Semi-Supervised Power Weighted Shortest Path Distances

Daniel Mckenzie [1] Steven Damelin [2]

[1] University of Georgia

[2] American Mathematical Society

October 20th 2018

Overview: Clustering Euclidean Data

- **Clustering:** Given $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$, find partition into clusters:

$$\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_k \tag{1}$$

- In this talk we consider:
  - **SS Clustering** Given $\mathcal{Y} \subset \mathcal{X}$ with $\mathcal{Y} = \mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_k$ *known*, find (1) with $\mathcal{Y}_a \subset \mathcal{X}_a$.
  - **Cluster Extraction** Given $\mathcal{Y}_a \subset \mathcal{X}_a$, find $\mathcal{X}_a$.
- We propose a distance $d^{ss,p}(\cdot,\cdot)$ on $\mathcal{X}$ incorporating labeled data $\mathcal{Y}$.
- We provide theoretical[1] and experimental evidence that using $d^{ss,p}(\cdot,\cdot)$ instead of Euclidean distance can improve accuracy of many algorithms.

[1] using results of Hwang, Damelin, and Hero 2016.

Graphical Approaches to Clustering

- Convert $\mathcal{X}$ to weighted graph $G = (V, E, W)$ with $V = \{v_1, \ldots, v_N\}$ and $W_{ij} = \varphi(d(\mathbf{x}_i, \mathbf{x}_j))$.
- Require $\varphi : \mathbb{R} \to \mathbb{R}$ to be non-increasing, continuous at 0, fast-decaying.
- Typical example: $\varphi(d(\mathbf{x}_i, \mathbf{x}_j)) = \exp\left(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma^2\right)$
- More refined example[2]:

$$W_{ij} = \begin{cases} \exp\left(-d(\mathbf{x}_i, \mathbf{x}_j)^2/\sigma_i\sigma_j\right) & \text{if } \mathbf{x}_j \text{ amongst } r\text{-NN of } \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

  Here $\sigma_i = \|\mathbf{x}_i - \mathbf{x}_{[\ell,i]}\|_2$ where $\mathbf{x}_{[\ell,i]}$ is ℓ-th nearest neighbor of $\mathbf{x}_i$.
- Usually $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$.

[2] Zelnik-Manor and Perona 2005.

Data Driven Metrics/Distances

- It makes sense to consider d dependent on $\mathcal{X}$.
- Nearest Neighbor metric[3], more generally density based distances[4].
- **Shortest Path Distances**[5][6] $d_{SP}(\mathbf{x}_i, \mathbf{x}_j) = \min_{\gamma} \sum_{j=0}^{m} \|\mathbf{x}_{i_{j+1}} - \mathbf{x}_{i_j}\|_2$

  where $\gamma = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}\} \subset \mathcal{Q} \subset \mathcal{X}$ and $\mathbf{x}_{i_0} := \mathbf{x}_i$, $\mathbf{x}_{i_{m+1}} := \mathbf{x}_j$
- Longest leg distance[7].
- Diffusion Distances[8].

[3] Cohen et al. 2015.
[4] Orlitsky and Sajama 2005.
[5] Vincent and Bengio 2003.
[6] Tenenbaum, De Silva, and Langford 2000.
[7] Little, Maggioni, and Murphy 2017.
[8] Coifman and Lafon 2006.

Semi-Supervised Power weighted Path Distances

- If available, it makes sense to incorporate labeled data $\mathcal{Y}$ into metric.
- For a fixed a, and $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, define a *path through* $\mathcal{Y}_a$ *from* $\mathbf{x}_i$ *to* $\mathbf{x}_j$ as any subset $\gamma := \{\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m}\} \subset \mathcal{Y}_a$.
- Power-weighted length of path (for $p > 1$):

$$\ell^p(\gamma) = \|\mathbf{x}_i - \mathbf{y}_{i_1}\|^p + \sum_{j=1}^{m-1} \|\mathbf{y}_{i_j} - \mathbf{y}_{i_{j+1}}\|^p + \|\mathbf{y}_{i_m} - \mathbf{x}_j\|^p$$

- For $a = 1, \ldots, k$ define $d^{a,p}(\mathbf{x}_i, \mathbf{x}_j) := \min_{\gamma} \ell^p(\gamma)$.
- Define $d^{ss,p}(\mathbf{x}_i, \mathbf{x}_j) := \min\{\min_{a} d^{a,p}(\mathbf{x}_i, \mathbf{x}_j), \|\mathbf{x}_i - \mathbf{x}_j\|_2^p\}$.
- Bijral *et. al*[9] consider a similar, but different approach.

