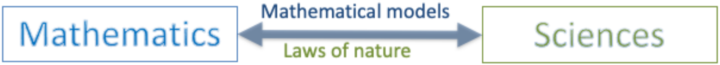
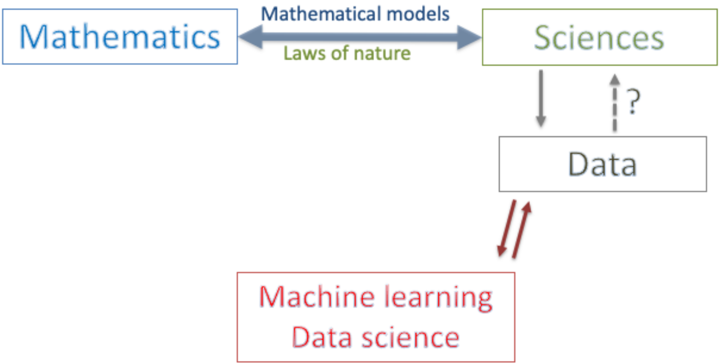
Manifold Learning, Explanations and Eigenflows

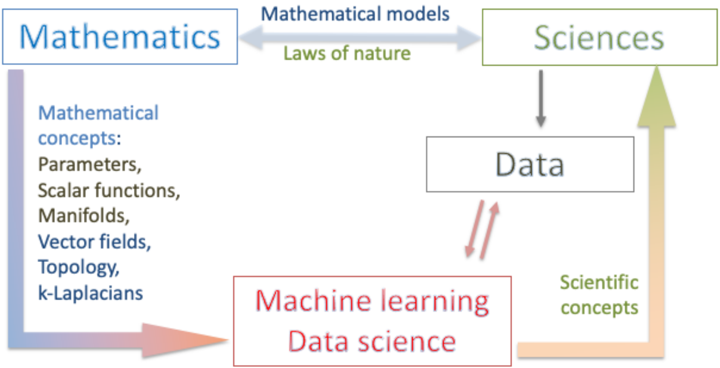
Marina Meilă
Yu-chia Chen, Samson Koelle, Hanyu Zhang and Ioannis Kevrekidis

University of Washington
mmp@stat.washington.edu

Non-Linear and High Dimensional Inference Workshop
Geometry and Statistics in Data Sciences
Institut Henri Poincaré
October 4, 2022


Mathematics
Mathematical models
Laws of nature
Sciences


Mathematics
Mathematical models
Laws of nature
Sciences
?
Data
Machine learning
Data science


Mathematics
Mathematical models
Laws of nature
Sciences
Mathematical concepts:
Parameters,
Scalar functions,
Manifolds,
Vector fields,
Topology,
k-Laplacians
Data
Machine learning
Data science
Scientific concepts

Outline

Manifold coordinates with Scientific meaning

Machine Learning 1-Laplacians, topology, vector fields

- 1-Laplacian $\Delta_1(\mathcal{M})$ estimation from samples
- Analysis of vector fields – Helmholtz-Hodge decomposition
- Harmonic Embedding Spectral Decomposition Algorithm
- Spectral Shortest Homologous Loop Detection

Outline

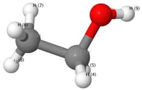
Manifold coordinates with Scientific meaning

Machine Learning 1-Laplacians, topology, vector fields

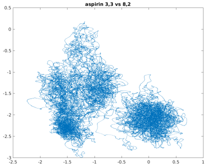
- 1-Laplacian $\Delta_1(\mathcal{M})$ estimation from samples
- Analysis of vector fields – Helmholtz-Hodge decomposition
- Harmonic Embedding Spectral Decomposition Algorithm
- Spectral Shortest Homologous Loop Detection

Motivation – understanding data from a Molecular Dynamics simulation

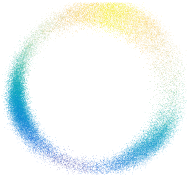
ethanol



original
data

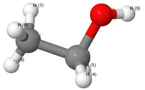


preprocessed

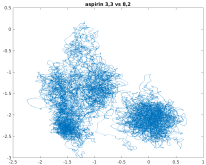


Motivation – understanding data from a Molecular Dynamics simulation

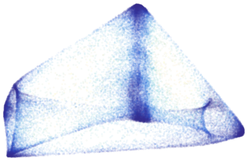
ethanol



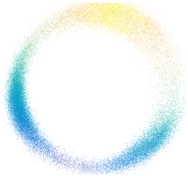
original data



after manifold learning

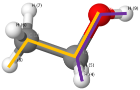


preprocessed

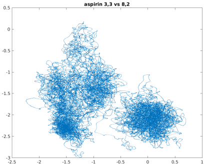


Motivation – understanding data from a Molecular Dynamics simulation

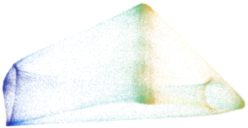
ethanol



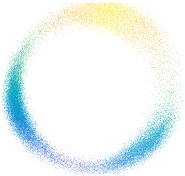
original data



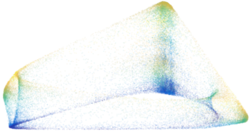
torsion 1



preprocessed



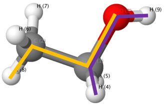
torsion 2



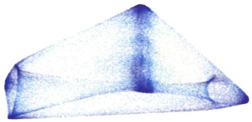
- 2 rotation angles (torsions) describe this manifold
- Can we discover these features automatically? Can we select these angles from a larger set of features with physical meaning?

scientific language (torsions)

data driven coordinates (from DiffMaps, Isomap)



$+$

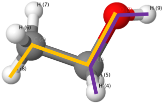
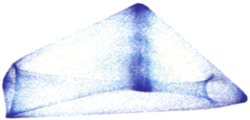
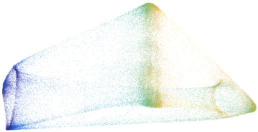


$=$

Idea Replace data driven coordinates with selected torsions

- Scientist: proposes a dictionary $\mathcal{G}$ with all variables of interest
- ML algorithm: outputs embedding ϕ,
- ManifoldLasso: finds new coordinates in $\mathcal{G}$ "equivalent" with ϕ $\leftarrow$ our algorithm

▶ Explanation
  ▶ $=$ find manifold coordinates from among scientific variables of interest
  ▶ in the language of the domain

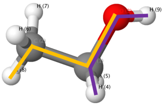
scientific language (torsions)		data driven coordinates (from DiffMaps, Isomap)		coordinates with scientific interpretation (selected torsions)
	$+$		$=$	

Idea Replace data driven coordinates with selected torsions

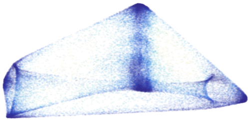
- Scientist: proposes a dictionary $\mathcal{G}$ with all variables of interest
- ML algorithm: outputs embedding ϕ,
- ManifoldLasso: finds new coordinates in $\mathcal{G}$ "equivalent" with ϕ $\leftarrow$ our algorithm

- Explanation
  - $=$ find manifold coordinates from among scientific variables of interest
  - in the language of the domain

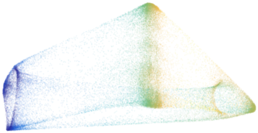
scientific language (torsions)	data driven coordinates (from DiffMaps, Isomap)	coordinates with scientific interpretation (selected torsions)



$+$



$=$

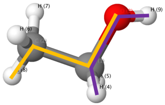


Idea Replace data driven coordinates with selected torsions

- Scientist: proposes a dictionary $\mathcal{G}$ with all variables of interest
- ML algorithm: outputs embedding ϕ,
- ManifoldLasso: finds new coordinates in $\mathcal{G}$ "equivalent" with ϕ $\leftarrow$ our algorithm

