Selecting the independent coordinates of manifolds with large aspect ratios

Yu-Chia Chen (yuchaz@uw.edu)
Department of Electrical & Computer Engineering
University of Washington
October 22, 2019

Information

- Joint work with Marina Meila.
- Please check our paper [CM19] for more detail:
  - Yu-Chia Chen and Marina Meila. "Selecting the independent coordinates of manifolds with large aspect ratios". NeurIPS'19. (To appear, also on `arXiv:1907.01651`).
- Paper, Slides & Poster can be downloaded here[1].
- Codes will be made available soon and can also be accessed here[1] when it is ready.

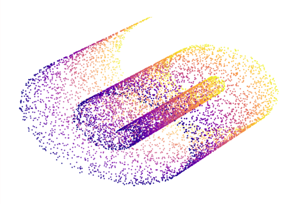
[1] `https://yuchaz.github.io/publication/2019-indep-coord-search`

Introduction

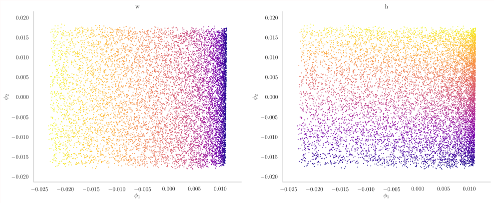
Introduction

Basic setup

Given data $\mathbf{X} \in \mathbb{R}^{n \times D}$ sampled from a *smooth* d-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^D$. Manifold Learning algorithms map $\mathbf{x}_i, i \in [n]$ to $\mathbf{y}_i = \phi(\mathbf{x}_i) \in \mathbb{R}^s$, where $d \leqslant s \ll D$, thus reducing the dimension of the data $\mathbf{X}$ while preserving (some of) its properties.

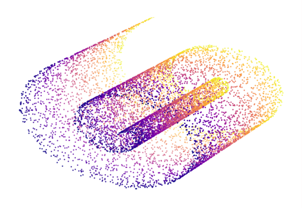


$\xrightarrow{\phi}$



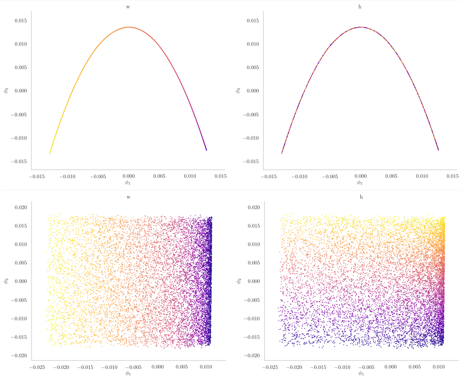
Independent eigen-coordinates search problem [GZKR08]

A family of Manifold learning algorithms **fail** apparently or suffer from **artifacts** when data manifold $\mathcal{M}$ is long and thin, e.g., when it has **aspect ratio** > 2 for 2D strip.

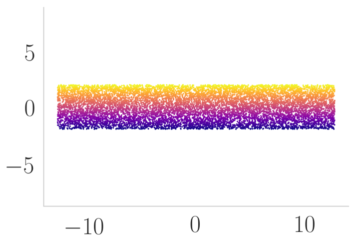


$\xrightarrow{\phi\,(\text{degenerate})}$

$\xrightarrow{\phi\,(\text{independent})}$

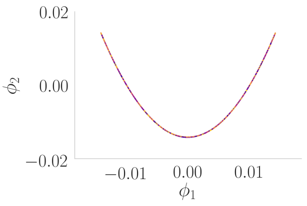


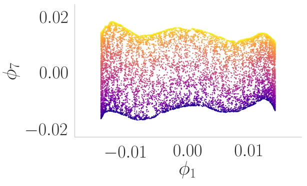
Example with Diffusion map embedding — I



$\xrightarrow{\phi\,(\text{degenerate})}$

$\xrightarrow{\phi\,(\text{independent})}$

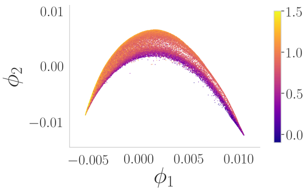




Example with Diffusion map embedding — II

$\xrightarrow{\phi\,(\text{degenerate})}$

$\xrightarrow{\phi\,(\text{independent})}$



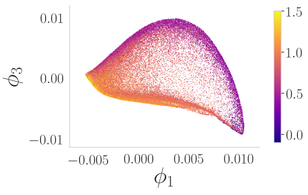


Image source[2].

[2] http://imgsrc.hubblesite.org/hu/db/images/hs-1999-25-a-full_tif.tif

Local, spectral embedding algorithms

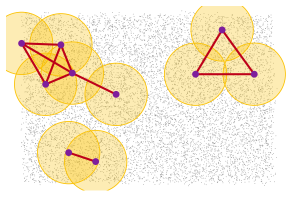
The family of algorithms suffer from the artifacts

- Laplacian eigenmap (LE) [BN03]
- Diffusion map (DM) [CL06]
- Locally linear embedding (LLE) [RS00]
- Local tangent space alignment (LTSA) [ZZ02]
- Hessian LLE (HLLE) [DG03]

In this work, we focus on *diffusion map* algorithm. But we will also discuss the possible extensions & challenges to LTSA & HLLE algorithms.

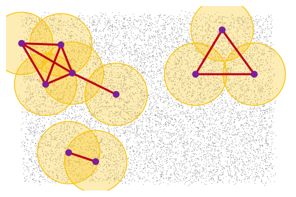
Algorithmic framework

1. Build neighborhood graph $G = (V, E)$.
   - ε-ball kernel.
   - k nearest neighbor (kNN) kernel.
   - self-tuning kernel (e.g., continuous kNN [BS16]).



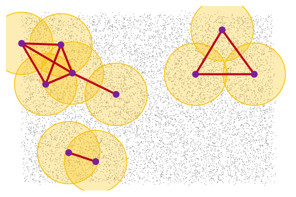
2. Construct matrix $\mathbf{M}$ from neighborhood graph G.
3. Solve the min-eigen problem of $\mathbf{M}$.
   - m-dimensional embedding is obtained from the 2nd to $m+1$th minimum eigenvectors, i.e., $\phi = [\phi_1, \cdots, \phi_m]$
   - In this work, we will show that the coordinates chosen by the above criteria will **not** give us an optimal embedding.

1. Build neighborhood graph $G = (V, E)$.
   - ε-ball kernel.
   - k nearest neighbor (kNN) kernel.
   - self-tuning kernel (e.g., continuous kNN [BS16]).



2. Construct matrix $\mathbf{M}$ from neighborhood graph G.
3. Solve the min-eigen problem of $\mathbf{M}$.
   - m-dimensional embedding is obtained from the 2^{nd} to $m+1^{th}$ minimum eigenvectors, i.e., $\phi = [\phi_1, \cdots, \phi_m]$
   - In this work, we will show that the coordinates chosen by the above criteria will **not** give us an optimal embedding.

