $q \in \mathbb{R}^{1 \times n}$ satisfying $qP = q$. Denote (right) eigenvector-eigenvalue pairs of $P$ by $\{(\psi_i, \lambda_i)\}_{i=1}^n$, sorted so that $1 = |\lambda_1| > |\lambda_2| \geq \ldots |\lambda_n| \geq 0$. The first $K$ eigenvectors of $P$ tend to concentrate on the $K$ most highly-connected components of the underlying graph, making them useful for clustering.

*Diffusion distances*, defined for $x_i, x_j \in X$ at time $t \geq 0$ by $D_t(x_i, x_j) = \sqrt{\sum_{k=1}^n |(P^t)_{ik} - (P^t)_{jk}|^2/q_k}$ [11, 12],

are a data-dependent distance metric well-suited to the clustering problem [2, 9]. If clusters are well-separated and coherent, within-cluster diffusion distances will be small relative to between-cluster diffusion distances [9]. The *diffusion map* $\Psi_t(x) = (\psi_1(x), |\lambda_2|^t\psi_2(x), \ldots, |\lambda_n|^t\psi_n(x))$ is a closely related concept [11, 12]. Indeed, diffusion distances can be identified as Euclidean distances in a data-dependent feature space consisting of diffusion map coordinates: $D_t(x,y) = \|\Psi_t(x) - \Psi_t(y)\|_2$ [11].

The time parameter $t$ is linked to the scale of structure that can be separated using diffusion distances [2, 11, 12]. Small $t$ enables the detection of fine-scale local structure, while larger $t$ enables the detection of macro-scale global structures. Thus, the diffusion time parameter may be varied to learn the rich multiscale structure encoded in HSI data [2].

## 2.2. The SRDL Clustering Algorithm

Nearby HSI pixels tend to come from the same cluster. Thus, incorporating spatial geometry (where pixels are located within the HSI) into a clustering algorithm may result in smoother clusters [6, 13, 14]. The SRDL algorithm (Algorithm 1) modifies the graph underlying $P$ to directly incorporate spatial geometry into the diffusion process [6]. More precisely, each $x$ is connected to its $N$ $\ell^2$-nearest neighbors, chosen from pixels within a $(2R+1) \times (2R+1)$ spatial square centered at $x$ in the original image, where $R \in \mathbb{N}$ is called the *spatial window*. Thus, nearby pixels may have edges between them, but pixels from different segments of the image will not. Diffusion distances computed using this modified graph are called *spatially-regularized diffusion distances* [6].

Fix $t \geq 0$ so that there is a single scale of latent cluster structure to be learned. Let $p(x)$ be a kernel density estimator (KDE) capturing empirical density:

$$p(x) = \frac{1}{Z} \sum_{y \in NN_N(x)} \exp\left( -\frac{\|x-y\|_2^2}{\sigma_0^2} \right),$$

where $NN_N(x)$ is the set of the $N$ points in $X$ closest to $x$ in $\ell^2$-distance, $\sigma_0$ is a KDE bandwidth, and $Z$ is a normalization constant so that $\sum_{x \in X} p(x) = 1$. Let $\rho_t(x)$ be defined as

$$\rho_t(x) = \begin{cases} \max\limits_{y \in X} D_t(x,y) & x = \operatorname{argmax}\limits_{y \in X} p(y), \\ \min\limits_{y \in X}\{D_t(x,y) | p(y) \geq p(x)\} & \text{otherwise.} \end{cases}$$

Thus, $\rho_t(x)$ returns the diffusion distance at time $t$ between $x$ and its $D_t$-nearest neighbor of higher density, capturing diffusion geometry underlying the HSI. Maximizers of $\mathcal{D}_t(x) := p(x)\rho_t(x)$ are, thus, points with large $p$-values (i.e., high density) and large $\rho_t$-values (i.e., high diffusion distances from any other high-density points). These maximizers are defined as *data modes* and are assigned unique cluster labels.

SRDL labels non-modal points in two stages. In the Labeling Stage 1, for each non-modal point $x \in X$, SRDL finds

---

**Algorithm 1:** Spatially-Regularized Diffusion Learning (SRDL) clustering algorithm [6]

**Input:** $X$ (data), $t$ (time step), $N$ (no. nearest neighbors) $\sigma$ (scale parameter), $\sigma_0$ (KDE bandwidth), $R$ (spatial window)

**Output:** $\mathcal{C}$ (clustering), $K$ (no. clusters)

1 Build KNN graph with $N$ $\ell^2$-nearest neighbors within a spatial window $R > 0$. Weight edges according to a Gaussian kernel with scale parameter $\sigma$. Compute $P$ from this graph;