- Explanation
  - $=$ find manifold coordinates from among scientific variables of interest
  - in the language of the domain

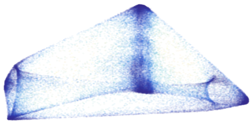
scientific language (torsions)



$\mathcal{G} = \{\text{functions } g\}$

$+$

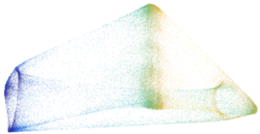
data driven coordinates (from DiffMaps, Isomap)



ϕ

$=$

coordinates with scientific interpretation (selected torsions)



$\{g_{j_1}, \ldots g_{j_d}\} \equiv g_S \subset \mathcal{G}$

Idea Replace data driven coordinates with selected torsions

- Scientist: proposes a dictionary $\mathcal{G}$ with all variables of interest
- ML algorithm: outputs embedding ϕ,
- ManifoldLasso: finds new coordinates in $\mathcal{G}$ "equivalent" with ϕ $\leftarrow$ our algorithm

- Explanation
  - $=$ find manifold coordinates from among scientific variables of interest
  - in the language of the domain

Idea: Sparse regression in function space

$$\underset{\text{manifold coordinates}}{\phi} = \overset{\text{homeomorphism}}{h} \circ \underset{\substack{\text{functions from } \mathcal{G} \\ \text{(new coordinates)}}}{g_S}$$

$$D\phi = DhDg_S$$

Leibnitz Rule

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

- sparse linear regression problem
- For every data i
  - $Y_i = \operatorname{grad} \phi(\xi_i)$,
  - $\mathbf{X}_i = \operatorname{grad} g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i\beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_{\beta} J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots j_s}$ from $\mathcal{G}$

Idea: Sparse regression in function space

$$\phi = h \circ g_S$$

manifold coordinates — functions from $\mathcal{G}$ (new coordinates)

$$D\phi = DhDg_S$$

Leibnitz Rule

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

- sparse linear regression problem
- For every data i
  - $Y_i = \text{grad}\,\phi(\xi_i)$,
  - $\mathbf{X}_i = \text{grad}\,g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i \beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_\beta J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\text{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots,j_s}$ from $\mathcal{G}$

Idea: Sparse regression in function space

$$\underbrace{\phi}_{\text{manifold coordinates}} = h \circ \underbrace{g_S}_{\text{functions from } \mathcal{G} \text{ (new coordinates)}}$$

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

$$D\phi = DhDg_S$$

Leibnitz Rule

- sparse linear regression problem
- For every data i
  - $Y_i = \text{grad}\,\phi(\xi_i)$,
  - $\mathbf{X}_i = \text{grad}\,g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i \beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_\beta J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\text{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots,j_s}$ from $\mathcal{G}$

Idea: Sparse regression in function space

$$\underset{\text{manifold coordinates}}{\phi} = h \circ \underset{\substack{\text{functions from } \mathcal{G} \\ \text{(new coordinates)}}}{g_S}$$

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

$$D\phi = DhDg_S$$

Leibnitz Rule

- sparse linear regression problem
- For every data i
  - $Y_i = \text{grad}\,\phi(\xi_i)$,
  - $\mathbf{X}_i = \text{grad}\,g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i\beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_\beta J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n}||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda\sum_j||\beta_j||, \quad (\text{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots,j_s}$ from $\mathcal{G}$

Idea: Sparse regression in function space

$$\phi = h \circ g_S$$

manifold coordinates — functions from $\mathcal{G}$ (new coordinates)

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

Functional (Group) Lasso

- optimize

$$\min_\beta J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots j_s}$ from $\mathcal{G}$

$$D\phi = DhDg_S$$

Leibniz Rule

- sparse linear regression problem
- For every data i
  - $Y_i = \operatorname{grad}\phi(\xi_i)$,
  - $\mathbf{X}_i = \operatorname{grad} g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i\beta_i$
- Constraint: subset S is same for all i

Idea: Sparse regression in function space

$$\underset{\text{manifold coordinates}}{\phi} = h \circ \underset{\text{functions from } \mathcal{G} \text{ (new coordinates)}}{g_S}$$

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

$$D\phi = DhDg_S$$

Leibnitz Rule

- sparse linear regression problem
- For every data i
  - $Y_i = \text{grad}\, \phi(\xi_i)$,
  - $\mathbf{X}_i = \text{grad}\, g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i \beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_{\beta} J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad \text{(ManifoldLasso)}$$

- support S of β selects $g_{j_1,\ldots j_s}$ from $\mathcal{G}$

Idea: Sparse regression in function space

$$\underset{\text{manifold coordinates}}{\phi} = h \circ \underset{\text{functions from } \mathcal{G} \text{ (new coordinates)}}{g_S}$$

Challenges

- sparse, non-linear regression problem
- ML coordinates ϕ defined up to diffeomorphism
- hence, h cannot take parametric form
- we cannot choose a basis for h
- will not ϕ_k depends on single g_j
- will not assume ϕ isometric

$$D\phi = DhDg_S$$

Leibnitz Rule

- sparse linear regression problem
- For every data i
  - $Y_i = \text{grad}\,\phi(\xi_i)$,
  - $\mathbf{X}_i = \text{grad}\, g_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial g_j}(\xi_i)$
  - Sparse linear system $Y_i = \mathbf{X}_i\beta_i$
- Constraint: subset S is same for all i

Functional (Group) Lasso

- optimize

$$\min_{\beta} J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\boldsymbol{\beta}_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

- support S of β selects $g_{j_1,\ldots j_s}$ from $\mathcal{G}$

ManifoldLasso Algorithm

Given Data $\xi_{1:n}$, $\dim \mathcal{M} = d$, embedding $\phi(\xi_{1:n})$, dictionary $\mathcal{G} = \{g_{1:p}\}$

1. Estimate tangent subspace at ξ_i by (weighted) PCA
2. Project dictionary functions gradients ∇g_j on tangent subspace, obtain $\mathbf{X}_{1:n} \in \mathbb{R}^{d \times p}$
3. Estimate gradients of $\phi_{1:k}$, obtain $Y_{1:n} \in \mathbb{R}^{d \times m}$

   By pull-back from embedding space ϕ
4. Solve GroupLasso($Y_{1:n}$, $\mathbf{X}_{1:n}$, d), obtain support S

$$\min_{\beta} J_\lambda(\beta) \;=\; \tfrac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

Output S

ManifoldLasso Algorithm

Given Data $\xi_{1:n}$, $\dim \mathcal{M} = d$, embedding $\phi(\xi_{1:n})$, dictionary $\mathcal{G} = \{g_{1:p}\}$

1. Estimate tangent subspace at ξ_i by (weighted) PCA
2. Project dictionary functions gradients ∇g_j on tangent subspace, obtain $\mathbf{X}_{1:n} \in \mathbb{R}^{d \times p}$
3. Estimate gradients of $\phi_{1:k}$, obtain $Y_{1:n} \in \mathbb{R}^{d \times m}$

   By pull-back from embedding space ϕ
4. Solve GroupLasso($Y_{1:n}, \mathbf{X}_{1:n}, d$), obtain support S

$$\min_{\beta} J_\lambda(\beta) \;=\; \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i \beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad \text{(ManifoldLasso)}$$

Output S

ManifoldLasso Algorithm

Given Data $\xi_{1:n}$, dim $\mathcal{M} = d$, embedding $\phi(\xi_{1:n})$, dictionary $\mathcal{G} = \{g_{1:p}\}$

