Landmark Diffusion Speeds up the Alternating Diffusion Map

Sing-Yuan Yeh

Advised by Prof. Hau-Tieng Wu, Prof. Ronen Talmon & Prof. Mao-Pei Tsui

Data Science Degree Program, NTU

April 23 2024

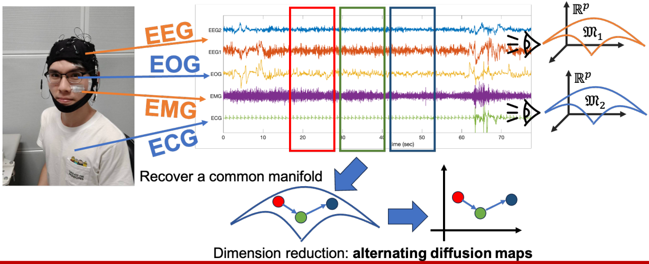
Table of Contents

Outline

Goal and Task

Goal

In application, the more we fuse physical signals, the more information we obtain. We hope to use more information to improve the accuracy of our models.



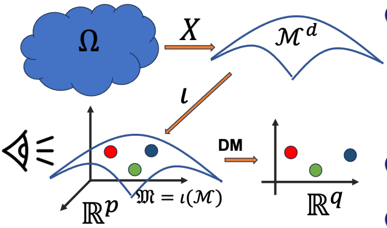
Task

Speed up alternating diffusion maps which capture the asymptotical behavior of AD.

Manifold setting and vanilla diffusion maps

Assumption: Assume the data is located on a manifold

1. $\iota : \mathcal{M} \to \mathbb{R}^p$ a smooth compact d-dim Riemanniam manifold.
2. $\mathcal{M}$-valued random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathcal{M}$ induced $\mu = X_* \mathbb{P}$.
3. Assume $d\mu$ is absolutely continuous w.r.t. dV. Denote p.d.f. $p = \frac{d\mu}{dV}$.



1. Sample n points $\{\iota(x_i)\}_{i=1}^n \subset \mathbb{R}^p$ i.i.d.
2. Construct the affinity matrix $W_{ij} = K\left(\frac{\|\iota(x_i)-\iota(x_j)\|}{\sqrt{\epsilon}}\right)$. e.g. $K(t) = \frac{1}{\sqrt{\pi}} e^{-t^2}$.
3. Let the degree matrix $D_{ii} = \sum_{j=1}^n W_{ij}$.
4. Define Markov matrix $M = D^{-1} W$.
5. Top q eigenvectors of M.

Behavior of vanilla diffusion maps

Coifman & Lafon (2006) and Singer (2006)

Suppose a function $f \in C^3(\mathcal{M})$ and $p \in C^2(\mathcal{M})$. Denote $\boldsymbol{f} \in \mathbb{R}^n$ with $\boldsymbol{f}_i = f(x_i)$. Then, with probability $1 - \mathcal{O}(n^{-2})$, we have

$$\left[\frac{D^{-1}W - I_n}{\epsilon}\boldsymbol{f}\right](i)$$

$$= \underbrace{\Delta f\left(x_i\right)}_{\text{spectral embedding}} + \frac{2\nabla f\left(x_i\right) \cdot \nabla p\left(x_i\right)}{p\left(x_i\right)} + \underbrace{\mathcal{O}(\epsilon^{1/2})}_{\text{Bias}} + \underbrace{\mathcal{O}\left(\frac{\sqrt{\log(n)}}{n^{1/2}\epsilon^{d/4+1/2}}\right)}_{\text{Variance}}$$

Finite spectral embedding

Almost isometric embedding via finite eigenfunctions ϕ_i where $\Delta\phi_i = -\lambda_i\phi_i$ with $0 = \lambda_0 \leq \lambda_1 \leq \cdots$.

Portegies (2015) and Berard, Besson & Gallot (1994)

Let $(\mathcal{M}, g)$ be a compact manifold. Fix $\epsilon > 0$. Then, there exists t_0 depending on manifold and ϵ such that for all $0 < t \leq t_0$ there exist N_0 depending on manifold, ϵ and t so that such that if $N > N_0$, the spectral embedding

$$x \mapsto 2t^{(d+2)/4}\sqrt{2}(4\pi)^{d/4}\left[\begin{array}{ccc} e^{-\lambda_1 t}\phi_1(x) & \ldots & e^{-\lambda_N t}\phi_N(x) \end{array}\right]^\top$$

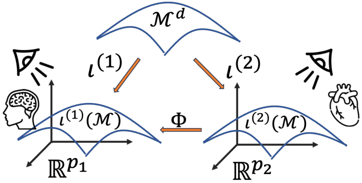
is almost isometric with the error controlled by ϵ

Outline

Alternating diffusion maps [Talmon & Wu]

Assumption: Assume the data is located on a manifold

1. $\iota^{(\ell)} : \mathcal{M} \to \mathbb{R}^{p_\ell}$ a smooth compact d-dim Riemanniam manifold.
2. Assume $d\mu$ is absolutely continuous w.r.t. $dV^{(\ell)}$. Denote p.d.f. $p^{(\ell)} = \frac{d\mu}{dV^{(\ell)}}$.



