Unsupervised Learning: Validation beyond Visualization

Marina Meilă
University of Washington
Yu-chia Chen, Samson Koelle, Hanyu Zhang, Weicheng Wu, Vlad Murad
and Ioannis Kevrekidis

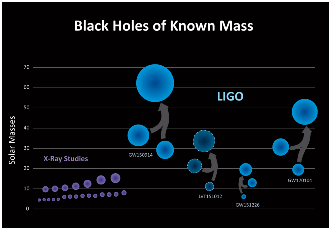
University of Washington
mmp@stat.washington.edu

Centre de Recherche INRIA de Paris

November 13, 2024

Unsupervised learning for the sciences – how do we know machine learning is right?

- Success of modern AI:
  - driven by predicting and acting
  - clear error measure
  - validation "easy" (e.g. speech recognition)
  - many local optima
- Unsupervised learning: clustering, dimension reduction
  - finding [geometric, causal] structure of data
  - formulating "error measure" is part of the problem
  - validation can be EXPENSIVE
  - uniqueness of solution matters
- Big scientific data
  - Allows us to ask more detailed questions (e.g "personalized medicine")
  - Big data contains more complex patterns
  - Machine Learning discovers patterns fast
- Often Hypotheses are cheap, experiments are expensive
- Validation is the bottleneck



Stability guarantees for clustering [M NeurIPS 2018], [Wan, M NIPS 2016],[M ICML 2006] [M, Zhang 2021], [M, Zhang 2023]

provable “correctness” for the practitioner

Manifold coordinates with physical meaning [M,Koelle,Zhang arXiv:1811.11891]

Interpretability in the language of the domain

Explainable or data driven coordinates?

The ManifoldLasso algorithm

Theoretical and experimental recovery results

Outline

Stability guarantees for clustering [M NeurIPS 2018], [Wan, M NIPS 2016],[M ICML 2006] [M, Zhang 2021], [M, Zhang 2023]

provable "correctness" for the practitioner

Manifold coordinates with physical meaning [M,Koelle,Zhang arXiv:1811.11891]

Interpretability in the language of the domain

Explainable or data driven coordinates?

The ManifoldLasso algorithm

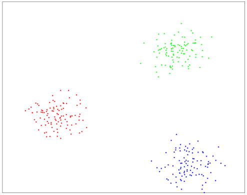
Theoretical and experimental recovery results

For the practitioner of clustering

- Clustering algorithm e.g. K-means, Spectral clustering produces clustering $\mathcal{C}$ with K clusters

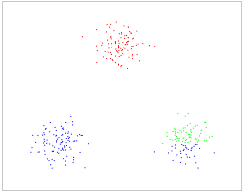
- IDEALLY WANTED: guarantee that $\mathcal{C}$ is correct/optimal
- WHAT WE CAN DO: guarantee that $\mathcal{C}$ is approximately correct/optimal
- WHEN $\mathcal{C}$ is good and stable

Good, stable



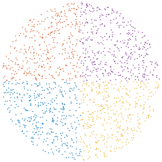
SS output: OI$=1e^{-4}$
OI = Optimality Interval

Bad



no guarantee

Unstable



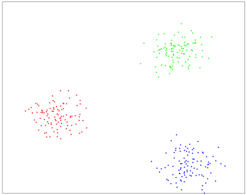
no guarantee

What is an Optimality Interval (OI)?

$\boxed{\mathrm{OI}(\mathcal{C}) = \epsilon}$ is a certificate that

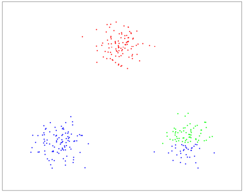
all good clusterings, including the optimal clustering, are contained in the Ball($\mathcal{C}$, ϵ)

Good, stable



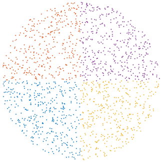
SS output: OI$=1e^{-4}$

Bad



no guarantee

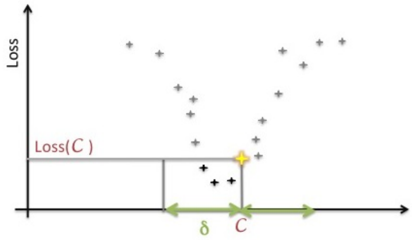
Unstable



no guarantee

What is an Optimality Interval (OI)?