2 Compute $\mathcal{D}_t(x) = p(x)\rho_t(x)$, where $p(x)$ and $\rho_t(x)$ are computed as described in Section 2.2;

3 Solve $K = \operatorname{argmin}_{1 \leq k \leq n-1} \mathcal{D}_t(x_{m_k})/\mathcal{D}_t(x_{m_{k+1}})$, where $\{x_{m_k}\}_{k=1}^n$ is a sorting of the points in $X$ according to $\mathcal{D}_t(x)$ in non-increasing order. Label data modes $\mathcal{C}(x_{m_k}) = k$ for $k = 1, \ldots, K$;

4 *Labeling Stage 1*: In order of non-increasing $p(x)$, for each $x \in X$, find $x^*$ and $\mathcal{C}^{(s)}(x)$ (if it exists), as described in Section 2.2. If $\mathcal{C}^{(s)}(x)$ exists and $\mathcal{C}(x^*) = \mathcal{C}^{(s)}(x)$, then set $\mathcal{C}(x) = \mathcal{C}(x^*)$. Otherwise, skip $x$ in Labeling Stage 1;

5 *Labeling Stage 2*: In order of non-increasing $p(x)$, for each $x \in X$, find $x^*$ and $\mathcal{C}^{(s)}(x)$ (if it exists), as described in Section 2.2. If $\mathcal{C}^{(s)}(x)$ exists, set $\mathcal{C}(x) = \mathcal{C}^{(s)}(x)$. Otherwise, set $\mathcal{C}(x) = \mathcal{C}(x^*)$.

---

$x^* = \operatorname{argmin}_{y \in X}\{D_t(x,y) \mid \mathcal{C}(y) \neq 0 \wedge p(y) \geq p(x)\}$: the $D_t$-nearest neighbor of $x$ that is higher density and already labeled. SRDL then establishes the *spatial consensus label* of $x$, namely the majority among the spatial neighbors of $x$, denoted $\mathcal{C}^{(s)}(x)$. If the label of $x^*$ agrees with the spatial consensus label of $x$, then SRDL assigns $x$ that label. If not, $x$ is skipped in the first labeling stage. In Labeling Stage 2, each unlabeled point is given its spatial consensus label (if one exists). If not, SRDL assigns the label the $D_t$-nearest neighbor that is higher-density and already labeled.

## 3. THE M-SRDL CLUSTERING ALGORITHM

SRDL has been shown to be effective on a number of synthetic and real HSIs [6], but it is limited in that it focuses on a single clustering scale. In this section, we introduce a multiscale extension of SRDL, which we call the Multiscale Spatially-Regularized Diffusion Learning (M-SRDL) clustering algorithm (provided in Algorithm 2). M-SRDL implements SRDL at an exponential range of time scales $t \in \{0, 1, 2, 2^2, \ldots, 2^T\}$, where $T = \lceil \log_2[\log_{|\lambda_2(P)|}(\frac{2\tau}{\min(q)})]\rceil$. This cut-off is chosen so that, if $t \geq 2^T$, $\max\limits_{x,y \in X} D_t(x,y) \leq \tau$. Hence, $\tau \ll 1$ can be interpreted as a threshold for how small diffusion distances are allowed to become before terminating cluster analysis. Importantly, since M-SRDL uses SRDL to

4689

**Algorithm 2:** Multiscale Spatially-Regularized Diffusion Learning (M-SRDL) clustering algorithm

**Input:** $X$ (data), $\tau$ (threshold) $\sigma$ (scale parameter), $\sigma_0$ (KDE bandwidth), $R$ (spatial window)

**Output:** $\mathcal{C}_{t^*}$ (optimal clustering), $K_{t^*}$ (no. clusters)

1 Build KNN graph with $N$ $\ell^2$-nearest neighbors within a spatial window of $R > 0$. Weight edges according to a Gaussian kernel with scale parameter $\sigma$. Compute $P$ and $q$ from this graph;

2 Compute $T = \lceil \log_2 [\log_{|\lambda_2(P)|} (\frac{2\tau}{\min(s)})] \rceil$. For each time step $t_i \in \{0, 1, 2, 2^2, \ldots, 2^T\}$, calculate $[\mathcal{C}_{t_i}, K_{t_i}] = \text{SRDL}(X, t_i, N, \sigma, \sigma_0, R)$;

3 For each $t \in J = \{t | K_{t_i} \in [2, \frac{n}{2})\}$, calculate $VI^{(\text{tot})}(\mathcal{C}_t) = \sum_{u \in J} VI(\mathcal{C}_t, \mathcal{C}_u)$ ;