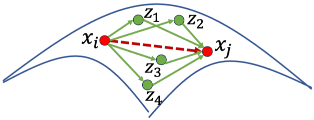
1. Sample n pairs $\left\{\left(\iota^{(1)}(x_i), \iota^{(2)}(x_i)\right)\right\}_{i=1}^{n}$.
2. Construct affinity matrices $W_{ij}^{(\ell)}$.
3. Let degree matrices $D_{ii}^{(\ell)} = \sum_{i=1}^{n} W_{ij}^{(\ell)}$.
4. Define alternative Markov matrix $M = D^{(1)^{-1}} W^{(1)} D^{(2)^{-1}} W^{(2)}$.

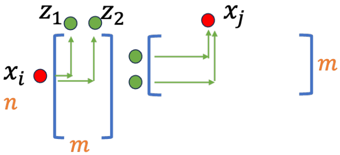
Computational complexity

The computational complexity and space complexity is $\mathcal{O}(n^3)$ where n is the number of data.

$$\underset{O(n^3)}{EVD}\left(\left[D^{(1)^{-1}}\right]\overset{O(n^3)}{\leftrightarrow}\left[W^{(1)}\right]\overset{O(n^3)}{\leftrightarrow}\left[D^{(2)^{-1}}\right]\overset{O(n^3)}{\leftrightarrow}\left[W^{(2)}\right]\right)$$

Landmark diffusion maps (Roseland) [Shen & Wu]





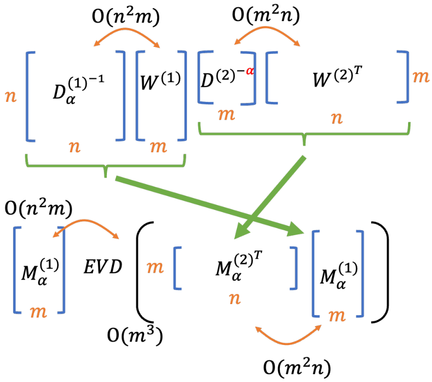
Denote $p_{\mathcal{Z}}$ is p.d.f. of landmark set $\mathcal{Z}$ with $|\mathcal{Z}| = m \ll n$.

$$\frac{1}{m}\sum_{k=1}^{m} \epsilon^{-d/2} K_\epsilon(x_i, z_k) K_\epsilon(z_k, x_j) = p_{\mathcal{Z}}(x_i) K_{2\epsilon}(x_i, x_j) + \mathcal{O}(\epsilon^{1/2}) + \mathcal{O}\left(\frac{\sqrt{\log(n)}}{n^{\beta/2}\epsilon^{d/4}}\right)$$

Landmark alternative diffusion maps (LAD)

1. Choose a landmark set $\mathcal{Z}$ with size m.
2. Build $n \times m$ affinity matrix $W^{(\ell)}_{ik} = K^{(\ell)}_{\epsilon}(x_i, z_k)$.
3. $m \times m$ matrix $D^{(2)}_{ii} = \texttt{diag}(W^{(2)} {W^{(2)}}^\top \mathbf{1}_m)$
4. $n \times m$ matrix $M^{(2)}_{\alpha} = W^{(2)} {D^{(2)}}^{-\alpha}$
5. $n \times n$ matrix $D^{(1)}_{\alpha;ii} = \texttt{diag}(W^{(1)} {M^{(2)}_{\alpha}}^\top \mathbf{1}_n)$
6. $n \times m$ matrix $M^{(1)}_{\alpha} = {D^{(1)}_{\alpha}}^{-1} W^{(1)}$
7. EVD on $m \times m$ matrix ${M^{(2)}_{\alpha}}^\top M^{(1)}_{\alpha} = V \Lambda V^{-1}$
8. $U = M^{(1)}_{\alpha} V$ and choose top q vectors as U_q
9. $e_i^\top U_q \Lambda_q^t \in \mathbb{R}^q$ as embedding point of (x_i, y_i) for all i.