OI($\mathcal{C}$) $= \epsilon$: all good clusterings are contained in the Ball($\mathcal{C}$, ϵ)



- $\mathcal{C}'$ is good if Loss($\mathcal{C}'$) $\leq$ Loss($\mathcal{C}$)$+\alpha$.
- ϵ is OI: for all good $\mathcal{C}'$, $d^{EM}(\mathcal{C}', \mathcal{C}) \leq \epsilon$
  in particular, $d^{EM}(\mathcal{C}^{\mathrm{opt}}, \mathcal{C}) \leq \epsilon$
- If OI exists, we say $\mathcal{C}$ is stable
- OI must be tractably computable in practice

The Sublevel Set (SS) method

Given

- clustering problem defined by Loss ,
  convex relaxation of Loss with space $\mathcal{X}$
- data and clustering $\mathcal{C}$ of data

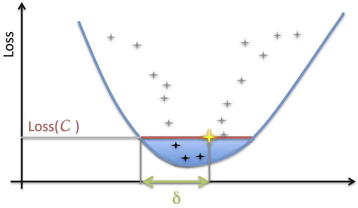
Question Is $\mathcal{C}$ good & stable? Wanted: OI for $\mathcal{C}$

Step 1 Use convex relaxation to define Sublevel Set problem

$$\text{SS} \quad \delta \;=\; \max_{X' \in \mathcal{X}} \|X(\mathcal{C}) - X'\|_F, \quad \text{s.t. } \text{Loss}(X') \leq \text{Loss}(\mathcal{C}).$$

Step 2 Prove that $\|X(\mathcal{C}) - X(\mathcal{C})'\|_F \leq \delta \Rightarrow d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon$ E.g. by [M, MLJ 2012]

Done: ϵ is a Optimality Interval (OI) for $\mathcal{C}$.



Two technical bits

1. SS is convex only if $||X' - X(\mathcal{C})||$ concave
   - Use $|| \, ||_F$ Frobenius norm. $||X(\mathcal{C})||_F^2 = K$ for any clustering.
2. Relating $\| \, \|_F$ to distance between clusterings.

$\|X(\mathcal{C}) - X(\mathcal{C})'\|_F^2 \leq \delta$ $\Rightarrow$ $d^{EM}(\mathcal{C}, \mathcal{C}') \leq \epsilon$

distance between matrices — "misclassification error" metric between clusterings

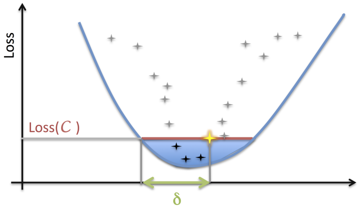
- Theorem proved in [M, MLJ, 2012] with $\epsilon = 2\delta p_{\max}$.
- The tightest result known. Upper/lower bounds between d^{EM}, $\| \, \|_F$ and Rand

- Proofs use geometry of convex sets + refined analysis for small distances
- Example from [Wan,M NIPS16] OI by existing results [Rohe et al, 2014] OI by our method

Summary of SS method

1. Cluster data
2. Set up and solve SS problem
3. If solution δ small enough, we have guarantee ϵ that $\mathcal{C}$ is approximately optimal and all other good clusterings are near it

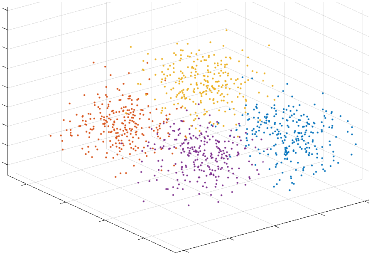
- without any model assumptions, practically applicable
- not all $\mathcal{C}$ can have guarantees



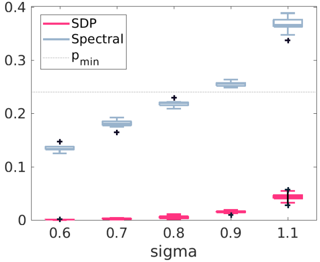
Results for K-means clusterings

$K = 4$ equal Gaussian clusters, $n = 1024$, $||\mu_k - \mu_l|| = 4\sqrt{2} \approx 5.67$

data for $\sigma = 0.9$

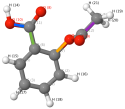


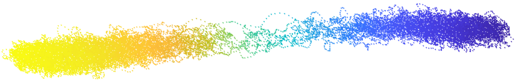
Values of ϵ vs cluster spread σ



Spectral=[M ICML06], SDP=[M NeurIPS 2018]

Aspirin ($C_9O_4H_8$) molecular simulation data [Chmiela et al. 2017]





$K = 2$
$p_{\min} = .26$
$p_{\max} = .74$

$n = 2118$ $\varepsilon = 0.065$ fast ADMM algorithm by Gang Cheng `https://github.`

For what clustering paradigms can we obtain OI's?