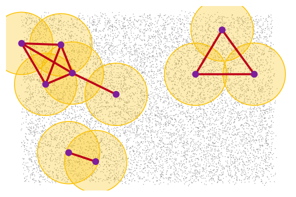
1. Build neighborhood graph $G = (V, E)$.
   - ε-ball kernel.
   - k nearest neighbor (kNN) kernel.
   - self-tuning kernel (e.g., continuous kNN [BS16]).



2. Construct matrix $\mathbf{M}$ from neighborhood graph G.
3. Solve the min-eigen problem of $\mathbf{M}$.
   - m-dimensional embedding is obtained from the 2nd to $m + 1$th minimum eigenvectors, i.e., $\phi = [\phi_1, \cdots, \phi_m]$
   - In this work, we will show that the coordinates chosen by the above criteria will **not** give us an optimal embedding.

Algorithmic framework

1. Build neighborhood graph $G = (V, E)$.
   - ε-ball kernel.
   - k nearest neighbor (kNN) kernel.
   - self-tuning kernel (e.g., continuous kNN [BS16]).



2. Construct matrix $\mathbf{M}$ from neighborhood graph G.
3. Solve the min-eigen problem of $\mathbf{M}$.
   - m-dimensional embedding is obtained from the 2^{nd} to $m+1^{\text{th}}$ minimum eigenvectors, i.e., $\phi = [\phi_1, \cdots, \phi_m]$
   - In this work, we will show that the coordinates chosen by the above criteria will **not** give us an optimal embedding.

Laplacian eigenmap/diffusion map algorithm

1. Build neighborhood graph $G(V, E)$ with 3ε-ball kernel.
   - $V = [n], E = \{(i, j) \in V^2 : \|\mathbf{x}_i - \mathbf{x}_j\| \leqslant 3\varepsilon\}$.
2. Compute kernel matrix $[\mathbf{K}]_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\varepsilon^2)$ and the *renormalized graph Laplacian* $\mathbf{L}$

$$\mathbf{L} = \mathbf{I} - \mathbf{W}^{-1}\mathbf{D}^{-1}\mathbf{K}\mathbf{D}^{-1}$$

   where $\mathbf{D} = \text{diag}(\mathbf{K}\mathbf{1}_n)$ and $\mathbf{W} = \text{diag}\left(\mathbf{D}^{-1}\mathbf{K}\mathbf{D}^{-1}\mathbf{1}_n\right)$

3. An m dimensional embedding is obtained from the 2^{nd} to $m + 1^{\text{th}}$ min-eigenvectors of the graph Laplacian $\mathbf{L}$.
   - Coordinates chosen by the above criteria will suffer from the IES artifacts.

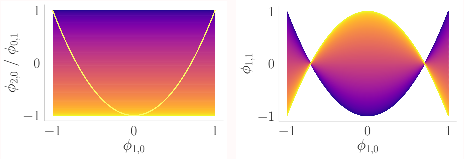
Introduction

Motivating example

The eigenvalues & eigen-functions of Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ (Neumann boundary condition) on 2D long strip, measurement is (width, height) = (W, H), are

$$\lambda_{k_1,k_2} = \left(\frac{k_1\pi}{W}\right)^2 + \left(\frac{k_2\pi}{H}\right)^2$$

$$\phi_{k_1,k_2}(w, h) = \cos\left(\frac{k_1\pi w}{W}\right)\cos\left(\frac{k_2\pi h}{H}\right)$$



$\phi_{1,0}, \phi_{0,1}$ are independent thus should be chosen, while, e.g., $\phi_{1,0}, \phi_{2,0}$ are not.

Motivating example – 2D strip

For example, let $H = 1, W = 2\pi$, we have

	$k_1 = 0$	$k_1 = 1$	$k_1 = 2$	$k_1 = 3$	$k_1 = 4$	$k_1 = 5$	$k_1 = 6$	$k_1 = 7$
$k_2 = 0$	0	1/2 1st	1 2nd	3/2 3rd	2 4th	5/2 5th	3 6th	7/2 8th
$k_2 = 1$	π 7th	...	...	...	...	...	...	...

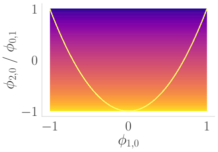
- Sort ϕ_k by λ_k, the first two eigenvalues are $\lambda_{1,0}$ and $\lambda_{2,0}$.
- $\lambda_{0,1}$ corresponds to the $\lceil W/H \rceil$-th ($= 7$-th here) eigenvalue.
- ϕ_1, ϕ_2 are orthogonal, but not functionally independent.
- $\phi_1, \phi_{\lceil W/H \rceil}$ are functionally independent, therefore $\{1, \lceil W/H \rceil\}$ should be chosen.

Motivating example – 2D strip

For example, let $H = 1$, $W = 2\pi$, we have

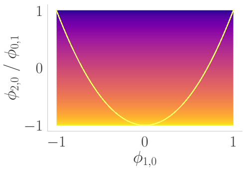
	$k_1 = 0$	$k_1 = 1$	$k_1 = 2$	$k_1 = 3$	$k_1 = 4$	$k_1 = 5$	$k_1 = 6$	$k_1 = 7$
$k_2 = 0$	0	1/2 1st	1 2nd	3/2 3rd	2 4th	5/2 5th	3 6th	7/2 8th
$k_2 = 1$	π 7th	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

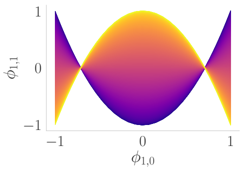
- Sort ϕ_k by λ_k, the first two eigenvalues are $\lambda_{1,0}$ and $\lambda_{2,0}$.
- $\lambda_{0,1}$ corresponds to the $\lceil W/H \rceil$-th ($= 7$-th here) eigenvalue.
- ϕ_1, ϕ_2 are orthogonal, but not functionally independent.
- $\phi_1, \phi_{\lceil W/H \rceil}$ are functionally independent, therefore $\{1, \lceil W/H \rceil\}$ should be chosen.

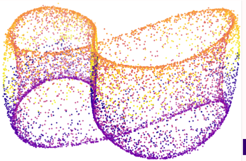


Situation when $\phi(\mathcal{M})$ can fail to be invertible

- *(Global) functional dependency*: **rank** $\mathbf{D}\phi < d$ on an open subset or all of $\mathcal{M}$ (yellow curve in top).
- The *knot*: **rank** $\mathbf{D}\phi < d$ at an isolated point (middle).
- The *crossing*: $\phi : \mathcal{M} \to \phi(\mathcal{M})$ is not invertible at $\mathbf{x}$, but $\mathcal{M}$ can be covered with open sets $\mathcal{U}$ such that the restriction $\phi : \mathcal{U} \to \phi(\mathcal{U})$ has full rank d (bottom).
- Existence of solution for LE/DM has been proved [Bat14].
  - However, s, the number of eigenfunctions needed, may exceed the *Whitney embedding dimension* ($\leqslant 2d$), and that s may depend on injectivity radius, aspect ratio, etc.