Illustration of size of matrices



Remark

The singular vectors of $M^{(1)}M^{(2)^\top}$ are same as $M^{(1)}$ multiply eigenvectors of $M^{(2)^\top}M^{(1)}$.

Behavior of LAD

Main Theorem (Yeh, Wu, Talmon & Tsui)

Suppose a function $f \in C^3(\mathcal{M})$ and $p \in C^2(\mathcal{M})$. Let $q_\alpha(x) := \frac{p_{\mathcal{Z}}^{(2)}(x)^{1-\alpha}}{p^{(2)}(x)^\alpha}$.
Then, with probability $1 - \mathcal{O}(n^{-2})$, we have

$$
\begin{aligned}
&\frac{1}{\epsilon}\left[\left(I_n - \left(D_\alpha^{(1)}\right)^{-1} W_\alpha^{(1)} M_\alpha^{(2)}\right)\boldsymbol{f}\right](i) \\
&\quad = \frac{\mu_{2,0}^{(2)}}{2d}\Delta^{(2)} f(x_i) + \frac{\mu_{2,0}^{(1)}}{2d}\sum_{j=1}^{d}\lambda_j \nabla_{E_j E_j}^{(2)^2} f(x_i) \\
&\qquad + \frac{\mu_{2,0}^{(1)}}{d}\sum_{j=1}^{d}\lambda_j\left(\frac{\nabla_{E_j}^{(2)} p^{(2)}(x_i)}{p^{(2)}} + \frac{\nabla_{E_j}^{(2)} q_\alpha(x_i)}{q_\alpha(x_i)}\right)\nabla_{E_j}^{(2)} f(x_i) \\
&\qquad + \frac{\mu_{2,0}^{(2)}}{d}\frac{\nabla^{(2)} p^{(2)}(x_i)\cdot\nabla^{(2)} f(x_i)}{p^{(2)}(x_i)} + \underbrace{\mathcal{O}(\epsilon^{1/2})}_{\text{Bias}} + \underbrace{\mathcal{O}\left(\frac{\sqrt{\log(n)}}{n^{1/2}\epsilon^{d/4+1}}\right)}_{\text{Variance}}.
\end{aligned}
$$

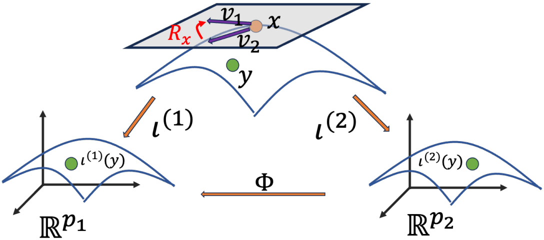
Rotation matrix

Consider a maps

$$\exp_x^{(1)^{-1}} \circ \iota^{(1)^{-1}} \circ \Phi \circ \iota^{(2)} \circ \exp_x^{(2)} : T_x\mathcal{M} \to T_x\mathcal{M}$$

Denote a matrix

$$R_x = \left[d\exp_x^{(1)}\Big|_0 \right]^{-1} \left[d\iota^{(1)} \right]^{-1} \nabla\Phi \left[d\iota^{(2)} \right] \left[d^{(2)} \exp_x^{(2)}\Big|_0 \right].$$



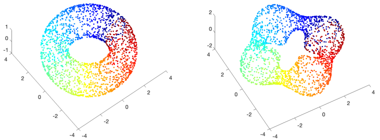
α-normalizer

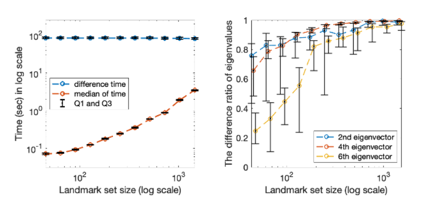
- Case 1 $\alpha = 0$: If $p^{(2)} = p^{(2)}_{\mathcal{Z}}$, $\iota^{(1)} = \iota^{(2)}$, then LAD is Roseland.
- Case 2 $\alpha = 1/2$: If $p^{(2)} = p^{(2)}_{\mathcal{Z}}$, then $q_\alpha = 1$ and LAD approach to AD.
- Case 3 $\alpha = 1$: LAD is independent of $p_{\mathcal{Z}}$.

Outline

Speed up AD

Sample 5000 pairs of data on the following two manifolds.





The x-axis represents the size of the landmark set, increasing from left to right.

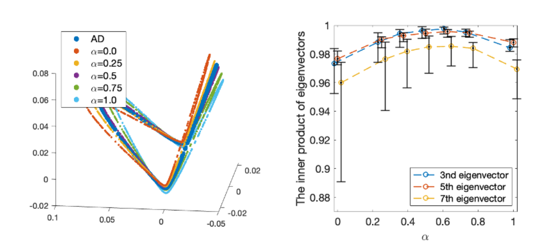
If we increase n to 1,000,000 and use LAD with $m = 1,000$, the computation time is 12.3 minutes.