1. Estimate tangent subspace at ξ_i by (weighted) PCA
2. Project dictionary functions gradients ∇g_j on tangent subspace, obtain $\mathbf{X}_{1:n} \in \mathbb{R}^{d \times p}$
3. Estimate gradients of $\phi_{1:k}$, obtain $Y_{1:n} \in \mathbb{R}^{d \times m}$

   By pull-back from embedding space ϕ
4. Solve GroupLasso($Y_{1:n}, \mathbf{X}_{1:n}, d$), obtain support S

$$\min_{\beta} J_\lambda(\beta) = \frac{1}{2}\sum_{i=1}^{n} ||Y_i - \mathbf{X}_i\beta_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad \text{(ManifoldLasso)}$$

Output S

ManifoldLasso Algorithm

Given Data $\xi_{1:n}$, $\dim \mathcal{M} = d$, embedding $\phi(\xi_{1:n})$, dictionary $\mathcal{G} = \{g_{1:p}\}$

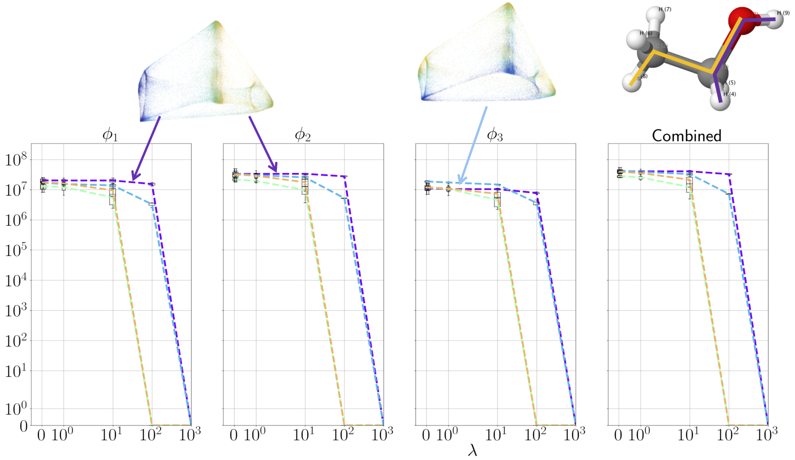
1. Estimate tangent subspace at ξ_i by (weighted) PCA
2. Project dictionary functions gradients ∇g_j on tangent subspace, obtain $\mathbf{X}_{1:n} \in \mathbb{R}^{d \times p}$
3. Estimate gradients of $\phi_{1:k}$, obtain $Y_{1:n} \in \mathbb{R}^{d \times m}$

   By pull-back from embedding space ϕ
4. Solve GroupLasso($Y_{1:n}, \mathbf{X}_{1:n}, d$), obtain support S

$$\min_{\beta} J_\lambda(\beta) \;=\; \tfrac{1}{2} \sum_{i=1}^{n} ||Y_i - \mathbf{X}_i \boldsymbol{\beta}_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

Output S

Ethanol MD simulation



regularization paths $\beta_{1:p}$ vs λ

Theory

- When is S unique? / When can $\mathcal{M}$ be uniquely parametrized by $\mathcal{G}$?
  Functional independence conditions on dictionary $\mathcal{G}$ and subset $g_{j_1,\ldots j_s}$
- Basic result

  $f_S = h \circ f_{S'}$ on U iff

$$\operatorname{rank}\begin{pmatrix} Df_S \\ Df_{S'} \end{pmatrix} = \operatorname{rank} Df_{S'} \quad \text{on } U$$

- When can GLasso recover S ?
  (Simple) Incoherence Conditions

$$\mu = \max_{i=1:n, j\in S, j'\notin S} \frac{|\mathbf{X}_{ji}^T \mathbf{X}_{j'i}|}{||\mathbf{X}_{ji}||\,||\mathbf{X}_{j'i}||} \quad \nu = \frac{1}{\min_{i=1:n} ||\mathbf{X}_{iS}^T \mathbf{X}_{iS}||_2} \quad nd\sigma^2 = \sum_{i,k} \epsilon_{ik}^2$$

Theorem If, $||\mathbf{X}_{1:p}|| = 1$, $\mu\nu\sqrt{d} + \frac{\sigma\sqrt{nd}}{\lambda} < 1$ then $\beta_j = 0$ for $j \notin S$.

Theory

- When is S unique? / When can $\mathcal{M}$ be uniquely parametrized by $\mathcal{G}$?
  Functional independence conditions on dictionary $\mathcal{G}$ and subset $g_{j_1,\dots j_s}$
- Basic result

  $f_S = h \circ f_{S'}$ on U iff

$$\text{rank}\begin{pmatrix} Df_S \\ Df_{S'} \end{pmatrix} = \text{rank}\, Df_{S'} \quad \text{on } U$$

- When can GLasso recover S ?
  (Simple) Incoherence Conditions

$$\mu = \max_{i=1:n, j \in S, j' \notin S} \frac{|\mathbf{X}_{ji}^T \mathbf{X}_{j'i}|}{\|\mathbf{X}_{ji}\| \|\mathbf{X}_{j'i}\|} \quad \nu = \frac{1}{\min_{i=1:n} ||\mathbf{X}_{iS}^T \mathbf{X}_{iS}||_2} \quad nd\sigma^2 = \sum_{i,k} \epsilon_{ik}^2$$

<u>Theorem</u> If, $\|\mathbf{X}_{1:p}\| = 1$, $\mu\nu\sqrt{d} + \frac{\sigma\sqrt{nd}}{\lambda} < 1$ then $\beta_j = 0$ for $j \notin S$.

Recovery for ManifoldLasso

Theorem 7 (Support recovery) *Assume that equation (30) holds, and that* $\sum_{i=1}^{n} ||x_{ij}||^2 = \gamma_j^2$ *for all* $j = 1 : p$. *Let* $\gamma_{\max} = \max_{j \notin S} \gamma_j$, $\kappa_S = \max_{i=1:n} \frac{\max_{j \in S} \|x_{ij}\|}{\min_{j \in S} \|x_{ij}\|}$. *Denote by* $\bar{\beta}$ *the solution of (31) for some* $\lambda > 0$. *If* $1 - (s-1)\mu > 0$ *and*

$$\gamma_{\max} \left(\frac{\mu}{1-(s-1)\mu} \frac{\kappa_S}{\min_{i=1}^{n} \min_{j' \in S} \|x_{ij'}\|} + \frac{\sigma\sqrt{d}}{\lambda\sqrt{n}} \right) \leq 1 \tag{37}$$

then $\bar{\beta}_{ij} = 0$ *for* $j \notin S$ *and all* $i = 1, \ldots n$.