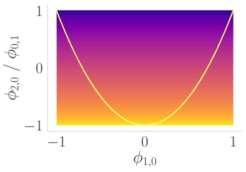


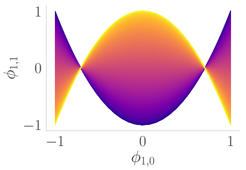


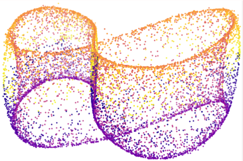


Situation when $\phi(\mathcal{M})$ can fail to be invertible

- *(Global) functional dependency*: $\mathrm{rank}\, D\phi < d$ on an open subset or all of $\mathcal{M}$ (yellow curve in top).
- The *knot*: $\mathrm{rank}\, D\phi < d$ at an isolated point (middle).
- The *crossing*: $\phi : \mathcal{M} \to \phi(\mathcal{M})$ is not invertible at $\mathbf{x}$, but $\mathcal{M}$ can be covered with open sets $\mathcal{U}$ such that the restriction $\phi : \mathcal{U} \to \phi(\mathcal{U})$ has full rank d (bottom).
- Existence of solution for LE/DM has been proved [Bat14].
  - However, s, the number of eigenfunctions needed, may exceed the *Whitney embedding dimension* ($\leqslant 2d$), and that s may depend on injectivity radius, aspect ratio, etc.







Riemannian metric

The pushforward Riemannian metric [PM13]

- Associate with $\phi(\mathcal{M})$ a *pushforward Riemannian metric* $g_{*\phi}$ that preserves the geometry of $(\mathcal{M}, g)$. Here $g_{*\phi}$ is defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle_{g_{*\phi}(\mathbf{x})} = \langle D\phi^{-1}(\mathbf{x})\mathbf{u}, D\phi^{-1}(\mathbf{x})\mathbf{v} \rangle_{g(\mathbf{x})}$$

$$\text{for all } \mathbf{u}, \mathbf{v} \in \mathcal{T}_{\phi(\mathbf{x})}\phi(\mathcal{M})$$

  - $\mathcal{T}_{\mathbf{x}}\mathcal{M}, \mathcal{T}_{\phi(\mathbf{x})}\phi(\mathcal{M})$ are tangent subspaces.
  - $D\phi^{-1}(\mathbf{x})$ maps vectors from $\mathcal{T}_{\phi(\mathbf{x})}\phi(\mathcal{M})$ to $\mathcal{T}_{\mathbf{x}}\mathcal{M}$.

- $g_{*\phi}(\mathbf{x}_i)$ in local coordinate is a PSD matrix $\mathbf{G}(i)$

$$\langle \mathbf{u}, \mathbf{v} \rangle_{g_{*\phi}(\mathbf{x}_i)} = \mathbf{u}^\top \mathbf{G}(i)\mathbf{v}$$

- Local Coordinate $\mathbf{U}(i)$ (tangent plane) on embedding $\mathbf{y}_i = \phi(\mathbf{x}_i)$ and distortion $\mathbf{\Sigma}(i)$ can be obtained by SVD of co-metric $\mathbf{H}(i)$ $=$ pseudo_inv($\mathbf{G}(i)$).
- Local coordinate $\mathbf{U}(i)$ projects onto coordinates set S is

$$\mathbf{U}_S(i) = \mathbf{U}(i)[S, :]$$

Algorithm 1: Riemannian metric estimation

RMetric($\mathbf{Y}, \mathbf{L}, d$)

for *all* $\mathbf{y}_i \in \mathbf{Y}, k = 1 \to m, l = 1 \to m$ **do**

  $[\tilde{\mathbf{H}}(i)]_{kl} = \sum_{j \neq i} L_{ij}(y_{jl} - y_{il})(y_{jk} - y_{ik})$

end

for $i = 1 \to n$ **do**

  $\mathbf{U}(i), \mathbf{\Sigma}(i) \leftarrow$ ReducedRankSVD($\tilde{\mathbf{H}}(i), d$)

  $\mathbf{H}(i) = \mathbf{U}(i)\mathbf{\Sigma}(i)\mathbf{U}(i)^\top$

  $\mathbf{G}(i) = \mathbf{U}(i)\mathbf{\Sigma}^{-1}(i)\mathbf{U}(i)^\top$

end

Return: $\mathbf{U}(i) \in \mathbb{R}^{m \times d}$ for $i \in [n]$

Loss function based on volume

Loss function

The loss function

$$\mathfrak{L}(S;\zeta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log\sqrt{\det\left(\mathbf{U}_S(i)^\top \mathbf{U}_S(i)\right)}}_{\mathfrak{R}_1(S)=\frac{1}{n}\sum_{i=1}^{n}\mathfrak{R}_1(S;i)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d} \log\|\mathbf{u}_k^S(i)\|_2}_{\mathfrak{R}_2(S)=\frac{1}{n}\sum_{i=1}^{n}\mathfrak{R}_2(S;i)} - \zeta\sum_{k\in S}\lambda_k \tag{1}$$

The chosen independent coordinates

$$S_*(\zeta) = \underset{S\subseteq[m];|S|=s;1\in S}{\operatorname{argmax}} \mathfrak{L}(S;\zeta)$$

Loss function

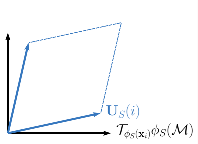
- The search space: $S_*(\zeta) = \text{argmax}_{S \subseteq [m]; |S|=s; 1 \in S}\, \mathfrak{L}(S; \zeta)$
  - S_* exists but cannot be computed analytically [Bat14].
  - Start with larger set $[m] = \{1, \cdots, m\}$ of eigenvector of $\mathbf{L}$, find coordinates $S \subseteq [m]$ with $|S| = s$ and force the slowest varying coordinate to **always** be chosen , i.e., $1 \in S$.
- $\mathfrak{R} = \mathfrak{R}_1 - \mathfrak{R}_2 = (\cdots)$
  - Projected volume of a unit parallelogram in $\mathcal{T}_{\phi_S(\mathbf{x}_i)}\phi_S(\mathcal{M})$

$$\text{Vol}(i; S) = \frac{\sqrt{\det\left(\mathbf{U}_S(i)^\top \mathbf{U}_S(i)\right)}}{\prod_{k=1}^{d} \|\mathbf{u}_k^S(i)\|_2}$$

  - Since ϕ_S is *not* an isometry $\to$ remove the local distortions $\mathbf{\Sigma}(i)$ introduced by ϕ from the estimated rank of ϕ at $\mathbf{x}$.
- Regularization term, consisting of the sum of eigenvalues $\sum_{k \in S} \lambda_k$ of the graph Laplacian $\mathbf{L}$, is added to penalize the high frequency coordinates.