Recover AD

Consider the canonical $\mathbb{S}^1$ and ellipse E, defined as

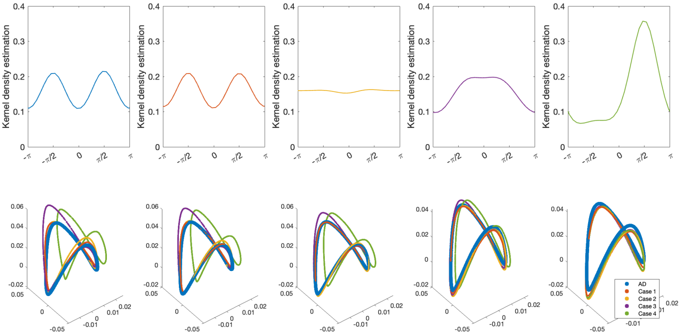
$$\mathbb{S}^1 = \{(\cos\theta, \sin\theta)\} \subset \mathbb{R}^2,$$
$$E = \{(2\cos\theta, \sin\theta)\} \subset \mathbb{R}^2,$$

where $\theta \in [-\pi, \pi)$. Sample 3000 pairs of points by non-uniformly p.d.f. $p^{(2)}(\theta) = \frac{1}{\pi}[\tan^{-1}(\frac{1}{2}\tan\theta)]$ and 1500 pairs of landmark points by $p^{(2)}_{\mathcal{Z}} = p^{(2)}$.



Independent of $p_{\mathcal{Z}}^{(2)}$

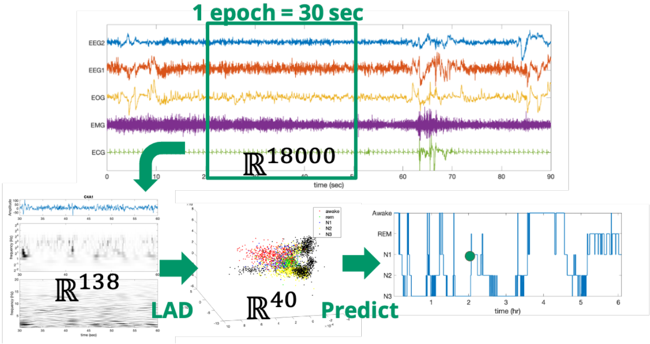
Sample 3000 pairs of points by non-uniformly p.d.f. $\frac{58}{50}[0.48\cos\theta + 0.52]$ and 1500 pairs of landmark points by following distribution (upper subfigures).



The lower five subfigures are $\alpha = 0, 0.25, 0.5, 0.75, 1$ from left to right.

Outline

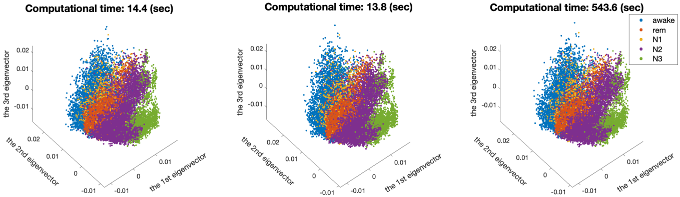
Methodology and data

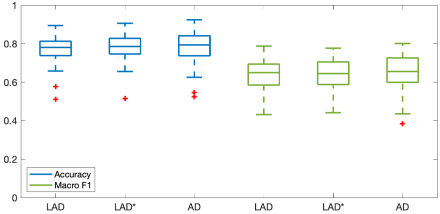


- 40 patients
- 240+ hours
- 5 labels (Awake, REM, N1, N2, N3)
- 29070 epochs

The dataset is from NCTS x TIDIS x CGMH

Embedding by AD and LAD



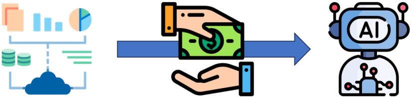


Boxplot results for leave-one-subject-out cross-validation of 40 patients.

From left to right is LAD, LAD*, AD.

Outline

Sample Complexity



With probability $1-\delta$, we have

Algorithm	Paper	Variance term
Vanilla DM	Dusun, Wu & Wu (2019)	$\mathcal{O}\left(\frac{\sqrt{-\log \epsilon}+\sqrt{-\log \delta}}{\sqrt{n}\epsilon^{d+2}}\right)$
LAD	in progress	??
VDM	struggle	??
Kernel-based MDP	Yeh et al. (2023)	$\mathcal{O}\left(\frac{\sqrt{-\log \delta}}{\sqrt{n}(1-\gamma)^{7/2}}\right)$

We are interested in the term the difference between ϕ_i and $v_{i,\epsilon,n}$, where ϕ_i is the eigenfunction of Δ and $v_{i,\epsilon,n}$ is the eigenvector of Vanilla DM.