"All" ways to map $\mathcal{C}$ to a matrix

space	matrix	definition	size
$\mathcal{X}$	$X(\mathcal{C})$	$X_{ij} = 1/n_k$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\tilde{\mathcal{X}}$	$\tilde{X}(\mathcal{C})$	$\tilde{X}_{ij} = 1$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\mathcal{Z}$	$Z(\mathcal{C})$	$Z_{ik} = 1/\sqrt{n_k}$ iff $i \in C_k$	$n \times K$, orthogonal

Theorem

[M NeurIPS 2018] If Loss has a convex relaxation involving one of $X, \tilde{X}, Z$, then

(1) There exists a convex SS problem

$$\text{(SS)} \quad \delta = \min_{X' \in \mathcal{X}_{\leq c}} \langle X(\mathcal{C}), X' \rangle \quad \text{(similarly for } \tilde{X}, Z\text{).}$$

(2) From optimal δ an OI ε can be obtained, valid when $\varepsilon \leq p_{\min}$.

$$X : X_{ij} = 1/n_k \text{ iff } i, j \in C_k \quad \varepsilon = (K - \delta) p_{\max}$$

$$\tilde{X} : \tilde{X}_{ij} = 1 \text{ iff } i, j \in C_k \quad \varepsilon = \frac{\sum_{k \in [K]} n_k^2 + (n - K + 1)^2 + (K - 1) - 2\delta}{2 p_{\min}}$$

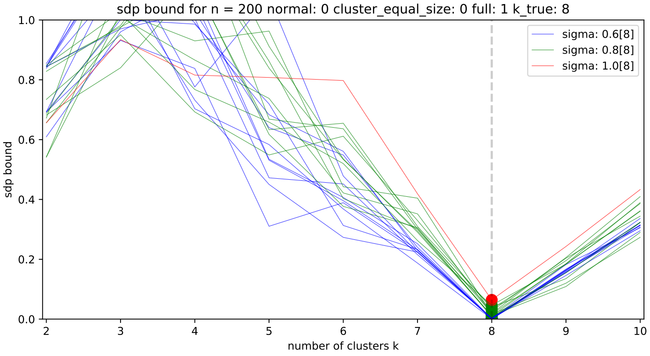
$$Z : Z_{ik} = 1/\sqrt{n_k} \text{ iff } i \in C_k \quad \varepsilon = (K - \delta^2/2) p_{\max}$$

Existence of guarantee depends only on space of convex relaxation.

Relation with other work

- Previous ideas on OI
  - Spectral bounds for Spectral Clustering [M,Shortreed,Xu AISTATS05]
  - Spectral bounds for K-means, NCut and other quadratic costs [M, ICML06 and JMVA 2018]
  - Spectral bounds for networks model based clustering: Stochastic Block Model and Preference Frame Model [Wan,M NIPS16] and comparisons [M, Wan, ISAIM16]
- Previous work we build on
  - Convex relaxations for clustering (MANY!) here we use SDP for K-means [Peng, Wei 2007]
  - Transforming bound on $||X - X'||_F$ into bound on d^{EM} [M MLJ 2012]
- Contrast with work on Clusterability and resilience, e.g. [Ben-David, 2015],[Bilu,Linial 2009]
  - clusterable data, resilient clustering $\approx$ stable $\mathcal{C}$
  - Assume $\exists$ stable $\mathcal{C}$, prove it can be found efficiently
  - Our work: given $\mathcal{C}$, prove it is stable

Stability and the selection of K [Cheng,M,Harchaoui (in prep)]



Recap: generic stability guarantees

$$\text{for any } \mathcal{C}' \in \mathcal{M}, \text{ if } \text{fit}(\mathcal{C}', Q) \leq \text{fit}(\mathcal{C}, Q) + \gamma \quad \text{then} \quad d(\mathcal{C}, \mathcal{C}') \leq \delta(\mathcal{C}, \gamma)$$

paradigm		fit$(\mathcal{C}, Q)$	$d(\mathcal{C}, \mathcal{C}')$	Ref
K-means	dataset	K-means loss	Earthmover's distance	[Zhang, M 2017]
Spectral	dataset	NCut	Earthmover's distance	[Zhang, M 2017]
...	dataset	Loss	Earthmover's distance	[Zhang, M 2017]
Network clustering	dataset	Difference in graph Laplacian	Earthmover's distance	[Wan, M 2016]
Gaussian mixture	distribution Q	$TV(P, Q)$	d_{param}	[Zhang, M 2023]

[1] [2]

[1] H.Zhang and M. Meila, Distribution free optimality intervals for clustering, arXiv 2107.14442

[2] Y.Wan and M.Meila, Graph clustering: block-models and model free result, NeuRIPS 2016