Corollary 8 *Assume that equation* (31) *and condition (37) hold. Let* $\kappa = \frac{\mu}{1-(s-1)\mu} \frac{\kappa_S}{\min_{i=1}^{n} \min_{j' \in S} \|x_{ij'}\|}$ *and* $\gamma_S = \|\bar{\bar{X}}_S\|$. *Denote by* $\hat{\beta}$ *the solution to problem* (31) *for some* $\lambda > 0$. *If (1)* $\lambda = c\frac{\gamma_{\max}\sigma\sqrt{d}}{1-\kappa\gamma \max}$, $c > 1$, *and (2)* $||\beta_j^*|| > \sigma\sqrt{d}(\gamma_{\max} + \gamma_S) + \lambda(1+\sqrt{s})$ *for all* $j \in S$, *then the support* S *is recovered exactly and*

$$||\hat{\beta}_j - \beta_j^*|| < \sigma\sqrt{d}(\gamma_{\max} + \gamma_S) + \lambda(1+\sqrt{s}) = \sigma\sqrt{d}\gamma_{\max}\left[1 + \gamma_S/\gamma_{\max} + c\frac{1+\sqrt{s}}{1-\kappa\gamma_{\max}}\right] \quad \text{for all } j \in S.$$

TangentSpaceLasso: ManifoldLasso without embedding

Simplification regress basis of $\mathcal{T}_\xi\mathcal{M}$ on gradients of g_j

Proposition 2 (after (**?**)). *Let $\mathcal{F}, f_j$ be dictionary and dictionary functions on the d–dimensional smooth manifold $\mathcal{M}$. Assume $f_j \in C^\ell$ with $\ell \geq d+1$. Suppose $S \subset [p]$, and denote by* $\operatorname{grad} f_S$ *the $\mathbb{R}^{d\times s}$ matrix of concatenated* $\operatorname{grad} f_j : f \in S$*. Then, if there is a subset $S' \subsetneq S$ such that the following rank condition holds globally:*

$$\operatorname{rank}\begin{pmatrix}\operatorname{grad} f_S \\ \operatorname{grad} f_{S'}\end{pmatrix} = \operatorname{rank}\operatorname{grad} f_{S'}\ . \quad (4)$$

Then there exists a function h which is C^ℓ almost everywhere in the image of $f_{S'}(\mathcal{M})$ such that $f_S = h \circ f_{S'}$

$$\mu_S = \sup_{\xi\in\mathcal{M}^\circ, j\in S, j'\notin S} |\mathbf{X}^T_{\{j\},\xi}\mathbf{X}_{\{j'\},\xi}| \quad (5)$$

$$\nu_S = \sup_{\xi\in\mathcal{M}^\circ \alpha\in\mathbb{R}^d:||\alpha||_2=1} \alpha^T(\mathbf{X}^T_{S,\xi}\mathbf{X}_{S,\xi})^{-1}\alpha. \quad (6)$$

Proposition 3. *Assume that*

1. *$\mathcal{M}$ is d–dimensional C^k compact manifold with strictly positive reach.*
2. *Data ξ are sampled from some density p on $\mathcal{M}$ with $p > 0$ all over $\mathcal{M}$.*
3. *$\xi \in \mathcal{M}^\circ$ with probability 1 under p.*

Let S be the 'true' support, $S(\widehat{\mathbf{B}})$ be the support selected by TSLasso*, μ_S and ν_S be defined by (5) and (6), and further assume*

4. *$|S| = d$.*
5. *Df_S has rank d on $\mathcal{M}^\circ$,*
6. *$\mu_S\nu_S d < 1$.*

Then if we adapt the tangent space estimation algorithm in (**?**) *with bandwidth choice $h = O(\log n/(n-1))^d$, with $n \geq ((1-\mu_S\nu_S d)/2\nu_S d)^{d/(k-1)}$ we have*

$$Pr(S(\widehat{\mathbf{B}}) \subset S) \geq 1 - O\left((\frac{1}{n})^{\frac{k}{d}}\right)\ .$$

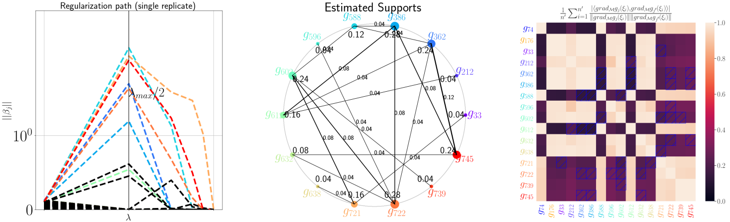
Experiments

Dataset	n	N_a	D	d	ϵ_N	m	n'	p	
SwissRoll	10000	NA	51	2	.18	2	100	51	synthetic
RigidEthanol	10000	9	50	2	3.5	3	100	12	
Ethanol	50000	9	50	2	3.5	3	100	12	skeleton $\mathcal{G}$
Malonaldehyde	50000	9	50	2	3.5	3	100	12	
Toluene	50000	16	50	1	1.9	2	100	30	
Ethanol	50000	9	50	2	3.5	3	100	756	exhaustive $\mathcal{G}$
Malonaldehyde	50000	9	50	2	3.5	3	100	756	
	ϕ						LASSO	$\lvert\mathcal{G}\rvert$	

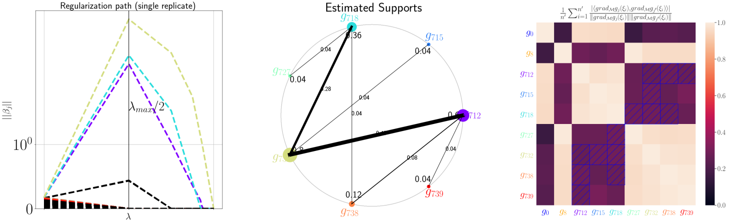
p = dictionary size, m = embedding dimension, n = sample size for manifold estimation, n' = sample size for MANIFOLDLASSO

Two-stage sparse recovery for exhaustive $\mathcal{G}$, $p = 756$

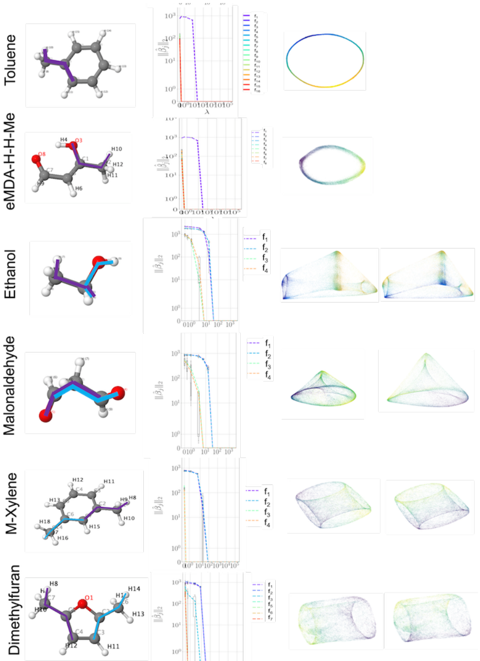
Ethanol



Malonaldehyde



Tangent Space Lasso experiments

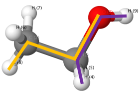


Summary of ManifoldLasso/FunctionalLasso

Technical contribution

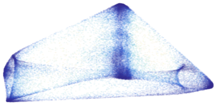
- non-linear sparse regression in function spaces
- Method to push/pull vectors through mappings ϕ
- ManifoldLasso: regression of data driven coordinates $\phi_{1:m}$ on domain-specific functions $\mathcal{G} = \{g_{1:p}\}$

scientific language (torsions)



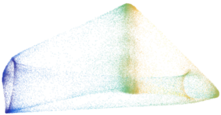
$+$

data driven coordinates



$=$

interpretable coordinates



- explains large scale structure with domain-relevant functions
- non-parametric; different from symbolic regression [Brunton et al. 2016, Rudy et al. 2019]
- transmissible knowledge, compare embeddings from different experiments
- extensions: estimated ∇g, simultaneous explanation of multiple manifolds

Summary of ManifoldLasso/FunctionalLasso

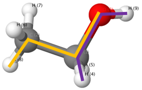
Technical contribution

- non-linear sparse regression in function spaces
- Method to push/pull vectors through mappings ϕ
- ManifoldLasso: regression of data driven coordinates $\phi_{1:m}$ on domain-specific functions $\mathcal{G} = \{g_{1:p}\}$

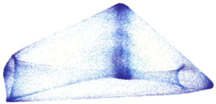
scientific language (torsions)

data driven coordinates

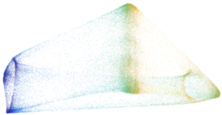
interpretable coordinates



$+$



$=$

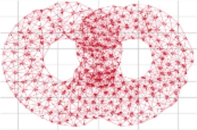


- explains large scale structure with domain-relevant functions
- non-parametric; different from symbolic regression [Brunton et al. 2016, Rudy et al. 2019]
- transmissible knowledge, compare embeddings from different experiments
- extensions: estimated ∇g, simultaneous explanation of multiple manifolds

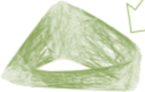
Learning with flows and vector fields [with Yu-chia Chen, Yoannis Kevrekidis]

Directed graph embedding
Manifold + vector field [NIPS 2011]

1-Laplacian estimation
[Arxiv:2103.07626]

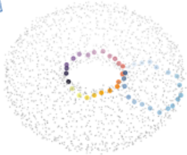


Helmholtz-Hodge
decomposition



Smoothed vector fields

Independent loops
[Arxiv:2107.10970]
[NeurIPS 2021]



Outline

Why Laplacians? Why higher order?