Outline

Reference

[1] C. Shen and H.-T. Wu, *Scalability and robustness of spectral embedding: landmark diffusion is all you need*, (2022).

[2] R. Talmon and H.-T. Wu, *Latent common manifold learning with alternating diffusion: Analysis and applications*, (2019).

Thank You for Your Attention!

Appendix: Behavior of AD

Shen & Wu (2019)

Suppose $f \in C^3(\mathcal{M})$. Fix normal coordinates around x associated with $g^{(\ell)}$ so that $\{E_i\}_{i=1}^d \subset T_x\mathcal{M}$ o.n. associated with $g^{(2)}$. Consider the SVD of $R_x = U_x \Lambda_x V_x^T$, where $\Lambda_x = \operatorname{diag}[\lambda_1, \dots, \lambda_d]$. Then, when ϵ is sufficiently small, the AD starting from $g^{(2)}$ satisfies

$$
\begin{aligned}
T_\epsilon f(x) = f(x) &+ \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \left[\nabla^{(2)}_{E_i,E_i} f(x) + \frac{2\nabla^{(2)}_{E_i} f(x) \nabla^{(2)}_{E_i} p^{(2)}(x)}{p^{(2)}(x)} \right] \\
&+ \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \left[\Delta^{(2)} f(x) + \frac{2\nabla^{(2)} f(x) \cdot \nabla^{(2)} p^{(2)}(x)}{p^{(2)}(x)} \right] + O\left(\epsilon^{3/2}\right).
\end{aligned}
$$

Appendix: Proof

Landmark kernel

Take $f \in C^3(\mathcal{M}), 0 < \gamma < 1/2$ and $x, y \in \mathcal{M}$ so that $y = \exp_x^{(2)} v$, where $v \in T_x\mathcal{M}$ and $\|v\|_{g^{(2)}} \leq 2\epsilon^\gamma$. Then, when ϵ is sufficiently small, the following holds:

$$\begin{aligned}
&\int_{\mathcal{M}} K_\epsilon^{(1)}(x, z) K_\epsilon^{(2)}(z, y) F(z) \mathrm{d}V^{(2)}(z) \\
=&\epsilon^{d/2} \left[F(x) A_{0,\epsilon}(v) + \epsilon^{1/2} A_{1,\epsilon}(F, v) + \epsilon A_{2,\epsilon}(F, v) \right] + O\left(\epsilon^{d/2+3/2}\right),
\end{aligned}$$

Leading term of Landmark kernel

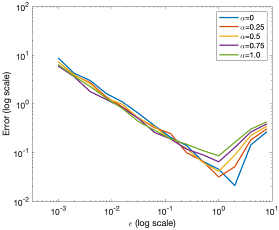
We remark that if the chosen kernels are both Gaussian, the α-landmark alternative kernel is Gaussian. Specifically, when $\tilde{K}^{(1)}$ and $\tilde{K}^{(2)}$ are both Gaussian, that is, $\tilde{K}^{(1)}(t) = \tilde{K}^{(2)}(t) = e^{-t^2}/\sqrt{\pi}$, we have

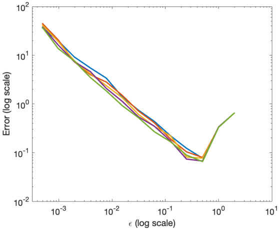
$$A_0(v) = \frac{\pi^{d/2}}{\sqrt{\det\left(I + \Lambda_x^2\right)}} e^{-\left\|\left(I+\Lambda_x^2\right)^{-1/2}\Lambda_x V_x^T\right\|^2/\epsilon},$$

which satisfies the exponential decay property of the kernel functions.

Appendix: Simulation results

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|\frac{1}{\epsilon}\left[\left(I_n-\left(\mathbf{D}_{\alpha,t}^{(1)}\right)^{-1}\mathbf{W}_{\alpha,t}^{(1)}\mathbf{M}_{\alpha,t}^{(2)}\right)\boldsymbol{f}\right](i)-\frac{f\left(x_i\right)-T_{\mathrm{lan},\epsilon,\alpha}f(x_i)}{\epsilon}\right|^2}$$





Left: $\mathbb{S}^1$ Right: $\mathbb{S}^2$.