[9] Bijral, Ratliff, and Srebro 2011.

Visualizing the geodesics

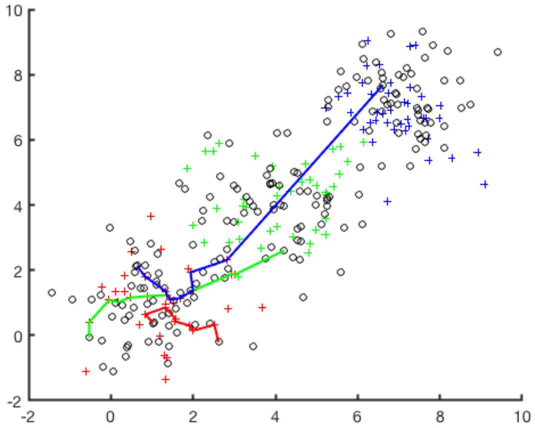

Figure: Three clusters drawn from three Gaussian distributions. Labeled data indicated by colored crosses. Paths shown are geodesics for $d^{1,2}$.

Visualizing the geodesics

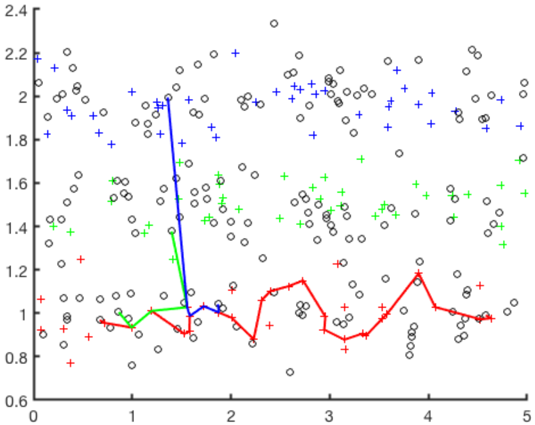

Figure: Data drawn from three thickened lines. Labeled data indicated by colored crosses. Paths shown are geodesics for $d^{1,2}$.

An appropriate generative model

- $\mathcal{M}_1, \ldots, \mathcal{M}_k \subset \mathbb{R}^D$ smooth, embedded, compact manifolds. $\dim(\mathcal{M}_a) = d_a \ll D$.
- $\mathcal{X}_a \sim \mathcal{M}_a$ uniformly i.i.d for $a = 1, \ldots, k$. Let $\mathcal{X} = \cup_{a=1}^k \mathcal{X}_a$
- $\text{dist}(\mathcal{M}_a, \mathcal{M}_b) := \min_{\mathbf{u} \in \mathcal{M}_a, \mathbf{v} \in \mathcal{M}_b} \|\mathbf{u} - \mathbf{v}\| \geq \delta$ for all $a \neq b$.
- Such data models are widely-studied[10] and are hypothesized to describe real-world data such as hand-written digits, faces, etc.
- Assume labeled data $\mathcal{Y} = \cup_{a=1}^k \mathcal{Y}_a$ with $\mathcal{Y}_a \subset \mathcal{X}_a$ selected at random.
- Let $m_a = |\mathcal{Y}_a|$ and assume $m_1 \approx m_2 \approx \ldots \approx m_k$.

[10] Arias-Castro 2011.

Analyzing the Shortest-Paths metric for this Data model

Two key parameters for clustering algorithms:

$$\epsilon_1 := \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_a} d^{ss,p}(\mathbf{x}_i, \mathbf{x}_j) \quad \text{(Max. in-cluster distance.)}$$

$$\epsilon_2 := \min_{\substack{\mathbf{x}_i \in \mathcal{X}_a, \mathbf{x}_j \in \mathcal{X}_b \\ a \neq b}} d^{ss,p}(\mathbf{x}_i, \mathbf{x}_j) \quad \text{(Min. between-cluster distance.)}$$

Want ϵ_1 small and ϵ_2 large. We are able to show that:

Lemma (Damelin & M.)

$\epsilon_2 \geq \delta^p$ *where* $\delta := \min_{\substack{\mathbf{u} \in \mathcal{M}_a, \mathbf{v} \in \mathcal{M}_b \\ a \neq b}} \|\mathbf{u} - \mathbf{v}\|$

Theorem (Damelin & M.)

$\epsilon_1 = O(m^{(1-p)/d}) \to 0$ *where* $m = |\mathcal{Y}|$.

