Unsupervised Geometric Learning: Theory and Applications

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

Unsupervised and Active Learning

Unsupervised learning assigns labels to datapoints $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ without training examples. In **active learning**, an algorithm may query a small number of parsimoniously chosen labels to help label the full dataset. Both unsupervised and active learning are important when large training sets are costly or unavailable. Unsupervised and active learning may be challenging when the underlying data is non-spherical, exhibits poor class separation, is corrupted by noise, or embedded in a high-dimensional space.

The classical **spectral clustering algorithm** [1; 2] computes the eigenfunctions of a graph Laplacian defined on a graph generated from X via weight matrix $W_{ij} = \exp(-\rho(x_i, x_j)^2/\sigma^2)$, for some metric ρ and constant σ. The eigenfunctions are consequently used as features in K-means [3].

Algorithm 1: Classical Spectral Clustering

Input: $\{x_i\}_{i=1}^n$ (Data) , $\sigma > 0$ (Scaling parameter)
Output: Y (Labels)

1: Compute the weight matrix $W \in \mathbb{R}^{n\times n}$ with $W_{ij} = \exp(-\rho(x_i, x_j)^2/\sigma^2)$.
2: Compute the diagonal degree matrix $D \in \mathbb{R}^{n\times n}$ with $D_{ii} = \sum_{j=1}^n W_{ij}$.
3: Form the symmetric normalized Laplacian $L_{\text{SYM}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$.
4: Compute the eigendecomposition $\{(\phi_k, \lambda_k)\}_{k=1}^n$, sorted so that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.
5: Estimate the number of clusters K as $\hat{K} = \arg\max_k \lambda_{k+1} - \lambda_k$.
6: For $1 \leq i \leq n$, let $\mathbf{v}_i = (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_{\hat{K}}(x_i))/||(\phi_1(x_i), \phi_2(x_i), \ldots, \phi_{\hat{K}}(x_i))||_2$ define the (row normalized) spectral embedding.
7: Compute labels Y by running K-means on the data $\{\mathbf{v}_i\}_{i=1}^n$ using $\hat{K}$ as the number of clusters.

Spectral clustering may struggle for elongated, poorly separated data, is highly sensitive to σ, and is poor at estimating K; see below.

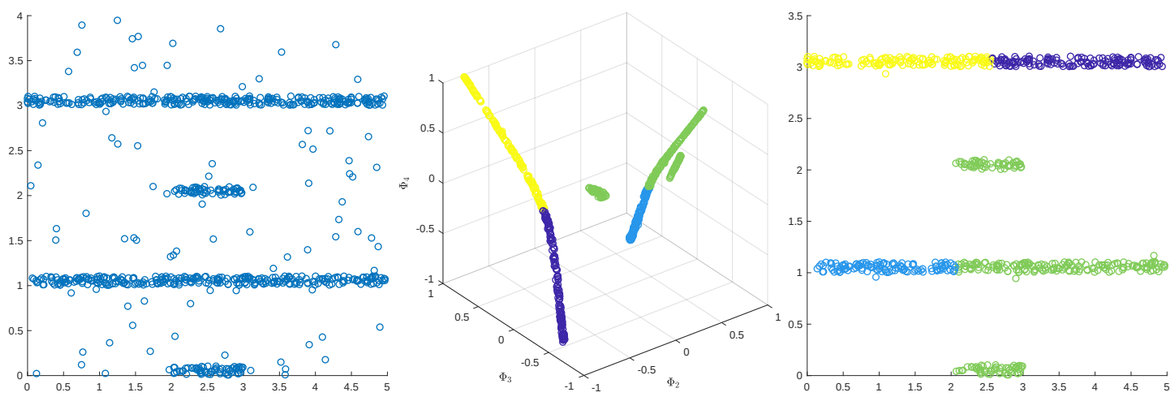

Figure 1: *Left to Right*: Raw data; embedding into $\mathbb{R}^3$, labels learned from spectral clustering.