Previous results for Gaussian mixtures

- Recovery guarantees under model assumptions [Vempala Wang 2004, Dasgupta Shulman 2007]
- Parametric stability
  - For e.g. Gaussian mixtures
  - If P, P' are close as distributions
    ... P, P' have similar parameters
  - [Liu, Moitra, 2021] "Settling the robust learnability of mixtures of Gaussians"

Theorem 4.1. *Let ϵ' be a parameter that is sufficiently small in terms of k. There is a sufficiently small function $f(k)$ and a sufficiently large function $F(k)$ such that if*

$$\mathcal{M} = w_1 N(\mu_1, I + \Sigma_1) + \cdots + w_k N(\mu_k, I + \Sigma_k)$$

is a mixture of Gaussians with

- $\|\mu_i\|_2, \|\Sigma_i\|_2 \leq \Delta$ *for all* i
- $\|\mu_i - \mu_j\|_2 + \|\Sigma_i - \Sigma_j\|_2 \geq c$ *for all* $i \neq j$
- $w_1, \dots, w_k \geq w_{\min}$

for parameters $w_{\min}, c \geq \epsilon'^{f(k)}$ and $\Delta \leq \epsilon'^{-f(k)}$ and we are given estimates $\overline{h}_i(X)$ for the Hermite polynomials for all $i \leq F(k)$ such that

$$\left\| v(\overline{h_i}(X) - h_i(X)) \right\|^2 \leq \epsilon'$$

where h_i are the Hermite polynomials for the true mixture $\mathcal{M}$, then there is an algorithm that returns $\mathrm{poly}(1/\epsilon')^{O_1(k)}$ candidate mixtures, at least one of which satisfies

$$\|w_i - \widetilde{w_i}\| + \|\mu_i - \widetilde{\mu_i}\|_2 + \left\|\Sigma_i - \widetilde{\Sigma_i}\right\|_2 \leq \epsilon'^{f(k)}$$

for all i.

- Any hope to do something that can inform practice?
- Yes, partway

Parametric stability with computable bounds [Zhang, M 2023]

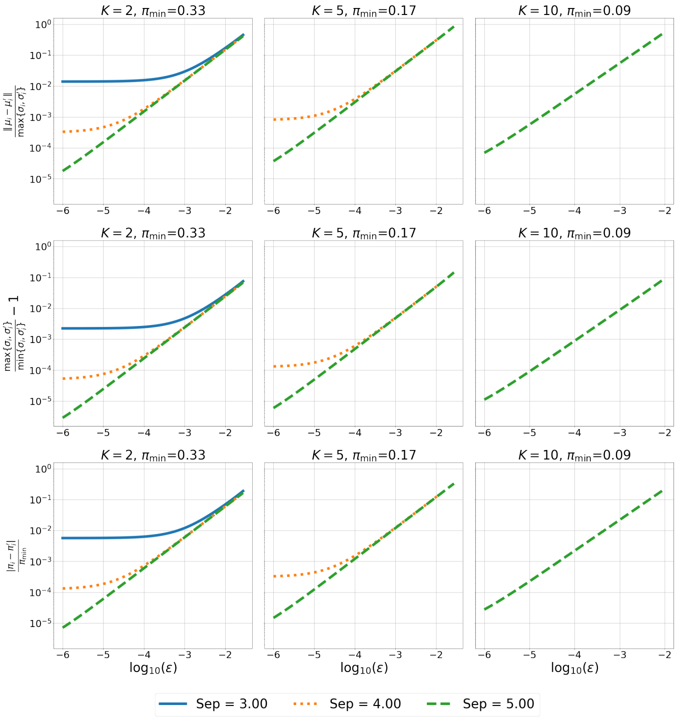
- $\mathcal{M}_{K,w_{\min},c}$ = Spherical Gaussian mixtures with
  fixed K number of components
  fixed minimal/maximal component weight $w_{\min}, w_{\max}$
  minimal separation $c = \min_{i,j\in[K],\, i\neq j} \frac{\|\mu_i - \mu_j\|}{\sigma_i+\sigma_j} \geq c$

$$P = \sum_{i=1}^{K} w_i N(\mu_i, \sigma_i^2 I)$$

- W.r.t. population goodness-of-fit $TV(Q,P)$
- Guarantees for distances in parameter space

$$d_{\mathrm{param}}(P,P') = \underbrace{\min_{\tau\in\Pi_K} \max_{i\in[K]} |w_i - w_{\tau(i)}|}_{\text{Difference in } w} + \underbrace{\frac{\|\mu_i - \mu'_{\tau(i)}\|}{\max(\sigma_i, \sigma'_{\tau(i)})}}_{\text{Difference in } \mu} + \underbrace{\left|\max\left\{\frac{\sigma_i}{\sigma'_{\tau(i)}}, \frac{\sigma'_{\tau(i)}}{\sigma_i}\right\} - 1\right|}_{\text{Difference in } \sigma}$$