Loss function

- The search space: $\mathcal{S}_*(\zeta) = \text{argmax}_{\mathcal{S}\subseteq[m];|\mathcal{S}|=s;1\in\mathcal{S}} \mathfrak{L}(\mathcal{S};\zeta)$
  - $\mathcal{S}_*$ exists but cannot be computed analytically [Bat14].
  - Start with larger set $[m] = \{1, \cdots, m\}$ of eigenvector of $\mathbf{L}$, find coordinates $\mathcal{S} \subseteq [m]$ with $|\mathcal{S}| = s$ and force the slowest varying coordinate to **always** be chosen , i.e., $1 \in \mathcal{S}$.
- $\mathfrak{R} = \mathfrak{R}_1 - \mathfrak{R}_2 = (\cdots)$
  - Projected volume of a unit parallelogram in $\mathcal{T}_{\phi_\mathcal{S}(\mathbf{x}_i)}\phi_\mathcal{S}(\mathcal{M})$



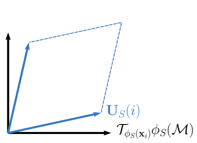
$$\text{Vol}(i;\mathcal{S}) = \frac{\sqrt{\det\left(\mathbf{U}_\mathcal{S}(i)^\top \mathbf{U}_\mathcal{S}(i)\right)}}{\prod_{k=1}^{d} \|\mathbf{u}_k^\mathcal{S}(i)\|_2}$$

  - Since $\phi_\mathcal{S}$ is *not* an isometry $\rightarrow$ remove the local distortions $\mathbf{\Sigma}(i)$ introduced by ϕ from the estimated rank of ϕ at $\mathbf{x}$.
- Regularization term, consisting of the sum of eigenvalues $\sum_{k\in\mathcal{S}} \lambda_k$ of the graph Laplacian $\mathbf{L}$, is added to penalize the high frequency coordinates.

Loss function

- The search space: $S_*(\zeta) = \text{argmax}_{S \subseteq [m]; |S| = s; 1 \in S} \mathfrak{L}(S; \zeta)$
  - S_* exists but cannot be computed analytically [Bat14].
  - Start with larger set $[m] = \{1, \cdots, m\}$ of eigenvector of $\mathbf{L}$, find coordinates $S \subseteq [m]$ with $|S| = s$ and force the slowest varying coordinate to **always** be chosen , i.e., $1 \in S$.
- $\mathfrak{R} = \mathfrak{R}_1 - \mathfrak{R}_2 = (\cdots)$
  - Projected volume of a unit parallelogram in $\mathcal{T}_{\phi_S(\mathbf{x}_i)} \phi_S(\mathcal{M})$

$$\text{Vol}(i; S) = \frac{\sqrt{\det\left(\mathbf{U}_S(i)^\top \mathbf{U}_S(i)\right)}}{\prod_{k=1}^{d} \|\mathbf{u}_k^S(i)\|_2}$$



  - Since ϕ_S is *not* an isometry $\rightarrow$ remove the local distortions $\mathbf{\Sigma}(i)$ introduced by ϕ from the estimated rank of ϕ at $\mathbf{x}$.
- Regularization term, consisting of the sum of eigenvalues $\sum_{k \in S} \lambda_k$ of the graph Laplacian $\mathbf{L}$, is added to penalize the high frequency coordinates.

Pseudo-code for brute-force search

Algorithm 2: Independent Eigencoordinates Search

IndEigenSearch($\mathbf{X}, \varepsilon, d, s, \zeta$)

$\mathbf{Y} \in \mathbb{R}^{n \times m}, \mathbf{L}, \lambda \in \mathbb{R}^m \leftarrow$ DiffMap($\mathbf{X}, \varepsilon$)

$\mathbf{U}(i), \cdots, \mathbf{U}(n) \leftarrow$ RMetric($\mathbf{Y}, \mathbf{L}, d$)

for $S \in \{S' \subseteq [m] : |S'| = s, 1 \in S'\}$ **do**

 $\mathfrak{R}_1(S) \leftarrow 0; \mathfrak{R}_2(S) \leftarrow 0$

 for $i = 1, \cdots, n$ **do**

  $\mathbf{U}_S(i) \leftarrow \mathbf{U}(i)[S, :]$

  $\mathfrak{R}_1(S) \mathrel{+}= \frac{1}{2n} \cdot \log\det\left(\mathbf{U}_S(i)^\top \mathbf{U}_S(i)\right)$

  $\mathfrak{R}_2(S) \mathrel{+}= \frac{1}{n} \cdot \sum_{k=1}^{d} \log \|u_k^S(i)\|_2$

 end

 $\mathfrak{L}(S; \zeta) = \mathfrak{R}_1(S) - \mathfrak{R}_2(S) - \zeta \sum_{k \in S} \lambda_k$

end

$S_* = \operatorname{argmax}_S \mathfrak{L}(S; \zeta)$

Return: Independent eigencoordinates set S_*

Limit of loss $\mathcal{L}$

Assumptions

1. The manifold $\mathcal{M}$ is compact of class $\mathcal{C}^3$, and there exists a set S, with $|S| = s$ so that ϕ_S is a smooth embedding of $\mathcal{M}$ in $\mathbb{R}^s$.
2. The data are sampled from a distribution on $\mathcal{M}$ continuous w.r.t. $\mu_{\mathcal{M}}$ with density p.
3. The estimate of $\mathbf{H}_S$ in Algorithm 1 computed w.r.t. the embedding ϕ_S is consistent.

Discussion

- From [Bat14] that Assumption 1 is satisfied for the LE/DM embedding.
- Assumptions 2, 3 are minimal requirements ensuring that limits of our quantities exist.

Let $j_S(\mathbf{y}) = 1/\operatorname{Vol}(\mathbf{U}_S(\mathbf{y})\mathbf{\Sigma}_S^{1/2}(\mathbf{y}))$ and $\tilde{j}_S(\mathbf{y}) = \prod_{k=1}^{d} \left(\|u_k^S(\mathbf{y})\|\sigma_k(\mathbf{y}))^{1/2}\right)^{-1}$, we study the limit of $\mathfrak{L}$ (Theorem 1 next page).

1. The manifold $\mathcal{M}$ is compact of class $\mathcal{C}^3$, and there exists a set S, with $|S| = s$ so that ϕ_S is a smooth embedding of $\mathcal{M}$ in $\mathbb{R}^s$.
2. The data are sampled from a distribution on $\mathcal{M}$ continuous w.r.t. $\mu_{\mathcal{M}}$ with density p.
3. The estimate of $\mathbf{H}_S$ in Algorithm 1 computed w.r.t. the embedding ϕ_S is consistent.

Discussion

- From [Bat14] that Assumption 1 is satisfied for the LE/DM embedding.
- Assumptions 2, 3 are minimal requirements ensuring that limits of our quantities exist.

Let $j_S(\mathbf{y}) = 1/\operatorname{Vol}(\mathbf{U}_S(\mathbf{y})\boldsymbol{\Sigma}_S^{1/2}(\mathbf{y}))$ and $\tilde{j}_S(\mathbf{y}) = \prod_{k=1}^{d} \left(\|u_k^S(\mathbf{y})\|\sigma_k(\mathbf{y}))^{1/2}\right)^{-1}$, we study the limit of $\mathfrak{L}$ (Theorem 1 next page).