We propose methods that enjoy performance guarantees with respect to:

1. **Labeling accuracy**: How good are the clustering labels?
2. **Estimating** K: Under what conditions does an unsupervised algorithm correctly estimate K?
3. **Parameter robustness**: How to learn the parameters of the algorithm without cross validation?

Spectral Clustering with Ultrametric Path Distances

We propose to use *ultrametric path distances* for spectral clustering.

Definition. *For $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, let G be the complete graph on X with edges weighted by Euclidean distance between points. For $x_i, x_j \in X$, let $\mathcal{P}(x_i, x_j)$ denote the set of all paths connecting x_i, x_j in G. The* longest-leg path distance (LLPD) *is:*

$$\rho_{\ell\ell}(x_i, x_j) = \min_{\{y_l\}_{l=1}^L \in \mathcal{P}(x_i, x_j)} \max_{l=1,2,\ldots,L-1} \|y_{l+1} - y_l\|_2.$$

The LLPD robustly compresses within-cluster distances while exacerbating between-cluster distances. The following simplified result characterizes the performance of LLPD spectral clustering [4]. Let $n_1, \ldots, n_K$, be the number of samples from intrinsically d-dimensional clusters, $n_{\min} = \min_{k=1,\ldots,K} n_k$ and let $\tilde{n}$ be the number of D-dimensional noise points, $d \leq D$. Denoise the data by removing points with LLPD to their k_{nse} nearest neighbor $\leq \theta$. Let the clusters be separated by minimum distance δ.

Theorem. *Under a suitable low-dimensional data model, suppose that $\tilde{n} \leq \left(\frac{C_2}{C_1}\right)^{\frac{k_{nse}D}{k_{nse}+1}} n_{min}^{\frac{D}{d+1}\left(\frac{k_{nse}}{k_{nse}+1}\right)}$. Let $f_\sigma(x) = e^{-x^2/\sigma^2}$ be the Gaussian kernel and assume $k_{nse} = O(1)$. Suppose the clusters are of comparable size. If n_{min} is large enough and σ satisfies $C_1 n_{min}^{-\frac{1}{d+1}} \leq \theta \leq C_2 \tilde{n}^{-\left(\frac{k_{nse}+1}{k_{nse}}\right)\frac{1}{D}}, C_3\theta \leq \sigma \leq C_4\delta$, then with high probability the denoised LDLN data X_N satisfies:*

(i) the largest gap in the eigenvalues of L_{SYM} is $\lambda_{K+1} - \lambda_K$.
(ii) spectral clustering with LLPD with K principal eigenvectors achieves perfect accuracy on X_N.

The constants $\{C_i\}_{i=1}^4$ depend on geometric properties of the data, but do not depend on $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

Consider high dimensional image data from the COIL database [5]. Below, we plot the eigenvalues for a graph Laplacian constructed with Euclidean distances and the LLPD, for a range of σ values. We see that there is a large gap between the 16th and 17th eigenvalues when the LLPD Laplacian is used, indicating that LLPD spectral clustering correctly estimates there are 16 clusters, for a range of σ values.