The Laplacian $\mathcal{L}_0 \in \mathbb{R}^{n \times n}$ is central to Manifold Learning

- embedding data by Diffusion Maps [Coifman, Lafon 2006]
- Spectral Clustering

- $\mathcal{L}_0$ related to Riemannian metric – captures geometry of $\mathcal{M}$
- Function approximation
- Smoothing, semi-supervised learning (Laplacian regularization) on manifolds

Higher order Laplacians $\Delta_1, \ldots \Delta_k$ also capture geometry and topology of $\mathcal{M}$

This talk

- estimate first order Laplacian (Helmholtzian) $\mathcal{L}_1(\mathcal{M})$ from data
- calculate Helmholtz-Hodge decomposition of $\mathcal{L}_1(\mathcal{M})$ from data
- Smoothing, function approximation, semi-supervised learning (Laplacian regularization) for vector fields on manifolds
- Manifold prime decomposition ($\approx$ Spectral clustering)
- find short loop bases

Why Laplacians? Why higher order?

The Laplacian $\mathcal{L}_0 \in \mathbb{R}^{n \times n}$ is central to Manifold Learning

- embedding data by Diffusion Maps [Coifman, Lafon 2006]
- Spectral Clustering

- $\mathcal{L}_0$ related to Riemannian metric – captures geometry of $\mathcal{M}$
- Function approximation
- Smoothing, semi-supervised learning (Laplacian regularization) on manifolds

Higher order Laplacians $\Delta_1, \ldots \Delta_k$ also capture geometry and topology of $\mathcal{M}$

This talk

- estimate first order Laplacian (Helmholtzian) $\mathcal{L}_1(\mathcal{M})$ from data
- calculate Helmholtz-Hodge decomposition of $\mathcal{L}_1(\mathcal{M})$ from data
- Smoothing, function approximation, semi-supervised learning (Laplacian regularization) for vector fields on manifolds
- Manifold prime decomposition ($\approx$ Spectral clustering)
- find short loop bases

Why Laplacians? Why higher order?

The Laplacian $\mathcal{L}_0 \in \mathbb{R}^{n \times n}$ is central to Manifold Learning

- embedding data by Diffusion Maps [Coifman, Lafon 2006]
- Spectral Clustering

- $\mathcal{L}_0$ related to Riemannian metric – captures geometry of $\mathcal{M}$
- Function approximation
- Smoothing, semi-supervised learning (Laplacian regularization) on manifolds

Higher order Laplacians $\Delta_1, \ldots \Delta_k$ also capture geometry and topology of $\mathcal{M}$

This talk

- estimate first order Laplacian (Helmholtzian) $\mathcal{L}_1(\mathcal{M})$ from data
- calculate Helmholtz-Hodge decomposition of $\mathcal{L}_1(\mathcal{M})$ from data
- Smoothing, function approximation, semi-supervised learning (Laplacian regularization) for vector fields on manifolds
- Manifold prime decomposition ($\approx$ Spectral clustering)
- find short loop bases

Why Laplacians? Why higher order?

The Laplacian $\mathcal{L}_0 \in \mathbb{R}^{n\times n}$ is central to Manifold Learning

- embedding data by Diffusion Maps [Coifman, Lafon 2006]
- Spectral Clustering

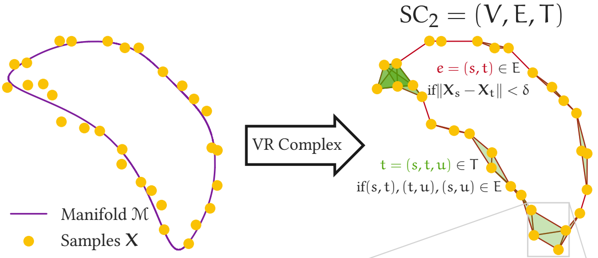
- $\mathcal{L}_0$ related to Riemannian metric – captures geometry of $\mathcal{M}$
- Function approximation
- Smoothing, semi-supervised learning (Laplacian regularization) on manifolds

Higher order Laplacians $\Delta_1, \ldots \Delta_k$ also capture geometry and topology of $\mathcal{M}$

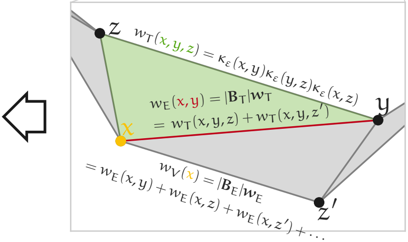
This talk

- estimate first order Laplacian (Helmholtzian) $\mathcal{L}_1(\mathcal{M})$ from data
- calculate Helmholtz-Hodge decomposition of $\mathcal{L}_1(\mathcal{M})$ from data
- Smoothing, function approximation, semi-supervised learning (Laplacian regularization) for vector fields on manifolds
- Manifold prime decomposition ($\approx$ Spectral clustering)
- find short loop bases

Estimating the 1-Laplacian with samples from $\mathcal{M}$

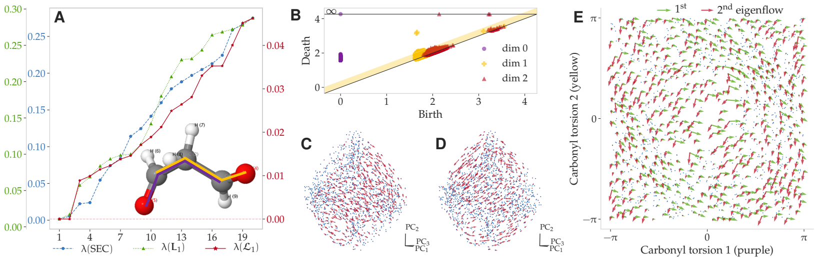


$$\mathcal{L}_1^{\text{down}} = \mathbf{B}_E^\top \boldsymbol{W}_V^{-1} \mathbf{B}_E \boldsymbol{W}_E$$
$$\mathcal{L}_1^{\text{up}} = \boldsymbol{W}_E^{-1} \mathbf{B}_T \boldsymbol{W}_T \mathbf{B}_E^\top$$
$$\Downarrow$$
$$\mathcal{L}_1 = a \cdot \mathcal{L}_1^{\text{down}} + b \cdot \mathcal{L}_1^{\text{up}}$$



$$\boxed{\mathcal{C}_1 \cong \mathbb{R}^{n_E} = \text{gradient} \oplus \text{harmonic} \oplus \text{curl}}$$

$\mathcal{L}_1$ estimation for Molecular Dynamics data (malonaldehyde)



graph Laplacian $w_t = 1$, [Berry, Giannakis 2020], [Chen,M,Kevrekidis 2020]

Outline

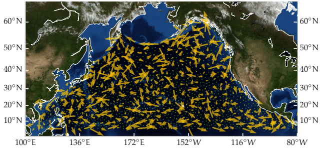
Eigenfunctions of $\mathcal{L}_1$ – what are they useful for?