- Results also for $\mathcal{M}_{w_{\min}}$, $\mathcal{M}_{w_{\min},w_{\max},c}$ (K not fixed), $\underline{\mathcal{M}_{K,w_{\min},c}}$ (K fixed)



$K=2$, $\pi_{\min}$=0.33
$K=5$, $\pi_{\min}$=0.17
$K=10$, $\pi_{\min}$=0.09
$\frac{\|\mu_i-\mu_i'\|}{\max\{\sigma_i,\sigma_i'\}}$
$\frac{\max\{\sigma_i,\sigma_i'\}}{\min\{\sigma_i,\sigma_i'\}}-1$
$\frac{|\pi_i-\pi_i'|}{\pi_{\min}}$
$\log_{10}(\epsilon)$
Sep = 3.00
Sep = 4.00
Sep = 5.00

Summary + What next?

- Stability guarantees/Optimality Intervals (OI) for any Loss-based clustering paradigm that admits convex relaxation [M, NIPS 2017]
- Guarantees are distribution free, computable, informative
- "Testing" data distribution clusterable [M, Zhang, arXiv:2107.14442]
- Parametric stability for Gaussian Mixtures (in population) [Zhang, M, arXiv:2302.00242] (population version)

- Model selection heuristic [Cheng, M, Harchaoui, Zhang, in preparation]
- Finite sample bounds for mixture models
- How sharp are the OIs (Optimality Intervals) ? Agnostic vs model based bounds

- Validation for other problems with discrete hidden variables
  - sparse linear regression
  - hierarchical clustering
  - ...topic models, graphical models, ...

Outline

Stability guarantees for clustering [M NeurIPS 2018], [Wan, M NIPS 2016],[M ICML 2006] [M, Zhang 2021], [M, Zhang 2023]

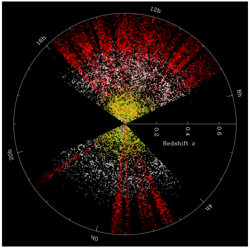
- provable “correctness” for the practitioner

Manifold coordinates with physical meaning [M,Koelle,Zhang arXiv:1811.11891]

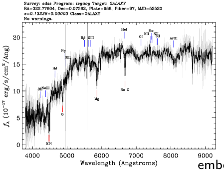
- Interpretability in the language of the domain
- Explainable or data driven coordinates?
- The ManifoldLasso algorithm
- Theoretical and experimental recovery results

Spectra of galaxies measured by the Sloan Digital Sky Survey (SDSS)

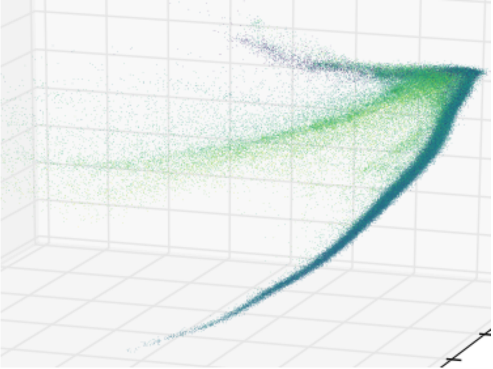
www.sdss.org



www.sdss.org



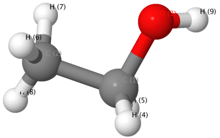
- Preprocessed by Jacob VanderPlas and Grace Telford
- $n = 675,000$ spectra $\times$ $D = 3750$ dimensions



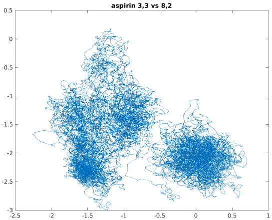
embedding by James McQueen `megaman.github.io` [McQueen, M, VanderPlas, Zhang JMLR 2

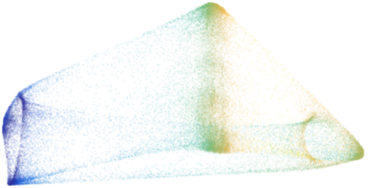
Molecular configurations

ethanol molecule



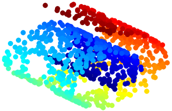
- Data from Molecular Dynamics (MD) simulations of small molecules by [Chmiela et al. 2016]
- $n \approx 200,000$ configurations $\times$ $D \sim 12$ dimensions