4 Solve $\mathcal{C}_{t^*} = \text{argmin}\{VI^{(\text{tot})}(\mathcal{C}_t) \mid t \in J\}$ and let $K_{t^*}$ be the number of clusters in $\mathcal{C}_{t^*}$;

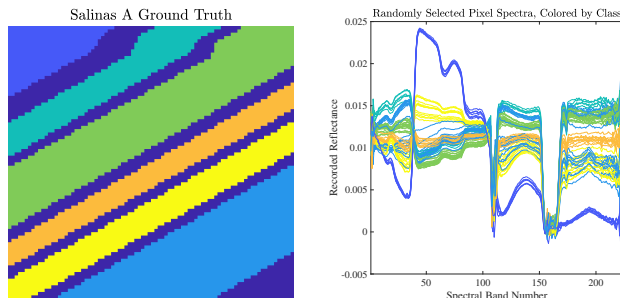extract clusterings, spatial regularization is incorporated into label assignments at all scales.

After extracting multiscale cluster structure from the HSI, M-SRDL solves $\mathcal{C}_{t^*} = \text{argmin}_{t \in J} \sum_{u \in J} VI(\mathcal{C}_t, \mathcal{C}_u)$, where $J = \{t_i | K_{t_i} \in [2, \frac{n}{2})\}$ is the set of time steps during which SRDL extracts nontrivial clusterings of $X$ and $VI$ is the *variation of information* (VI) distance metric between clusterings of $X$ [7]. The clustering $\mathcal{C}_{t^*}$ can be interpreted as the $VI$-barycenter of nontrivial clusterings generated by SRDL. Thus, M-SRDL integrates information gained from spatially-regularized partitions at multiple scales into a single representative clustering of $X$ [2, 6].

## 4. NUMERICAL EXPERIMENTS

In this section, we present analysis of M-SRDL on the real-world Salinas A HSI (Fig. 1) [15], which was generated by the Airborne Visible/Infrared Imaging Spectrometer over farmland in Salinas Valley, California, USA. This HSI encodes 224 spectral bands across an $83 \times 86$ image.

In Fig. 2, we evaluate M-SRDL on the Salinas A HSI

**Fig. 1**: Ground truth labels and spectra of a random sample of pixels from the Salinas A HSI [15]



and compare its clusterings against those of the *M-LUND algorithm* [2]. M-LUND is a multiscale extension of *the LUND algorithm* [9], which is identical to SRDL except that LUND does not include spatial regularization into its predictions. More precisely, LUND relies on a standard KNN graph to build $P$ and does not use spatial consensus to label non-modal points. M-LUND learns multiscale structure from $X$ by varying the input $t$ of LUND and outputs the $VI$-barycenter of learned nontrivial cluster structure as a clustering exemplar. Thus, the one difference between these two algorithms and their predictions is that M-SRDL incorporates spatial geometry into its labeling and M-LUND does not. If spatial regularization changes neither $\lambda_2(P)$ nor $\min(q)$, M-SRDL and M-LUND have identical computational complexity [2, 6]. M-SRDL and M-LUND were implemented using the same parameters $N = 100$, $\sigma = 1.30$, $\sigma_0 = 3.6 \times 10^{-3}$, and $\tau = 10^{-5}$. We set $R = 12$ when evaluating M-SRDL.

Both M-LUND and M-SRDL successfully extracted multiscale structure from the HSI (Fig. 2). Indeed, clusters are observed to merge as $t$ increases, reflecting that fewer diffusion map coordinates contribute to diffusion distances. However, spatial regularization clearly improved spatial smoothness of clusters, especially near cluster boundaries in the 1st and 2nd rows of Fig. 2. Indeed, while M-LUND assigned noisy labels to cluster boundaries, the spatial separation between clusters was strong for M-SRDL clusterings. Both methods assigned the coarsest clustering ($K_{t^*} = 2$ for M-LUND and $K_{t^*} = 3$ for M-SRDL) as $VI$-barycenters. The normalized mutual information [7] between these clusterings and the ground truth labels were 0.3016 for M-SRDL and 0.2836 for M-LUND respectively. Thus, spatial regularization resulted in a partition closer to the ground truth labels.