- Helmholtz-Hodge Decomposition classifies eigenfunctions of $\mathcal{L}_1$

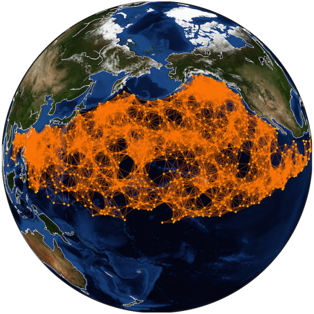
$$\mathcal{C}_1 \cong \mathbb{R}^{n_E} \cong \underbrace{\mathrm{Im}\,\mathcal{L}_1^{\mathrm{down}}}_{\text{gradient}} \oplus \underbrace{\mathrm{Null}\,\mathcal{L}_1}_{\text{harmonic}} \oplus \underbrace{\mathrm{Im}\,\mathcal{L}_1^{\mathrm{up}}}_{\text{curl}}$$

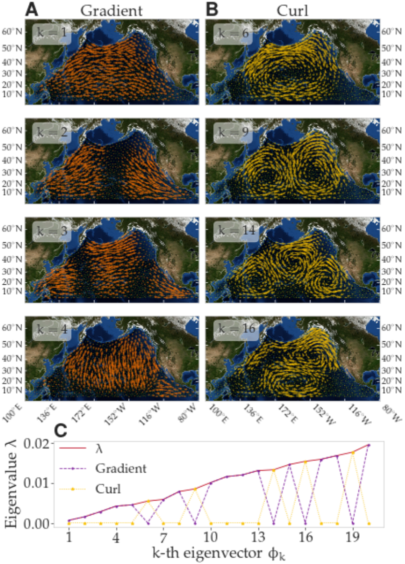
- Analysis of vector fields on $\mathcal{M}$
  - Decompose onto harmonic, gradient, curl
  - Smooth, predict, extend, complete a flow
- Analysis of $\mathcal{M}$
  - $\mathcal{H}_1 = Null\mathcal{L}_1$ Space of loops on $\mathcal{M}$ (1st co-homology space)
  - $\dim \mathcal{H}_1 = \beta_1$ number of (independent loops)
  - Find shortest loop basis

Helmholtz-Hodge decomposition for ocean buoys data

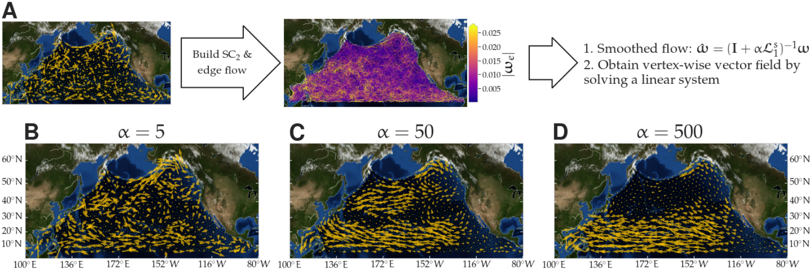


simplicial complex (V, E, T)

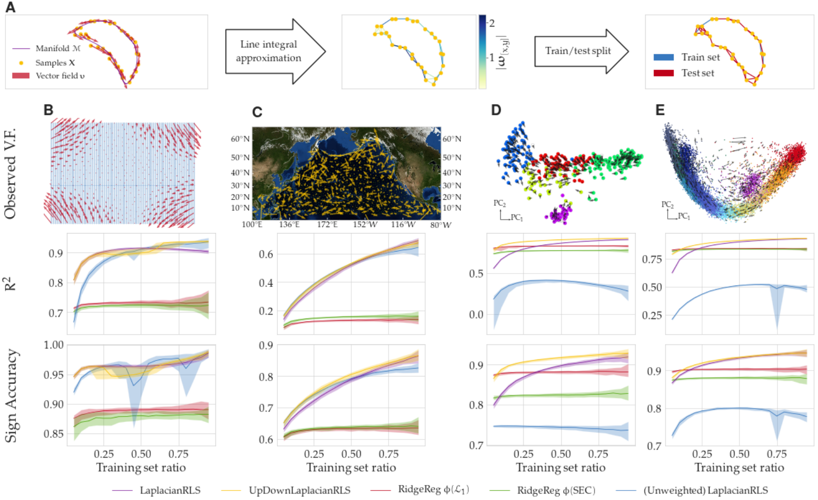




Flow Smoothing



Flow Completion – Semi-Supervised Learning (SSL)



Outline

Manifold coordinates with Scientific meaning

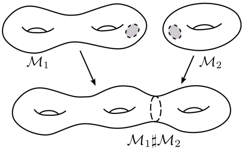
Machine Learning 1-Laplacians, topology, vector fields

- 1-Laplacian $\Delta_1(\mathcal{M})$ estimation from samples
- Analysis of vector fields – Helmholtz-Hodge decomposition
- Harmonic Embedding Spectral Decomposition Algorithm
- Spectral Shortest Homologous Loop Detection

Connected sum and manifold (prime) decomposition

The connected sum **?** $\mathcal{M} = \mathcal{M}_1 \sharp \mathcal{M}_2$:

1. removing two d-dimensional "disks" from $\mathcal{M}_1$ and $\mathcal{M}_2$ (shaded area)
2. gluing together two manifolds at the boundaries



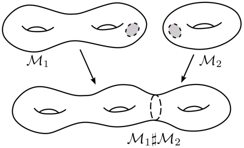
Existence of prime decomposition: factorize a manifold $\mathcal{M} = \mathcal{M}_1 \sharp \cdots \sharp \mathcal{M}_\kappa$ into $\mathcal{M}_i$'s so that $\mathcal{M}_i$ is a prime manifold

- $d = 2$: classification theorem of surfaces **?**
- $d = 3$: the uniqueness of the prime decomposition was shown by Kneser-Milnor theorem **?**
- $d \geq 5$: **?** proved the existence of factorization (but they might not be unique)

Connected sum and manifold (prime) decomposition

The connected sum **?** $\mathcal{M} = \mathcal{M}_1 \sharp \mathcal{M}_2$:

1. removing two d-dimensional "disks" from $\mathcal{M}_1$ and $\mathcal{M}_2$ (shaded area)
2. gluing together two manifolds at the boundaries



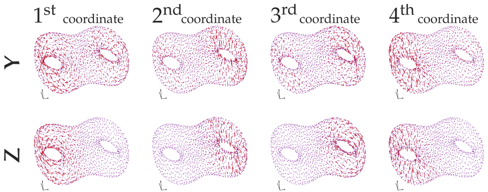
Existence of prime decomposition: factorize a manifold $\mathcal{M} = \mathcal{M}_1 \sharp \cdots \sharp \mathcal{M}_\kappa$ into $\mathcal{M}_i$'s so that $\mathcal{M}_i$ is a prime manifold

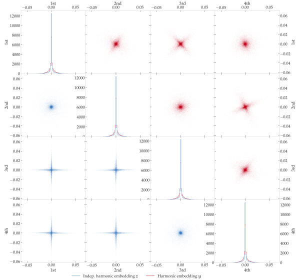
- $d = 2$: classification theorem of surfaces **?**
- $d = 3$: the uniqueness of the prime decomposition was shown by Kneser-Milnor theorem **?**
- $d \geq 5$: **?** proved the existence of factorization (but they might not be unique)

The decomposition of the higher-order homology embedding constructed from the k-Laplacian [Chen,M NeurIPS 2021]