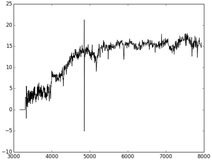


Embedding in 2 dimensions by different manifold learning algorithms

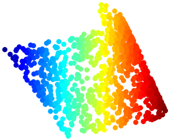
Original data
(Swiss Roll with hole)



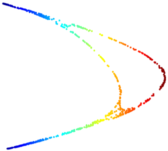
Galaxy spectrum



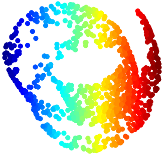
Hessian Eigenmaps
(HE)**X**



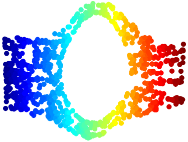
Diffusion Maps (DM)**X?**



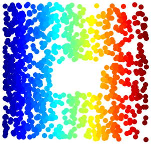
Local Linear Embedding
(LLE)**X**



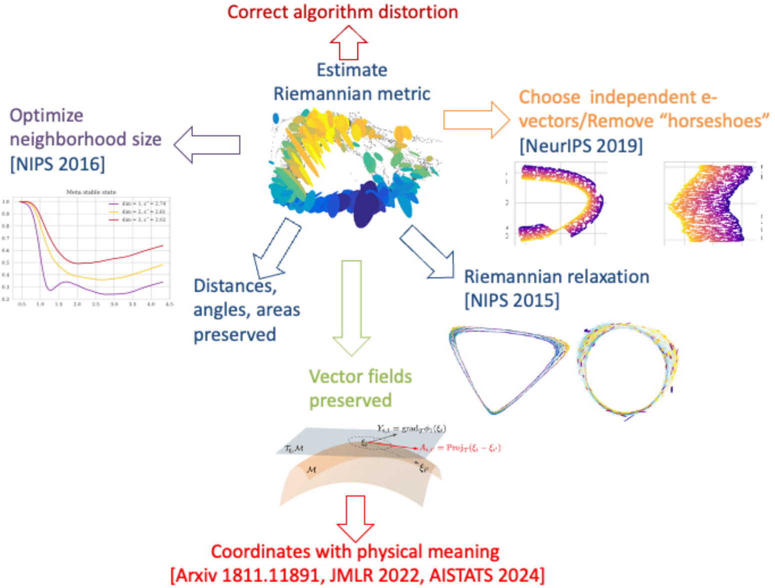
Isomap✓



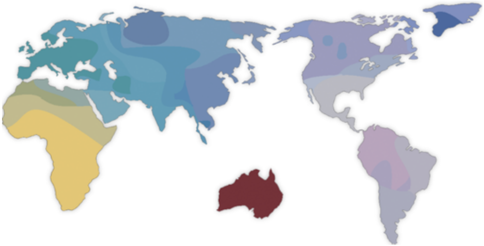
Local Tangent Space
Alignment (LTSA)✓



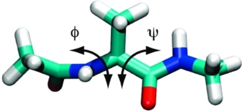
Manifold learning: beyond the embedding algorithm

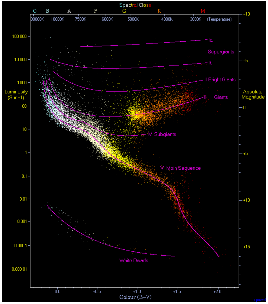


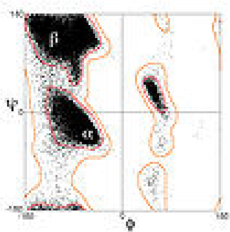
Coordinates with scientific meaning



[Cavalli-Sforza, Menozzi, Piazza, *"The history and geography of human genes"*, 1996]

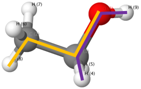




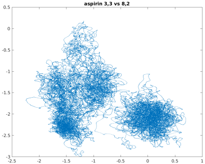


Motivation – understanding data from a Molecular Dynamics simulation

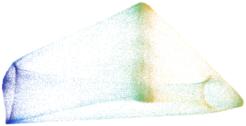
ethanol



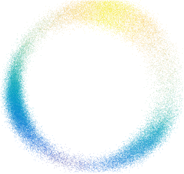
original data



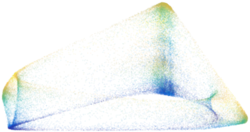
after ML
torsion 1



preprocessed



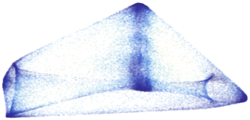
torsion 2



- 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?