Software to replicate numerical experiments is available at https://github.com/sampolk/MultiscaleDiffusionClustering.
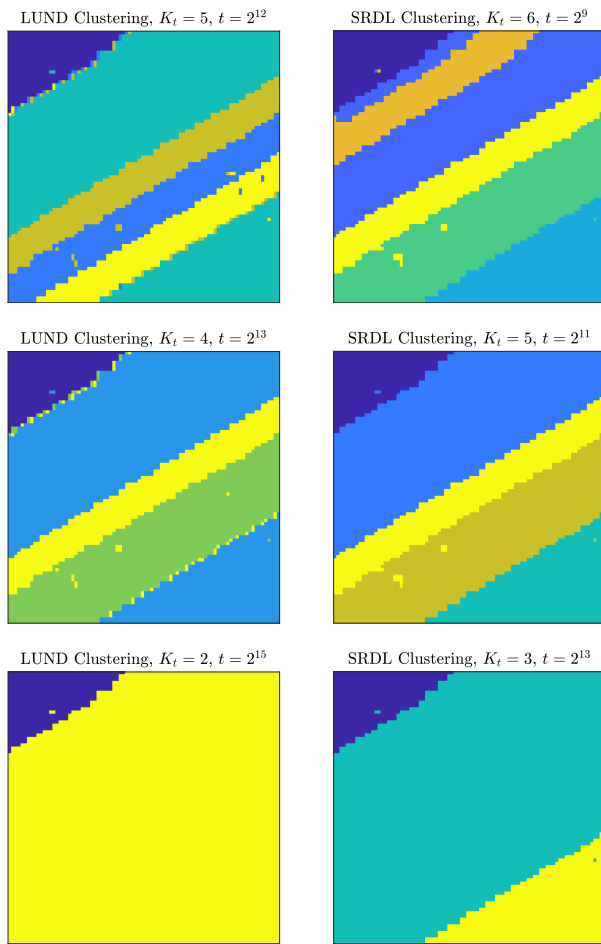
## 5. CONCLUSIONS

We conclude that spatially-regularized diffusion geometry is well-equipped for finding latent multiscale structure in high-dimensional images like HSIs and that spatial regularization improves cluster coherence. Future work includes extending past performance guarantees for M-LUND [2] to account for spatial regularization, which we observe improves the spatial coherence of cluster structure in the Salinas A HSI. Also, we expect that M-SRDL can be adapted for active learning, wherein a carefully chosen subset of pixels are queried for ground truth labels and exploited for semisupevised label propagation to the entire HSI [14, 16, 17, 18].

## 6. REFERENCES

[1] M. Eismann, *Hyperspectral Remote Sensing*, SPIE Press, 2012.

**Fig. 2**: Multiscale cluster structure learned by M-LUND [2] and M-SRDL from the Salinas A HSI. Spatial regularization yields smoother clusters, especially near cluster boundaries.



LUND Clustering, $K_t = 5$, $t = 2^{12}$ — SRDL Clustering, $K_t = 6$, $t = 2^9$

LUND Clustering, $K_t = 4$, $t = 2^{13}$ — SRDL Clustering, $K_t = 5$, $t = 2^{11}$

LUND Clustering, $K_t = 2$, $t = 2^{15}$ — SRDL Clustering, $K_t = 3$, $t = 2^{13}$

[2] J. M. Murphy and S. L. Polk, "A multiscale environment for learning by diffusion," *arXiv preprint arXiv:2102.00500*, 2021.

[3] 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, 2014.

[4] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pižurica, and W. Philips, "Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 14, no. 11, pp. 2142–2146, 2017.

[5] S. Lee and M.M. Crawford, "Hierarchical clustering approach for unsupervised image classification of hyperspectral data," in *IGARSS*. IEEE, 2004, vol. 2, pp. 941–944.

[6] J. M. Murphy and M. Maggioni, "Spectral-spatial diffusion geometry for hyperspectral image clustering," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 7, pp. 1243–1247, 2020.

[7] M. Meilă, "Comparing clusterings–an information based distance," *J. Multivar. Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

[8] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1, Springer Ser. Stat., 2001.

[9] M. Maggioni and J. M. Murphy, "Learning by unsupervised nonlinear diffusion," *J. Mach. Learn. Res.*, vol. 20, no. 160, pp. 1–56, 2019.

[10] Z. Meng, E. Merkurjev, A. Koniges, and A.L. Bertozzi, "Hyperspectral video analysis using graph clustering methods," *Image Process. Line*, vol. 7, pp. 218–245, 2017.

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

[12] 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.

[13] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, 2009.

[14] J. M. Murphy, "Spatially regularized active diffusion learning for high-dimensional images," *Pattern Recognit. Lett.*, vol. 135, pp. 213–220, 2020.

[15] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *JPL Airborne Geosci.*, 1999, pp. 217–227.

[16] M. Maggioni and J. M. Murphy, "Learning by active nonlinear diffusion," *Found. Data Sci.*, vol. 1, no. 3, pp. 271, 2019.

[17] 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.

[18] Z. Zhang, E. Pasolli, M. M. Crawford, and J. C. Tilton, "An active learning framework for hyperspectral image classification using hierarchical segmentation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 9, no. 2, pp. 640–654, 2015.