- $\mathcal{L}_1$ is $n_E \times n_E$, operates on edges flows of neighborhood graph
- Null $\mathcal{L}_1 = \mathcal{H}_1$ harmonic space, $\beta_1 = dim\mathcal{H}_1$
- **Y** is basis of $\mathcal{H}_1$ harmonic flows
- **Y** NOT UNIQUE

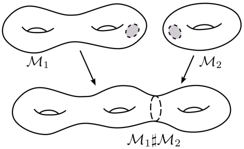
Harmonic Eigenfunctions Y (raw) vs. Z (decoupled)





Connected sum as a matrix perturbation: Assumptions

1. Points are sampled from a decomposable manifold
   - κ-fold connected sum: $\mathcal{M} = \mathcal{M}_1 \sharp \cdots \sharp \mathcal{M}_\kappa$
   - $\mathcal{H}_k(\mathrm{SC})$ (discrete) and $H_k(\mathcal{M}, \mathbb{R})$ (continuous) are isomorphic. Also for every $\mathcal{M}_i$
     - Works for any consistent method to build $\mathcal{L}_k$
     - We use our prior work **?** for $\mathcal{L}_1$

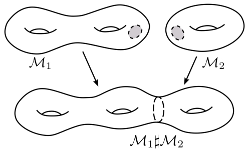


2. No k-homology class is created/destroyed during the connected sum
   - If $\dim(\mathcal{M}) > k$, then $\mathcal{H}_k(\mathcal{M}_1 \sharp \mathcal{M}_2) \cong \mathcal{H}_k(\mathcal{M}_1) \oplus \mathcal{H}_k(\mathcal{M}_2)$ **?**
   - *[Technical] The eigengap of $\mathcal{L}_k$ is the min of each $\hat{\mathcal{L}}_k^{(ii)}$:* $\delta = \min\{\delta_1, \cdots, \delta_\kappa\}$

3. Sparsely connected manifold
   - Not too many triangles are created/destroyed during connected sum (for $k = 1$)
   - Empirically, the perturbation is small even when $\mathcal{M}$ is not sparsely connected
   - *[Technical] Perturbations of ℓ-simplex set Σ_ℓ are small (ϵ_ℓ and ϵ'_ℓ are small) for $\ell = k, k-1$*

Connected sum as a matrix perturbation: Assumptions

1. Points are sampled from a decomposable manifold
    - κ-fold connected sum: $\mathcal{M} = \mathcal{M}_1 \sharp \cdots \sharp \mathcal{M}_\kappa$
    - $\mathcal{H}_k(\mathrm{SC})$ (discrete) and $H_k(\mathcal{M}, \mathbb{R})$ (continuous) are isomorphic. Also for every $\mathcal{M}_i$
        - Works for any consistent method to build $\mathcal{L}_k$
        - We use our prior work **?** for $\mathcal{L}_1$

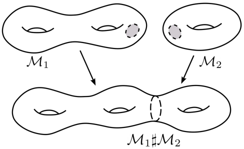


2. No k-homology class is created/destroyed during the connected sum
    - If $\dim(\mathcal{M}) > k$, then $\mathcal{H}_k(\mathcal{M}_1 \sharp \mathcal{M}_2) \cong \mathcal{H}_k(\mathcal{M}_1) \oplus \mathcal{H}_k(\mathcal{M}_2)$ **?**
    - *[Technical] The eigengap of $\mathcal{L}_k$ is the min of each $\hat{\mathcal{L}}_k^{(ii)}$:* $\delta = \min\{\delta_1, \cdots, \delta_\kappa\}$

3. Sparsely connected manifold
    - Not too many triangles are created/destroyed during connected sum (for $k = 1$)
    - Empirically, the perturbation is small even when $\mathcal{M}$ is not sparsely connected
    - *[Technical] Perturbations of ℓ-simplex set Σ_ℓ are small (ϵ_ℓ and ϵ'_ℓ are small) for $\ell = k, k-1$*

Connected sum as a matrix perturbation: Assumptions

1. Points are sampled from a decomposable manifold
   - κ-fold connected sum: $\mathcal{M} = \mathcal{M}_1 \sharp \cdots \sharp \mathcal{M}_\kappa$
   - $\mathcal{H}_k(\mathrm{SC})$ (discrete) and $H_k(\mathcal{M}, \mathbb{R})$ (continuous) are isomorphic. Also for every $\mathcal{M}_i$
     - Works for any consistent method to build $\mathcal{L}_k$
     - We use our prior work **?** for $\mathcal{L}_1$



2. No k-homology class is created/destroyed during the connected sum
   - If $\dim(\mathcal{M}) > k$, then $\mathcal{H}_k(\mathcal{M}_1 \sharp \mathcal{M}_2) \cong \mathcal{H}_k(\mathcal{M}_1) \oplus \mathcal{H}_k(\mathcal{M}_2)$ **?**
   - *[Technical] The eigengap of $\mathcal{L}_k$ is the min of each $\hat{\mathcal{L}}_k^{(ii)}$:* $\delta = \min\{\delta_1, \cdots, \delta_\kappa\}$

3. Sparsely connected manifold
   - Not too many triangles are created/destroyed during connected sum (for $k = 1$)
   - Empirically, the perturbation is small even when $\mathcal{M}$ is not sparsely connected
   - *[Technical] Perturbations of ℓ-simplex set Σ_ℓ are small (ϵ_ℓ and ϵ'_ℓ are small) for $\ell = k, k-1$*

Subspace perturbation

Theorem 1
Under Assumptions 1–3

$$\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + \left(1+\sqrt{\epsilon_k'}\right)^2 \sqrt{\epsilon_{k-1}'} + 4\sqrt{\epsilon_{k-1}}\right]^2 (k+1)^2; \text{ and}$$

$$\left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + 2\epsilon_k + 4\sqrt{\epsilon_k}\right]^2 (k+2)^2,$$

and there exists a unitary matrix $\mathbf{O} \in \mathbb{R}^{\beta_k \times \beta_k}$ *such that*

$$\left\|\mathbf{Y}_{N_k,:} - \hat{\mathbf{Y}}_{N_k,:}\mathbf{O}\right\|_F^2 \leq \frac{8\beta_k \left[\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 + \left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2\right]}{\min\{\delta_1, \cdots, \delta_\kappa\}}. \tag{1}$$

- **Assu. 2:** no topology is destroyed/created
- **Assu. 3:** sparsely connected
- N_k: bound only simplexes that are **not** altered during connected sum

Subspace perturbation

Theorem 1

Under Assumptions 1–3

$$\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + \left(1+\sqrt{\epsilon_k'}\right)^2 \sqrt{\epsilon_{k-1}' + 4\sqrt{\epsilon_{k-1}}}\right]^2 (k+1)^2; \text{ and}$$

$$\left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + 2\epsilon_k + 4\sqrt{\epsilon_k}\right]^2 (k+2)^2,$$

and there exists a unitary matrix $\mathbf{O} \in \mathbb{R}^{\beta_k \times \beta_k}$ *such that*

$$\left\|\mathbf{Y}_{N_k,:} - \hat{\mathbf{Y}}_{N_k,:}\mathbf{O}\right\|_F^2 \leq \frac{8\beta_k\left[\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 + \left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2\right]}{\min\{\delta_1, \cdots, \delta_\kappa\}}. \tag{1}$$

- **Assu. 2:** no topology is destroyed/created
- **Assu. 3:** sparsely connected
- N_k: bound only simplexes that are **not** altered during connected sum