Explaining a manifold with domain specific coordinates

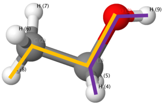
data driven coordinates (e.g. DiffMaps)



$\phi_{\xi_1}, \phi_{\xi_2}, \ldots \phi_{\xi_n}$

$+$

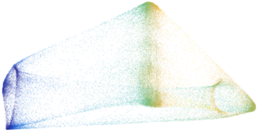
scientific language (torsions)



$\mathcal{F} = \{f_1, f_2 \ldots f_p\}$

$=$

interpretable coordinates



subset $f_{j_1}, \ldots f_{j_d} \subset \mathcal{F}$

- Explanation = finding manifold coordinates from among scientific variables of interest
  - Manifold learning algorithm outputs a data embedding ϕ,
  - $+$ Scientist proposes a dictionary $\mathcal{F}$ with all variables of interest,
  - ManifoldLasso finds new coordinates in $\mathcal{F}$ which are "equivalent" with ϕ

Solution by sparse regression in function space

Wanted: Change of variable

$$\underset{\text{data driven coordinates}}{\phi} = \underset{\downarrow}{h} \circ \underset{\text{selected functions from } \mathcal{G} \text{ (collective coordinates)}}{f_S}$$

Challenges

- sparse, non-linear regression problem
- coordinates ϕ depend on data, algorithm parameters
- hence, h cannot take parametric form
- we cannot choose a basis for h
- cannot assume ϕ_k depends on single f_j
- cannot assume ϕ isometric

Idea: Chain Rule

$$D\phi = DhDg_S$$

- sparse linear regression problem
- $y_i = X_i \beta_i$ for every data i
  - $y_i = \text{grad}\, \phi(\xi_i)$,
  - $X_i = \text{grad}\, f_{1:p}(\xi)$
  - $\beta_{ij} = \frac{\partial h}{\partial f_j}(\xi_i)$
- Constraint: subset S is same for all i

Solution by Group Lasso

- optimize

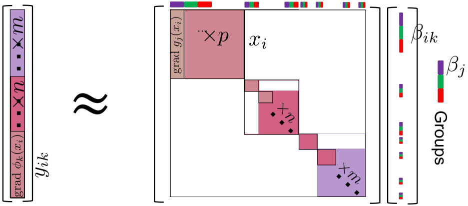
$$\min_{\beta} J_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^{n} ||y_i - X_i \boldsymbol{\beta}_i||_2^2 + \lambda \sum_j ||\beta_j||, \quad (\textsc{ManifoldLasso})$$

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

ManifoldLassoin matrix form

$$y_{ik} = \nabla\phi_k(\xi_i) \quad X_i = \nabla f_{1:p}(\xi) \quad \beta_{ijk} = \frac{\partial hk}{\partial f_j}(\xi_i)$$

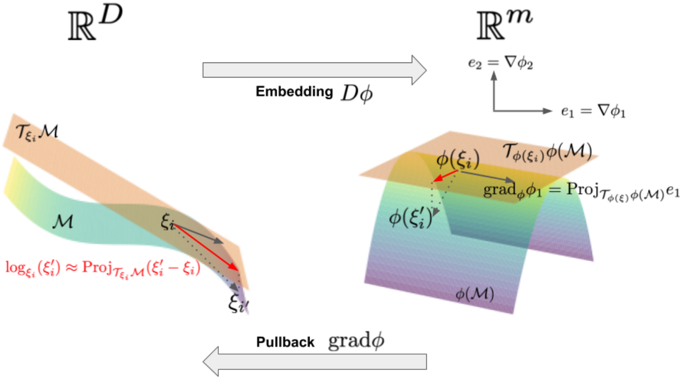
$$J_\lambda(\boldsymbol{\beta}) \;=\; \tfrac{1}{2}\sum_{i=1}^{n} ||y_i - X_i\boldsymbol{\beta}_i||_2^2 + \lambda\sum_{j=1}^{p} ||\boldsymbol{\beta}_{\boldsymbol{j}}||$$



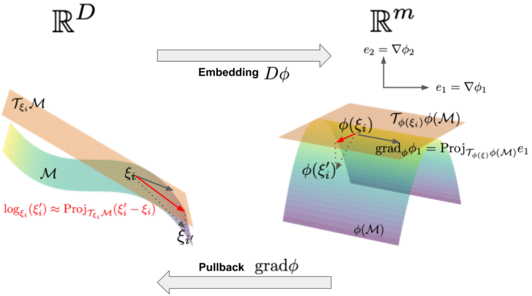
$\beta_j = \mathrm{vec}(\beta_{ijk},\ i = 1:n, k = 1:m) \in \mathbb{R}^{mn}, \quad \beta_{ik} = \mathrm{vec}(\beta_{ijk},\ j = 1:p) \in \mathbb{R}^p.$