Bounding minimal between-cluster distance

- Recall that:

$$d^{ss,p}(\mathbf{x}_i,\mathbf{x}_j) := \min\{\min_c d^{c,p}(\mathbf{x}_i,\mathbf{x}_j), \|\mathbf{x}_i - \mathbf{x}_j\|_2^p\}$$

$$d^{c,p}(\mathbf{x}_i,\mathbf{x}_j) = \min_\gamma \left(\|\mathbf{x}_i - \mathbf{y}_{i_1}\|^p + \sum_{j=1}^{m-1} \|\mathbf{y}_{i_j} - \mathbf{y}_{i_{j+1}}\|^p + \|\mathbf{y}_{i_m} - \mathbf{x}_j\|^p \right)$$

  Where $\gamma = \{\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m}\} \subset \mathcal{Y}_c$.

- If $\mathbf{x}_i \in \mathcal{X}_a$ and $\mathbf{x}_j \in \mathcal{X}_b$
  1. $\|\mathbf{x}_i - \mathbf{x}_j\|^p \geq \delta^p$
  2. Either $c \neq a$ or $c \neq b$, so $\|\mathbf{x}_i - \mathbf{y}_{i_1}\|^p \geq \delta^p$ or $\|\mathbf{y}_{i_1} - \mathbf{x}_j\|^p \geq \delta^p$
- Hence $\epsilon_2 \geq \delta^p$

Intrinsic Path distances

- Let g_a denote restriction of Euclidean (Riemannian) metric to $\mathcal{M}_a$. For any $\mathbf{u}, \mathbf{v} \in \mathcal{M}_a$ can define *intrinsic distance:*

$$d_{\mathcal{M}_a}(\mathbf{u}, \mathbf{v}) = \inf_{\lambda} \int_0^1 \sqrt{g_a(\lambda'(t), \lambda'(t))} dt$$

  where $\lambda : [0, 1] \to \mathcal{M}_a$ with $\lambda(0) = \mathbf{u}$ and $\lambda(1) = \mathbf{v}$.
- For $\mathbf{u}, \mathbf{v} \in \mathcal{M}_a$ define:

$$d^p_{S,a}(\mathbf{x}_i, \mathbf{x}_j) := \min_{\gamma} \left(d_{\mathcal{M}_a}(\mathbf{u}, \mathbf{y}_{i_1})^p + \sum_{j=1}^{m-1} d_{\mathcal{M}_a}(\mathbf{y}_{i_j}, \mathbf{y}_{i_{j+1}})^p + d_{\mathcal{M}_a}(\mathbf{y}_{i_m}, \mathbf{v})^p \right)$$

Where again $\gamma = \{\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m}\} \subset \mathcal{Y}_a$

Bounding maximal in-cluster distance 1

Lemma

For any $\mathcal{M} \subset \mathbb{R}^D$ with induced Riemannian metric and any $\mathbf{u}, \mathbf{v} \in \mathcal{M}$:

$$\|\mathbf{u} - \mathbf{v}\| \leq d_{\mathcal{M}}(\mathbf{u}, \mathbf{v})$$

Corollary

For all $a = 1, \ldots, k$ and all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}_a$:

$$d^{a,p}(\mathbf{x}_i, \mathbf{x}_j) \leq d^p_{S,a}(\mathbf{x}_i, \mathbf{x}_j)$$

Will show $d^p_{S,a}(\mathbf{x}_i, \mathbf{x}_j) = O(m^{(1-p)/d})$

Bounding maximal in-cluster distance 2

Theorem (From Theorem 1 in Hwang, Damelin, and Hero 2016)

Assume $\mathcal{Y}_a$ drawn uniformly i.i.d from $\mathcal{M}_a$, with $|\mathcal{Y}_a| = m_a$. Define $r_{m_a} := m_a^{(1-p)/pd}$ and fix $\epsilon > 0$:

$$\mathbb{P}\left(\sup_{\substack{\mathbf{u},\mathbf{v}\in\mathcal{M}_a \\ d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v})\geq r_{m_a}}} \left| \frac{d_{S,a}^{p}(\mathbf{u},\mathbf{v})}{m^{(1-p)/d}\nu_{\mathcal{M}_a}^{(p-1)/d} d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v})} - C(d_a,p) \right| > \epsilon \right) = o_{m_a}(1) \tag{2}$$

where $\nu_{\mathcal{M}_a} = Vol(\mathcal{M}_a)$

Rearranging, with probability $1 - o(1)$:

$$\begin{aligned}\max_{\substack{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}_a \\ d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v})\geq r_{m_a}}} d_{S,a}^{p}(\mathbf{x}_i,\mathbf{x}_j) &\leq \left(C(d_a,p)+\epsilon\right)\nu_{\mathcal{M}_a}^{(p-1)/d} m^{(1-p)/d} \max_{\mathbf{u},\mathbf{v}\in\mathcal{M}_a} d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v}) \\ &= \widetilde{C}_a m^{(1-p)/d}\end{aligned}$$