Limit of $\mathfrak{L} - \mathfrak{R} = \mathfrak{R}_1 - \mathfrak{R}_2$

Theorem 1 (Limit of $\mathfrak{R}$)

Under Assumptions 1–3,

$$\lim_{n\to\infty} \frac{1}{n} \sum_i \ln \mathfrak{R}(S, \mathbf{x}_i) = \mathfrak{R}(S, \mathcal{M})$$

and

$$\mathfrak{R}(S, \mathcal{M}) = -\int_{\phi_S(\mathcal{M})} \ln \frac{j_S(\mathbf{y})}{\tilde{j}_S(\mathbf{y})} p(\phi_S^{-1}(\mathbf{y})) j_S(\mathbf{y}) \, d\mu_{\phi_S(\mathcal{M})}(\mathbf{y}) \coloneqq -D(pj_S \| p\tilde{j}_S)$$

- $D(\cdot\|\cdot)$ is a KL divergence, where the measures defined by $pj_S, p\tilde{j}_S$ normalize to different values.
- Because $j_S \geqslant \tilde{j}_S$ the divergence D is always positive

LIMIT OF $\mathfrak{L}$ — REGULARIZATION TERM

Spectral convergence of L [BN07, vLBB08]
The smoothness penalty converges to

$$\phi_k^\top \mathbf{L}\phi_k \to \int_{\mathcal{M}} \|\operatorname{grad}\phi_k(\mathbf{x})\|_2^2 d\mu(\mathcal{M}) \tag{2}$$

Since ϕ_k satisfies the Neumann boundary condition (for LE/DM).

Discussion on the extension to LLE, LTSA and HLLE

1. Unlike LE/DM, no theory has been developed for Assumption 1.
2. LLE, LTSA and HLLE converge to different differential operators (with different boundary conditions) [TJ18], one has to modify the regularization term in (2) to get a better estimate of *smoothness*.

Spectral convergence of L [BN07, vLBB08]
The smoothness penalty converges to

$$\phi_k^\top \mathbf{L}\phi_k \to \int_{\mathcal{M}} \|\operatorname{grad}\phi_k(\mathbf{x})\|_2^2 d\mu(\mathcal{M}) \tag{2}$$

Since ϕ_k satisfies the Neumann boundary condition (for LE/DM).

Discussion on the extension to LLE, LTSA and HLLE

1. Unlike LE/DM, no theory has been developed for Assumption 1.
2. LLE, LTSA and HLLE converge to different differential operators (with different boundary conditions) [TJ18], one has to modify the regularization term in (2) to get a better estimate of *smoothness*.

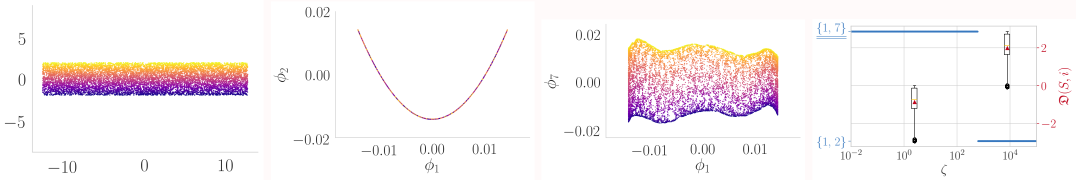
Experiments

Experiments

Synthetic datasets

2D LONG STRIP

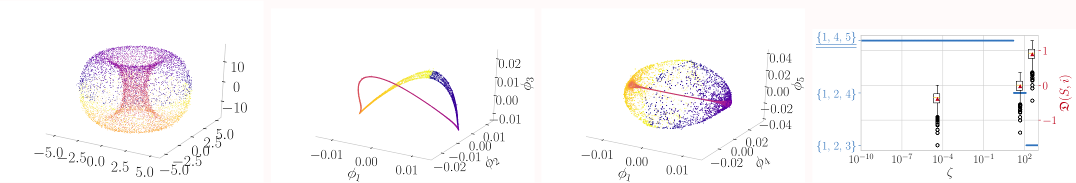
- The synthetic *2D long strip* with aspect ratio $W/H = 2\pi$.
- From the analysis before, the corresponding slowest varying unique eigendirections are $S_* = \{1, \lceil W/H \rceil\} = \{1, 7\}$.



(a) Original data $\mathbf{X}$ **(b)** $\phi_{[2]}$ **(c)** ϕ_{S_*} **(d)** Regularization path

High torus

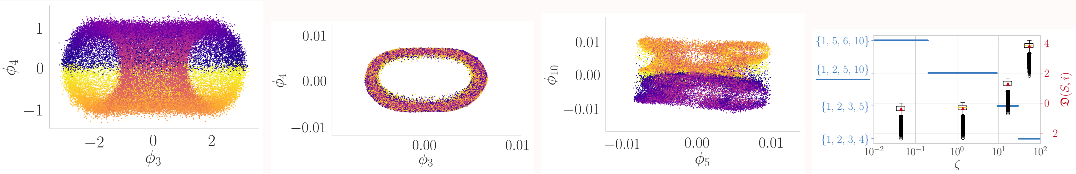
- The synthetic *High torus* dataset: example of the minimum embedding dimension s is greater than the intrinsic dimension d.
- $S_* = \{1, 4, 5\}$



(e) Original data $\mathbf{X}$ **(f)** $\phi_{[3]}$ **(g)** ϕ_{S_*} **(h)** Regularization path

Three torus

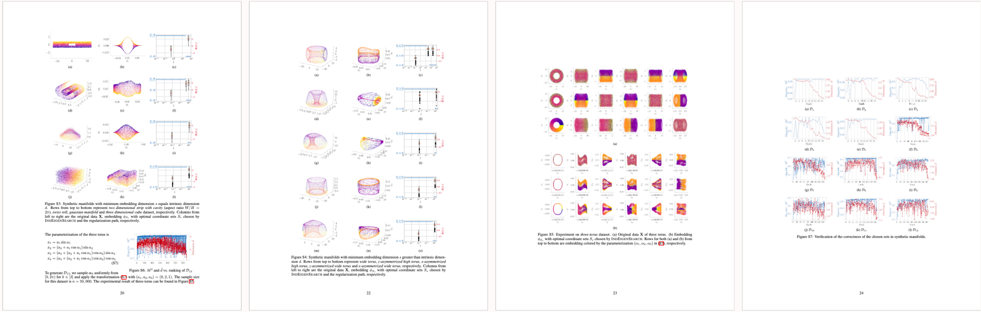
- The synthetic *Three torus* dataset: example of manifold having higher intrinsic dimension d, which cannot be visualized easily.
- $S_* = \{1, 2, 5, 10\}$



(i) Original data $\mathbf{X}_{\{3,4\}}$ **(j)** $\phi_{\{3,4\}}$ **(k)** $\phi_{S^*_{\{3,4\}}}$ **(l)** Regularization path

Extra experiments

Experiments on more synthetic datasets can be found on the paper [CM19].