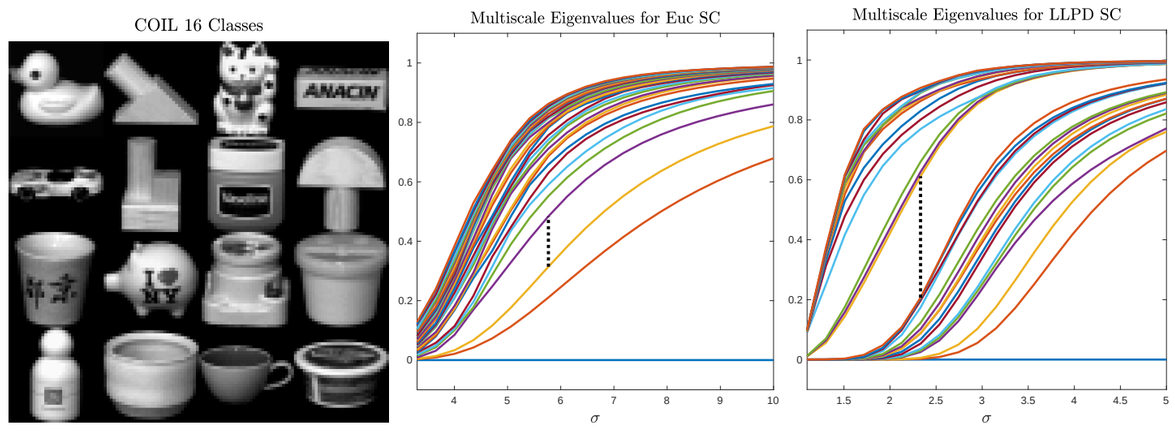

Figure 2: *Left to Right*: Representative COIL images; Euclidean multiscale eigenvalues; LLPD multiscale eigenvalues.

LUND: Learning by Unsupervised Nonlinear Diffusion

While longest leg path distances are powerful for unsupervised clustering, they are not robust to high density bottlenecks between clusters. *Diffusion distances* are a family of multiscale metrics that resolve such structures [6; 7]. For an ergodic Markov transition matrix P with stationary distribution π defined on $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$, the diffusion distance between x_i, x_j at time t is

$$D_t(x_i, x_j) = \sqrt{\sum_{\ell=1}^n (P_{i\ell}^t - P_{j\ell}^t)^2 \frac{1}{\pi(\ell)}}.$$

The *learning by unsupervised nonlinear diffusion (LUND)* algorithm enjoys performance guarantees and has shown state-of-the-art clustering performance on hyperspectral images [8; 9].

Algorithm 2: LUND

Input: X (data), σ_0 (kernel density bandwidth), σ (diffusion scaling parameter), t (time parameter), τ (threshold)
Output: Y (cluster assignments), $\hat{K}$ (estimated number of clusters)

1: Build Markov transition matrix P using scale parameter σ.
2: Compute an empirical density estimate $p(x)$ for all $x \in X$ using kernel bandwidth σ_0.
3: Compute $\rho_t(x)$, the distance to x's D_t-nearest neighbor of higher density, for all $x \in X$.
4: Compute $\mathcal{D}_t(x) = \rho_t(x)p(x)$ for all $x \in X$.
5: Sort X according to $\mathcal{D}_t(x)$ in descending order as $\{x_{m_i}\}_{i=1}^n, n = |X|$.
6: Compute $\hat{K} = \inf\{k \ |\mathcal{D}_t(x_{m_k})/\mathcal{D}_t(x_{m_{k+1}}) > \tau$.
7: Assign $Y(x_{m_i}) = i, \ i = 1, \ldots, \hat{K}$, and $Y(x_{m_i}) = 0, \ i = \hat{K}+1, \ldots, n$.
8: In order of decreasing $p(x)$ value, assign each point the same label as its nearest neighbor of higher density.

For well-separated, coherent clusters $X = \bigcup_{k=1}^K X_k$ (quantified by geometric constants $\{C_i\}_{i=1}^5$), there is a range of t for which diffusion distances are small within a cluster and large between a cluster. Let

$$D_t^{\text{in}} = \max_{k=1,\ldots,K} \max_{x,y\in X_k} D_t(x, y), D_t^{\text{btw}} = \min_{k\neq k'} \min_{x\in X_k, y\in X_{k'}} D_t(x, y).$$

Theorem. *Let $X = \bigcup_{k=1}^K X_k$ and let P be a corresponding Markov transition matrix on X, inducing diffusion distances $\{D_t\}_{t\geq 0}$. Then there exist constants $\{C_i\}_{i=1}^5 \geq 0$ such that the following holds: for any $\epsilon > 0$, and for any t satisfying $C_1 \ln\left(\frac{C_2}{\epsilon}\right) < t < C_3\epsilon$, we have $D_t^{in} \leq C_4\epsilon, D_t^{btw} \geq C_5 - C_4\epsilon$.*

These estimates translate to performance guarantees on the LUND algorithm itself. Let $\mathcal{M}$ be the density maximizers of distinct classes.