Bounding maximal in-cluster distance 3

Hence with probability $1 - o(1)$:

$$\max_{\substack{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}_a \\ d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v}) \geq r_{m_a}}} d^{a,p}(\mathbf{x}_i, \mathbf{x}_j) \leq \max_{\substack{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}_a \\ d_{\mathcal{M}_a}(\mathbf{u},\mathbf{v}) \geq r_{m_a}}} d^p_{S,a}(\mathbf{x}_i, \mathbf{x}_j) = \widetilde{C}_a m^{(1-p)/d}$$

So:

$$\begin{aligned}
\max_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}_a} d^{ss,p}(\mathbf{x}_i, \mathbf{x}_j) &\leq \max_{\mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}_a} \min\left(d^{a,p}(\mathbf{x}_i, \mathbf{x}_j), \|\mathbf{x}_i - \mathbf{x}_j\|^p\right. \\
&\leq \max\left(\widetilde{C}_a m^{(1-p)/d}, r^p_{m_a}\right) \\
&= O(m_a^{(1-p)/d}) \text{ as } r_{m_a} = m_a^{(1-p)/pd}
\end{aligned}$$

Experimental Results: Set-up

- For each data set $\mathcal{X} \subset \mathbb{R}^D$ draw $\mathcal{Y} \subset \mathcal{X}$ at random. $m_1 = m_2 = \ldots = m_k$.
- Computed Euclidean (E) distances: $A_{ij}^{(1)} = \|\mathbf{x}_i - \mathbf{x}_j\|$ and Path-Based (P-B) distances: $A_{ij}^{(2)} = d^{ss,p}(\mathbf{x}_i, \mathbf{x}_j)$ with $p = 2$.
- Use ZMP local scaling (for $\alpha = 1, 2$):

$$W_{ij}^{(\alpha)} = \begin{cases} \exp\left(-\left(A_{ij}^{(\alpha)}\right)^2 / \sigma_i \sigma_j\right) & \text{if } \mathbf{x}_j \text{ amongst } r\text{-NN of } \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases}$$

Experimental Results: Algorithms

Algorithms for SS Clustering:

1. TV-based partitioning with a Regional Force (`TVRF`) (Yin and Tai 2018)
2. Normalized Spectral Clustering (`Spectral`) (Ng, Jordan, and Weiss 2002)
3. Iterated SS Cluster Pursuit (`ISSCP`) (Lai and Mckenzie 2018)

Algorithms for Cluster Extraction:

1. Local Spectral Diffusion (`LOSP++`) (He et al. 2016)
2. Heat Kernel Diffusion (`HKGrow`) (Kloster and Gleich 2014)
3. SS Cluster Pursuit (`SSCP`) (Lai and Mckenzie 2018)

Experimental Results: Three Lines

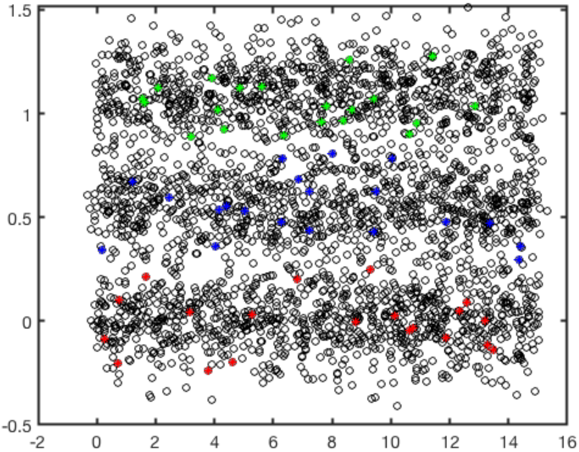

Figure: Three Lines. Artificial data set with 3000 points and three, equally sized clusters. Labeled data shown with colored crosses.

Experimental Results: Three Lines

	TVRF		ISSCP		Spectral	
	E	P-B	E	P-B	E	P-B
1%	43.13%	50.45%	**67.82**%	66.71%	34.78%	35.21%
2%	69.41%	72.79%	**77.75**%	76.47%	34.83%	34.86%
3%	81.47%	82.38%	**83.37**%	82.45%	34.79%	34.77%
4%	83.92%	**85.95**%	83.43%	81.53%	34.53%	34.70%
5%	**90.13**%	89.67%	87.09%	85.21%	34.55%	34.65%

Table: Comparing accuracy—Euclidean (E) versus Path-Based (P-B)—for three SS clustering algorithms on ‘Three Lines’ data set. Amount of labeled data varying from 1% to 5%.