Subspace perturbation

Theorem 1

Under Assumptions 1–3

$$\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + \left(1+\sqrt{\epsilon_k'}\right)^2\sqrt{\epsilon_{k-1}'} + 4\sqrt{\epsilon_{k-1}}\right]^2 (k+1)^2; \text{ and}$$

$$\left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2 \leq \left[2\sqrt{\epsilon_k'} + \epsilon_k' + 2\epsilon_k + 4\sqrt{\epsilon_k}\right]^2 (k+2)^2,$$

and there exists a unitary matrix $\mathbf{O} \in \mathbb{R}^{\beta_k \times \beta_k}$ *such that*

$$\left\|\mathbf{Y}_{N_k,:} - \hat{\mathbf{Y}}_{N_k,:}\mathbf{O}\right\|_F^2 \leq \frac{8\beta_k\left[\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 + \left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2\right]}{\min\{\delta_1, \cdots, \delta_\kappa\}}. \tag{1}$$

- **Assu. 2:** no topology is destroyed/created
- **Assu. 3:** sparsely connected
- N_k: bound only simplexes that are **not** altered during connected sum

Subspace perturbation

Theorem 1

Under Assumptions 1–3

$$\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 \le \left[2\sqrt{\epsilon_k'} + \epsilon_k' + \left(1+\sqrt{\epsilon_k'}\right)^2\sqrt{\epsilon_{k-1}'} + 4\sqrt{\epsilon_{k-1}}\right]^2 (k+1)^2; \text{ and}$$

$$\left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2 \le \left[2\sqrt{\epsilon_k'} + \epsilon_k' + 2\epsilon_k + 4\sqrt{\epsilon_k}\right]^2 (k+2)^2,$$

and there exists a unitary matrix $\mathbf{O} \in \mathbb{R}^{\beta_k \times \beta_k}$ *such that*

$$\left\|\mathbf{Y}_{N_k,:} - \hat{\mathbf{Y}}_{N_k,:}\mathbf{O}\right\|_F^2 \le \frac{8\beta_k\left[\left\|\mathrm{DiffL}_k^{\mathrm{down}}\right\|^2 + \left\|\mathrm{DiffL}_k^{\mathrm{up}}\right\|^2\right]}{\min\{\delta_1, \cdots, \delta_\kappa\}}. \tag{1}$$

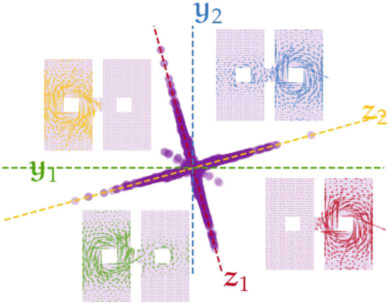
- **Assu. 2:** no topology is destroyed/created
- **Assu. 3:** sparsely connected
- N_k: bound only simplexes that are **not** altered during connected sum

Harmonic Embedding Spectral Decomposition Algorithm

In Simplicial complex (V, E, T), weights $\mathbf{W}_V, \mathbf{W}_E, \mathbf{W}_T$

1. Compute $\mathcal{L}_1$
2. Eigendecomposition
$$\beta_1, \mathbf{Y} \leftarrow Null(\mathcal{L}_1)$$
3. Independent Component Analysis
$$\mathbf{Z} \leftarrow \textsc{ICAnoprewhite}(\mathbf{Y})$$

Out $\mathbf{Z}$



Outline

Manifold coordinates with Scientific meaning

Machine Learning 1-Laplacians, topology, vector fields

- 1-Laplacian $\Delta_1(\mathcal{M})$ estimation from samples
- Analysis of vector fields – Helmholtz-Hodge decomposition
- Harmonic Embedding Spectral Decomposition Algorithm
- Spectral Shortest Homologous Loop Detection

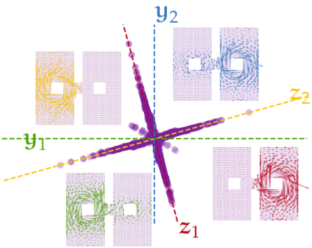
Spectral Shortest Homologous Loop Detection

In $\mathbf{Z} = [\mathbf{z}_1, \ldots \mathbf{z}_{\beta_1}]$, (V, E), edge lengths d_E

for $l = 1 : \beta_1$

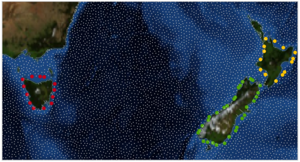
1. Remove edges e with low $|\mathbf{Z}_{le}|$, keep top $1/\beta_1$ fraction E_{keep}
2. Construct $G_l = (V, E_{keep})$, edge weights d_E
3. Repeat for a lot of edges in E_{keep}
   - 3.1 select $e = (t, s_0) \in E_{keep}$
   - 3.2 find shortest path s_0 to t
     $P_e \leftarrow \textsc{Dijkstra}(V, E_{keep} \setminus \{e\}, s_0, t, d_E)$
4. $C_l \leftarrow \text{argmin}_e \, \text{length}(\text{loop}(P_e))$

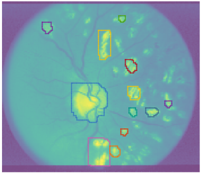
Out loops $C_{1:\beta_1}$

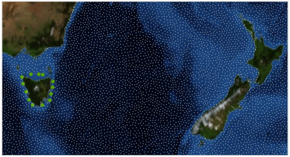


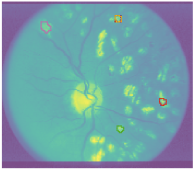
Shortest loop basis on real data

RNA single cell | sculpture | ocean buoys | retina









Summary – Manifold Learning beyond embedding algorithm

- Manifolds, vector fields, ...
  - historically used for modeling scientific data
  - represented analytically

NOW representations learned from data
  - machine learning needs to handle new mathematical concepts
  - need to output results in scientific language

- Generic method for Interpretation in the language of the domain
  - by finding coordinates from among domain-specific functions
  - non-parametric and non-linear
- Extended manifold learning from scalar functions to vector fields
  - first 1-Laplacian estimator
  - continuous limit derived
  - natural extensions of smoothing, semi-supervised learning to vector field data
  - perturbation result for prime manifold decomposition
  - algorithm for shortest loop basis

Summary – Manifold Learning beyond embedding algorithm

- Manifolds, vector fields, . . .
  - historically used for modeling scientific data
  - represented analytically

  NOW representations learned from data
  - machine learning needs to handle new mathematical concepts
  - need to output results in scientific language

- Generic method for Interpretation in the language of the domain
  - by finding coordinates from among domain-specific functions
  - non-parametric and non-linear
- Extended manifold learning from scalar functions to vector fields
  - first 1-Laplacian estimator
  - continuous limit derived
  - natural extensions of smoothing, semi-supervised learning to vector field data
  - perturbation result for prime manifold decomposition
  - algorithm for shortest loop basis

Summary – Manifold Learning beyond embedding algorithm

- Manifolds, vector fields, ...
  - historically used for modeling scientific data
  - represented analytically

  NOW representations learned from data
  - machine learning needs to handle new mathematical concepts
  - need to output results in scientific language

- Generic method for Interpretation in the language of the domain
  - by finding coordinates from among domain-specific functions
  - non-parametric and non-linear
- Extended manifold learning from scalar functions to vector fields
  - first 1-Laplacian estimator
  - continuous limit derived
  - natural extensions of smoothing, semi-supervised learning to vector field data
  - perturbation result for prime manifold decomposition
  - algorithm for shortest loop basis

Samson Koelle, Yu-Chia Chen, Hanyu Zhang, Alon Milchgrub

Hugh Hillhouse (UW), Jim Pfaendtner (UW), Chris Fu (UW)
Stefan Chmiela, Cecilia Clementi, John Maddocks, Klaus Müller, Frank Noé, Christof Schütte
Alexandre Tkatchenko, Alvaro Vasquez-Mayagoitia

Thank you

References I