Experiments

Real datasets

Overview

	n	D	deg_{avg}	(s, d)	t (sec)	S^*
SDSS [AAA+09]	299k	3750	144.91	(2, 2)	106.05	**(1, 3)**
Aspirin [CTS+17]	212k	244	101.03	(4, 3)	85.11	**(1, 2, 3, 7)**
Ethanol	555k	102	107.27	(3, 2)	233.16	**(1, 2, 4)**
Malondialdehyde	993k	96	106.51	(3, 2)	459.53	(1, 2, 3)
CH_3Cl [FTP16]	23k	34	91.84	(3, 2)	8.37	**(1, 4, 6)**

Selected eigenvectors ↑

Overview

	n	D	deg_{avg}	(s, d)	t (sec)	S^*
SDSS [AAA+09]	299k	3750	144.91	(2, 2)	106.05	**(1, 3)**
Aspirin [CTS+17]	212k	244	101.03	(4, 3)	85.11	**(1, 2, 3, 7)**
Ethanol	555k	102	107.27	(3, 2)	233.16	**(1, 2, 4)**
Malondialdehyde	993k	96	106.51	(3, 2)	459.53	(1, 2, 3)
CH_3Cl [FTP16]	23k	34	91.84	(3, 2)	8.37	**(1, 4, 6)**

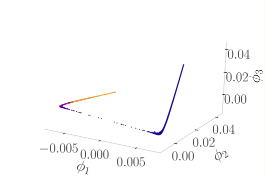
Selected eigenvectors ↑

Chloromethane molecular dynamics simulation [FTP16]

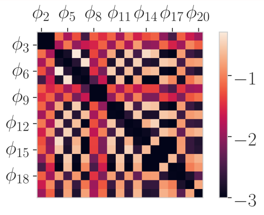
- MD simulation of the following reaction

$$CH_3Cl + Cl^- \leftrightarrow CH_3Cl + Cl^-$$

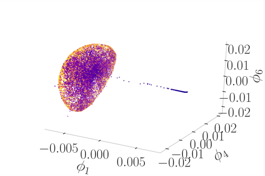
- Clusters, and a sparse connection between two clusters are visible.
- (m), (o) & (p) are colored by the distance between C and Cl, Cl, Cl, respectively.



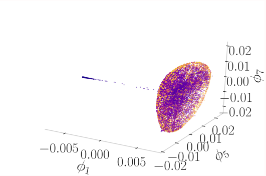
(m) $\phi_{[3]}$



(n) Heatmap of $\mathfrak{L}(\{1, i, j\})$



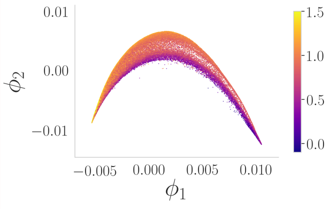
(o) ϕ_{S_*}



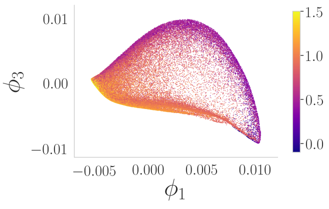
(p) ϕ_{S_2}

GALAXY SPECTRA FROM THE SDSS [AAA+09]

- Data can be downloaded here[3] and are preprocessed the same way as [MMVZ16].
- We sampled $n = 50,000$ points from the first 0.3 million points
  - correspond to closer galaxies.
- Embeddings are colored by the blue spectrum magnitude
  - correlated to the number of young stars in the galaxy.



(q) $\phi_{[2]}$



(r) ϕ_{S_*}

[3] http://sdss.org

Experiments

Application

Initializer for UMAP [MHM18]

The UMAP algorithm works as follow,

- Build a local fuzzy simplicial complex $\mathsf{SC}_k = (V, E, \Sigma_2, \cdots, \Sigma_k)$ from the data $\mathbf{X}$.
  - In their construction, only 1-skeleton of the simplicial set is considered in the loss function, so essentially it represents a graph $\mathsf{G} = (V, E) = \mathsf{SC}_1$.
- Initialize the embedding $\mathbf{Y}_0 \leftarrow \text{DiffusionMap}(\mathsf{G})$
- Optimize the following loss function by gradient descent.

$$\mathbf{Y}_* = \underset{\mathbf{Y}}{\text{argmin}}\, C\left(p(\mathbf{X}), q(\mathbf{Y}); \mathsf{SC}_1\right)$$

  - Here $C(p, q; \mathsf{SC}_1)$ is the cross entropy defined on the simplicial set SC_1
  - p, q is the transition probability computed on $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Initializer for UMAP [MHM18]

The UMAP algorithm works as follow,

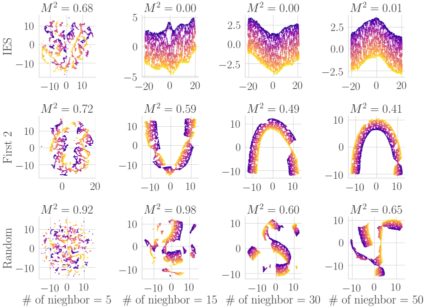
- Build a local fuzzy simplicial complex $SC_k = (V, E, \Sigma_2, \cdots, \Sigma_k)$ from the data $\mathbf{X}$.
  - In their construction, only 1-skeleton of the simplicial set is considered in the loss function, so essentially it represents a graph $G = (V, E) = SC_1$.
- Initialize the embedding $\mathbf{Y}_0 \leftarrow \text{DiffusionMap}(G)$
- Optimize the following loss function by gradient descent.

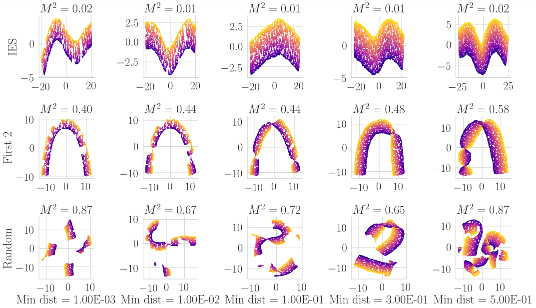
$$\mathbf{Y}_* = \operatorname*{argmin}_{\mathbf{Y}} C\left(p(\mathbf{X}), q(\mathbf{Y}); SC_1\right)$$

  - Here $C(p, q; SC_1)$ is the cross entropy defined on the simplicial set SC_1
  - p, q is the transition probability computed on $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Initializer for UMAP [MHM18]

- The bad initialization cannot always be fixed by more iterations.
  - In this simple example, it can. However, one needs way more iterations for it to converge.
- Figure below shows the experiment result of different initialization methods and choices of hyper-parameters with **fixed** iterations.





`kneigh`: # of neighbors in kNN graph.

`min_dist`: minimum separation in **Y**.