Experimental Results: Subset of MNIST

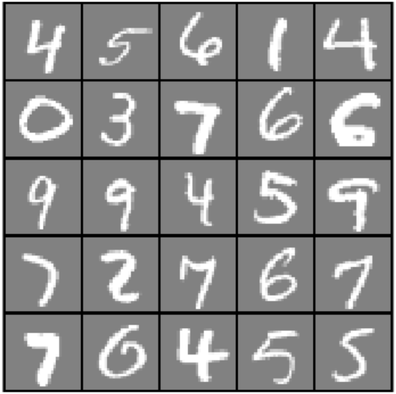

Figure: MNIST. 500 pictures of each digit 0–4 chosen at random for total of 2500 data points. Images converted to vectors, PCA done, 50 highest principal components kept.

Experimental Results: Subset of MNIST

	TVRF		ISSCP		Spectral	
	E	P-B	E	P-B	E	P-B
1%	20.12%	47.46%	89.90%	**91.07**%	78.76%	79.20%
2%	79.26%	83.06%	91.98%	**92.47**%	78.76%	79.20%
3%	98.46%	**98.61**%	91.82%	92.68%	78.76%	79.20%
4%	98.52%	**98.58**%	97.04%	97.29%	78.76%	79.20%
5%	98.56%	**98.56**%	97.33%	97.70%	78.76%	79.20%

Table: Comparing accuracy—Euclidean (E) versus Path-Based (P-B)—for three SS clustering algorithms on 'MNIST' data set. Amount of labeled data varies from 1% to 5%.

Experimental Results: Columbia Object Image Library (COIL)

Chapelle *et. al.*[11] constructed a standard SSL data set by:

- Taking only red channel, downsampling to 16×16.
- Choosing 24 objects, grouped into 6 categories. 250 images per category.

[11] Chapelle, Scholkopf, and Zien 2006.

Experimental Results: COIL

	TVRF		Spectral	
	Euclidean	Path-Based	Euclidean	Path-Based
1%	57.27%	**60.53**%	33.67%	41.93%
2%	57.73%	**66.73**%	37.80%	42.00%
3%	68.20%	**77.00**%	37.80%	37.80%
4%	71.93%	**85.60**%	37.47%	37.40%
5%	81.53%	**89.13**%	37.93%	37.93%

Table: Comparing accuracy—Euclidean (E) versus Path-Based (P-B)—for two SS clustering algorithms on 'COIL' data set. Amount of labeled data varies from 1% to 5%

Experimental Results: One Line With Background

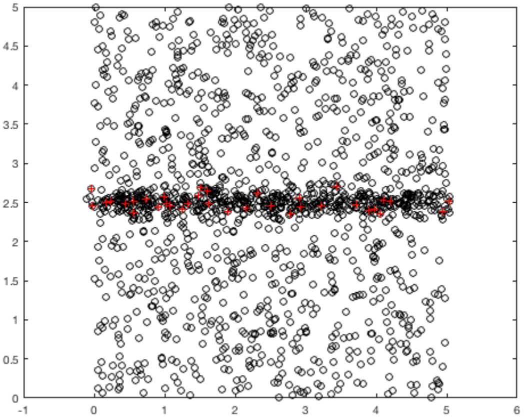

Figure: 500 data points drawn from thickened line and 1000 background points drawn uniformly in $[0,5]^{10} \subset \mathbb{R}^{10}$. 2-D projection shown

Experimental Results: One Line With Background

	SSCP		LOSP++		HKGrow	
	E	P-B	E	P-B	E	P-B
3%	**0.83**	0.81	0.55	0.43	0.58	0.76
4%	**0.84**	0.83	0.53	0.47	0.58	0.68
5%	**0.96**	0.93	0.72	0.51	0.68	0.95
6%	**0.94**	0.92	0.69	0.62	0.89	0.88
7%	**0.96**	0.93	0.77	0.68	0.81	0.81

Table: Comparing Jaccard index—Euclidean (E) versus Path-Based (P-B)— for three cluster extraction algorithms. Amount of labeled data varies from 3% of cluster of interest to 7% of cluster of interest

Concluding Remarks

Future Directions

1. Improve computational speed of computing $d^{ss,p}$.
2. Provide estimates on amount of labeled data ($|\mathcal{Y}|$) required.
3. Test on more SS Clustering and Cluster extraction algorithms.

Questions or Comments: danmac29@uga.edu

Thank you!