Theorem. *Suppose $X = \bigcup_{k=1}^K X_k$ as above. LUND labels all points accurately, and correctly estimates K, provided that $D_t^{in}/D_t^{btw} < \min(\mathcal{M})/\max(\mathcal{M})$.*

The LUND algorithm can be adapted to analyze high-dimensional hyperspectral data [10; 11], shown below.

Figure 3: *Left to Right*: Compressed Indian Pines HSI; Indian Pines ground truth; high confidence LUND labels; spatially regularized LUND labels.

Active Learning With Diffusion Geometry

The LUND approach may be modified into an active learning algorithm by *querying* the learned modes for labels, rather than assuming they belong to distinct classes [12]. This increases robustness to the types of distributions that can be learned with diffusion geometry, and is comparable to cluster-based active learning [13]. Some results comparing the resulting *learning by active nonlinear diffusion (LAND)* algorithm to related active learning methods for hyperspectral data appear below.

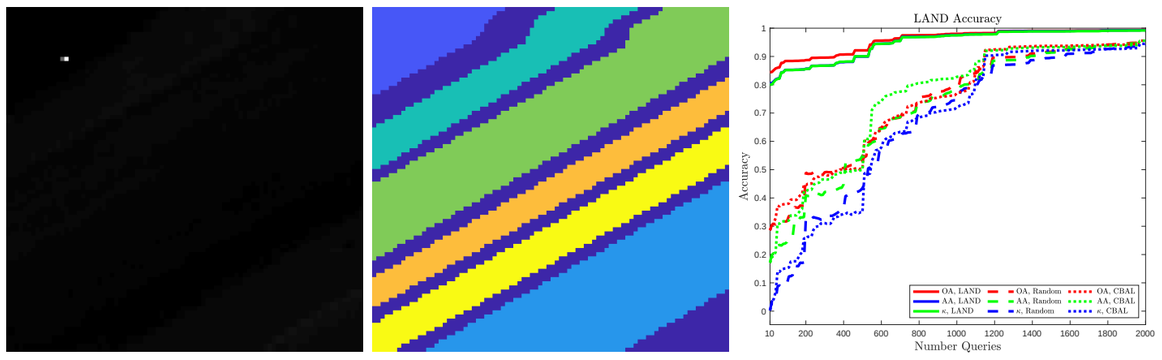

Figure 4: *Left to Right*: Compressed Salinas A HSI; Salinas A ground truth; active learning results as a function of number of queries.

Related methods based on querying points near the estimated cluster boundaries are also possible [14].

Code, Contact, Collaborators

- Code and papers: https://jmurphy.math.tufts.edu/
- Contact: jm.murphy@tufts.edu
- Partially joint with Anna Little (Michigan State) and Mauro Maggioni (Johns Hopkins)

References

[1] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[2] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.
[3] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
[4] A. Little, M. Maggioni, and J.M. Murphy. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *arXiv preprint arXiv:1712.06206*, 2017.
[5] S.A. Nene, S.K. Nayar, and H. Murase. Columbia object image library (coil-20). 1996.
[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.*, 102(21):7426–7431, 2005.
[7] R.R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.
[8] M. Maggioni and J.M. Murphy. Learning by unsupervised nonlinear diffusion. *arXiv preprint arXiv:1810.06702*, 2018.
[9] J.M. Murphy and M. Maggioni. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1829–1845, 2019.
[10] 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*, volume 10644, page 106440I. International Society for Optics and Photonics, 2018.
[11] J.M. Murphy and M. Maggioni. Spectral-spatial diffusion geometry for hyperspectral image clustering. *arXiv preprint arXiv:1902.05402*, 2019.
[12] M. Maggioni and J.M. Murphy. Diffusion geometric active learning. *Preprint*, 2019.
[13] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008.
[14] J.M. Murphy and M. Maggioni. Iterative active learning with diffusion geometry for hyperspectral images. In *Proc. 9th Workshop Hyperspectral Image Signal Process. Evol. Remote Sens.(WHISPERS)*, 2018.