Related works

Related works

- Analysis on the sufficient conditions for failure [GZKR08].
  - Focuses on rectangles/cubes.
  - Failure defines as obtaining a mapping $\mathbf{Y} = \phi(\mathbf{X})$ that is not *affinely equivalent* w.r.t. original data $\mathbf{X}$.
- Functionally independent coordinates [DTCK18].
  - If ϕ_k is a repeated eigendirection of $\phi_1, \cdots, \phi_{k-1}$, one can fit ϕ_k with *local linear regression* (LLR) on predictors $\phi_{[k-1]}$ with low leave-one-out errors r_k.
  - Sequentially fit LLR on ϕ_k and obtain the coordinates with first few largest r_k's.
- Sequential spectral decomposition [GTW07, BM17].
  - Modifying the matrix $\mathbf{M}_k$ constructed for finding each k-th coordinate. ϕ_k can be obtained by the first min-eigenvector of $\mathbf{M}_k$.

- Analysis on the sufficient conditions for failure [GZKR08].
  - Focuses on rectangles/cubes.
  - Failure defines as obtaining a mapping $\mathbf{Y} = \phi(\mathbf{X})$ that is not *affinely equivalent* w.r.t. original data $\mathbf{X}$.
- Functionally independent coordinates [DTCK18].
  - If ϕ_k is a repeated eigendirection of $\phi_1, \cdots, \phi_{k-1}$, one can fit ϕ_k with *local linear regression* (LLR) on predictors $\phi_{[k-1]}$ with low leave-one-out errors r_k.
  - Sequentially fit LLR on ϕ_k and obtain the coordinates with first few largest r_k's.
- Sequential spectral decomposition [GTW07, BM17].
  - Modifying the matrix $\mathbf{M}_k$ constructed for finding each k-th coordinate. ϕ_k can be obtained by the first min-eigenvector of $\mathbf{M}_k$.

Related works

- Analysis on the sufficient conditions for failure [GZKR08].
  - Focuses on rectangles/cubes.
  - Failure defines as obtaining a mapping $\mathbf{Y} = \phi(\mathbf{X})$ that is not *affinely equivalent* w.r.t. original data $\mathbf{X}$.
- Functionally independent coordinates [DTCK18].
  - If ϕ_k is a repeated eigendirection of $\phi_1, \cdots, \phi_{k-1}$, one can fit ϕ_k with *local linear regression* (LLR) on predictors $\phi_{[k-1]}$ with low leave-one-out errors r_k.
  - Sequentially fit LLR on ϕ_k and obtain the coordinates with first few largest r_k's.
- Sequential spectral decomposition [GTW07, BM17].
  - Modifying the matrix $\mathbf{M}_k$ constructed for finding each k-th coordinate. ϕ_k can be obtained by the first min-eigenvector of $\mathbf{M}_k$.

Time complexity

- Time complexity is $\mathcal{O}(nm^{s+3}) \to$ brute force search for small s.
- $\mathfrak{R}_1, \mathfrak{R}_2$ in (1) are submodular set functions $\to$ optimizing over difference of submodular functions for large s.
- [DTCK18] has quadratic dependency on sample size n, see, e.g., empirical runtime on the right.
- [GTW07, BM17] in general has quadratic to cubic time complexity. Convergence depends on the condition number of the system and eigen-solver used.

- Time complexity is $\mathcal{O}(nm^{s+3}) \to$ brute force search for small s.

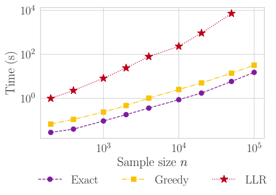
- $\mathfrak{R}_1, \mathfrak{R}_2$ in (1) are submodular set functions $\to$ optimizing over difference of submodular functions for large s.

- [DTCK18] has quadratic dependency on sample size n, see, e.g., empirical runtime on the right.

- [GTW07, BM17] in general has quadratic to cubic time complexity. Convergence depends on the condition number of the system and eigen-solver used.

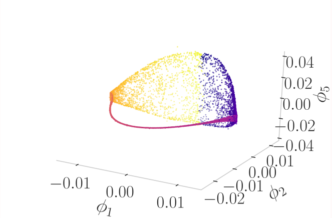
Time complexity

- Time complexity is $\mathcal{O}(nm^{s+3}) \to$ brute force search for small s.
- $\mathfrak{R}_1, \mathfrak{R}_2$ in (1) are submodular set functions $\to$ optimizing over difference of submodular functions for large s.
- [DTCK18] has quadratic dependency on sample size n, see, e.g., empirical runtime on the right.
- [GTW07, BM17] in general has quadratic to cubic time complexity. Convergence depends on the condition number of the system and eigen-solver used.

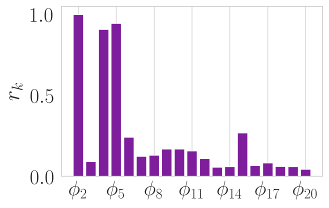


Comparison with [DTCK18]

- The embedding chosen by [DTCK18] is clearly shown to be suboptimal.
- This is because the algorithm searches in a sequential fashion; the noise eigenvector ϕ_2 in this example appears before the signal eigenvectors e.g., ϕ_4 and ϕ_5.



(u) ϕ_{S_*} by [DTCK18]



(v) Leave one out error

Conclusion

Conclusion

In this work, we

- Formulate the problem mathematically, show that a solution exists (for DM).
- Introduce a data driven loss $\mathfrak{L}$ and **Independent eigen-coordinates search (IES)** algorithm.
- Have experiments on real and synthetic data, showing the problem is pervasive.
- Analyze the limit of $\mathfrak{L}$ for $n \to \infty$.

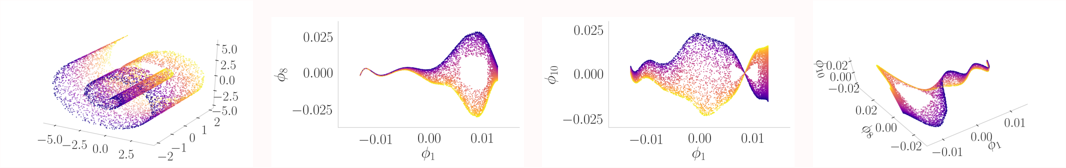
1. Extension of IES algorithm to LLE, LTSA & HLLE
    - Develop theories for Assumption 1.
    - Estimate gradient using coefficient Laplacian [TJ18].

2. Manifold optimization on the Grassmannian.
    - Instead of searching over fixed coordinates $S \subset [m]$, one can instead search over all possible projections.
    - $\mathfrak{L}$ will be a difference of convex function.

Future works

1. Extension of IES algorithm to LLE, LTSA & HLLE
   - Develop theories for Assumption 1.
   - Estimate gradient using coefficient Laplacian [TJ18].

2. Manifold optimization on the Grassmannian.
   - Instead of searching over fixed coordinates $S \subset [m]$, one can instead search over all possible projections.
   - $\mathfrak{L}$ will be a difference of convex function.



Thank you very much!

References I

[AAA+09] Kevork N Abazajian, Jennifer K Adelman-McCarthy, Marcel A Agüeros, Sahar S Allam, Carlos Allende Prieto, Deokkeun An, Kurt SJ Anderson, Scott F Anderson, James Annis, Neta A Bahcall, et al.
The seventh data release of the sloan digital sky survey.
The Astrophysical Journal Supplement Series, 182(2):543, 2009.

[Bat14] Jonathan Bates.
The embedding dimension of Laplacian eigenfunction maps.
Applied and Computational Harmonic Analysis, 37(3):516–530, November 2014.

[BM17] Yochai Blau and Tomer Michaeli.
Non-Redundant Spectral Dimensionality Reduction.
ArXiv, abs/1612.03412, 2017.

[BN03] Mikhail Belkin and Partha Niyogi.
Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.
Neural Computation, 15(6):1373–1396, June 2003.

References II

[BN07] Mikhail Belkin and Partha Niyogi.
Convergence of laplacian eigenmaps.
In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 129–136. MIT Press, 2007.

[BS16] Tyrus Berry and Timothy Sauer.
Consistent Manifold Representation for Topological Data Analysis.
arXiv:1606.02353 [math], June 2016.

[CL06] Ronald R. Coifman and Stéphane Lafon.
Diffusion maps.
Applied and Computational Harmonic Analysis, 21(1):5–30, July 2006.

[CM19] Yu-Chia Chen and Marina Meila.
Selecting the independent coordinates of manifolds with large aspect ratios.
arXiv:1907.01651 [cs, stat], July 2019.

References III

[CTS+17] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller.
Machine learning of accurate energy-conserving molecular force fields.
Science advances, 3(5):e1603015, 2017.

[DG03] David L. Donoho and Carrie Grimes.
Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data.
Proceedings of the National Academy of Sciences, 100(10):5591–5596, May 2003.

[DTCK18] Carmeline J Dsilva, Ronen Talmon, Ronald R Coifman, and Ioannis G Kevrekidis.
Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study.
Applied and Computational Harmonic Analysis, 44(3):759–773, 2018.

References IV

[FTP16] Kelly L. Fleming, Pratyush Tiwary, and Jim Pfaendtner.
New approach for investigating reaction dynamics and rates with ab initio calculations.
Jornal of Physical Chemistry A, 120(2):299–305, 2016.

[GTW07] Samuel Gerber, Tolga Tasdizen, and Ross Whitaker.
Robust Non-linear Dimensionality Reduction Using Successive 1-dimensional Laplacian Eigenmaps.
In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 281–288, New York, NY, USA, 2007. ACM.

[GZKR08] Yair Goldberg, Alon Zakai, Dan Kushnir, and Ya'acov Ritov.
Manifold learning: The price of normalization.
Journal of Machine Learning Research, 9(Aug):1909–1939, 2008.

[MHM18] Leland McInnes, John Healy, and James Melville.
Umap: Uniform manifold approximation and projection for dimension reduction.
arXiv preprint arXiv:1802.03426, 2018.

References V

[MMVZ16] James McQueen, Marina Meila, Jacob VanderPlas, and Zhongyue Zhang.
Megaman: Scalable manifold learning in python.
Journal of Machine Learning Research, 17(148):1–5, 2016.

[PM13] Dominique Perraul-Joncas and Marina Meila.
Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery.
arXiv:1305.7255 [stat], May 2013.

[RS00] Sam T. Roweis and Lawrence K. Saul.
Nonlinear Dimensionality Reduction by Locally Linear Embedding.
Science, 290(5500):2323–2326, December 2000.

[THJ11] Daniel Ting, Ling Huang, and Michael Jordan.
An Analysis of the Convergence of Graph Laplacians.
arXiv:1101.5435 [stat], January 2011.

References VI

[TJ18] Daniel Ting and Michael I. Jordan.
On Nonlinear Dimensionality Reduction, Linear Smoothing and Autoencoding.
arXiv:1803.02432 [stat], March 2018.

[vLBB08] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet.
Consistency of spectral clustering.
The Annals of Statistics, 36(2):555–586, April 2008.

[ZZ02] Zhenyue Zhang and Hongyuan Zha.
Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment.
SIAM Journal of Scientific Computing, 26:313–338, 2002.

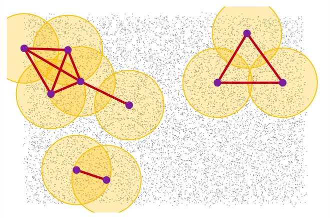
Backup slides

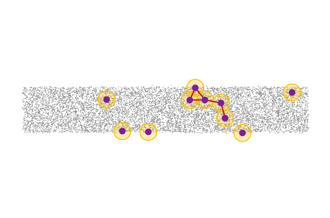
Backup slides

Some intuitions for IES problem

Some intuitions — I

- Two point clouds $\mathbf{X}_1 \overset{p}{\sim} \mathcal{M}_1, \mathbf{X}_2 \overset{p}{\sim} \mathcal{M}_2$ sampled from two manifold w.r.t. same density p.
- The neighborhood graph $G(V, E)$ should be built with similar ε.
- Short edges behave like noises.





It is possible to remove the defect by constructing a anisotropic kernel, however

- An isotropic kernel is needed for the convergence of graph Laplacian $\mathbf{L}$ [THJ11].
- Difficult to obtain/design such kernel since we do not know $\mathcal{M}$.

Some intuitions — II

Another way to think of it is to consider the Rayleigh quotient of the min-eigenvalue problem. The k-th minimum eigenvalue for graph Laplacian $\mathbf{L}$ is

$$\phi_k = \underset{\varphi \perp \phi_1 \cdots \phi_{k-1}; \|\varphi\|_2 = 1}{\text{argmin}} \varphi^\top \mathbf{L} \varphi = \underset{\varphi \perp \phi_1 \cdots \phi_{k-1}; \|\varphi\|_2 = 1}{\text{argmin}} \sum_{(i,j) \in E} (\varphi_i - \varphi_j)^2$$

- $\|\varphi\|_2 = 1$, in some sense it "normalizes" the manifold to equal aspect ratio.
- The density along the short edges are now sparser than the original density p.
- The term $(\varphi_i - \varphi_j)^2$ penalizes the function $\varphi(\cdot)$ parametrized short edges.

Backup slides

How to choose ζ

Regularization path and choosing ζ

- Define the *leave-one-out regret* of point i

$$\mathfrak{D}(S, i) = \mathfrak{R}(S_*^i; [n]\backslash\{i\}) - \mathfrak{R}(S; [n]\backslash\{i\})$$
$$\text{with } S_*^i = \text{argmax}_{S\subseteq[m];|S|=s;1\in S}\mathfrak{R}(S; i)$$

- $\mathfrak{D}(S, i)$ is the gain in $\mathfrak{R}$ if all the other points $[n]\backslash\{i\}$ choose the un-regularized optimal coordinates set in terms of point i.

- The optimal ζ' is then chosen by

$$\zeta' = \max_{\zeta\geqslant 0} \text{Percentile}\left(\{\mathfrak{D}(S_*(\zeta), i)\}_{i=1}^n, \alpha\right) \leqslant 0$$