Gradients in manifold setting

- gradients ∇ $\rightarrow$ manifold gradients grad in tangents subspace to $\mathcal{M}$
- grad f_j is in $\mathcal{T}_{\xi_i}\mathcal{M}$ (ambient space $\mathbb{R}^D$)
  - ∇f_j known analytically
- grad ϕ_k is in $\mathcal{T}_{\phi(\xi_i)}\phi(\mathcal{M})$ (embedding space $\mathbb{R}^m$)
  1. must estimate tangent subspace $\mathcal{T}_{\phi(\xi_i)}\phi(\mathcal{M})$
  2. must estimate grad $\phi_k(\phi(\xi_i))$ in tangent subspace $\mathcal{T}_{\phi(\xi_i)}\mathcal{M}$
  3. must pull-back grad $\phi_k(\phi(\xi_i))$ to $\mathcal{T}_{\xi_i}\mathcal{M}$



Second Idea: pulling back the ϕ gradients



Wanted $\quad Y_i = \text{grad}_{\mathcal{T}\mathcal{M}}\,\phi(\xi_i) \in \mathbb{R}^{m\times d}$

Estimate tangent subspace at ξ_i by (weighted) PCA

1. Estimate tangent subspace at $\phi(\xi_i)$ $\mathcal{T}_{\phi(\xi_i)}\phi(\mathcal{M})$ by SVD of push-forward Riemannian metric G

$$V_i, \Lambda_i = SVD(G_i, d)$$

2. in $\mathcal{T}_{\phi(\xi_i)}\phi(\mathcal{M})$, $\text{grad}\,\phi_k(\xi_i) = V_i V_i^T e_k$
3. Create neighbor matrices for ξ_i and $\phi(\xi_i)$.

$$A_i = \left[\text{Proj}_{\mathcal{T}_i\mathcal{M}}(\xi_{i'} - \xi_i)\right]_{i'\in\mathcal{N}_i} \quad B_i = \left[\text{Proj}_{\mathcal{T}_i\phi(\mathcal{M})}(\phi(\xi_{i'}) - \phi(\xi_i))\right]_{i'\in\mathcal{N}_i},$$

Solve linear system $\langle A_i, Y_i\rangle \approx \langle B_i, V_i V_i^T I\rangle$ [Luo,Safa,Wang2009]

ManifoldLasso Algorithm

Given Data $\xi_{1:n}$, intrinsic dimension d, embedding $\phi(\xi_{1:n})$

dictionary $\mathcal{F} = \{f_{1:p}\}$

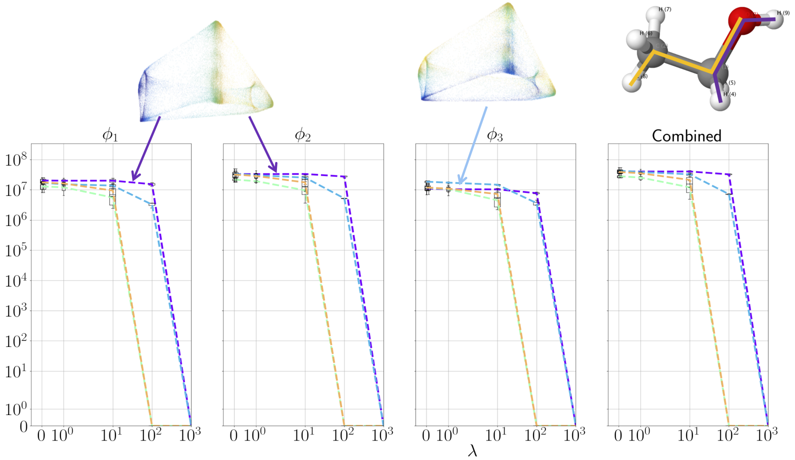
1. Estimate tangent subspace at ξ_i by (weighted) PCA
2. Project dictionary functions gradients ∇f_j on tangent subspace, obtain $X_{1:n} \in \mathbb{R}^{d \times p}$
3. Estimate gradients of $\phi_{1:k}$, obtain $y_{1:n} \in \mathbb{R}^{d \times m}$

   by pull-back from embedding space ϕ
4. Solve $\textsc{GroupLasso}(y_{1:n}, X_{1:n}, d)$, obtain support S

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

Output S

Ethanol MD simulation



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

Theory

[Koelle et al., arXiv:1811.11891, JMLR 2022, AISTATS 2024]

- When is S unique? / When can $\mathcal{M}$ be uniquely parametrized by $\mathcal{F}$?
  Functional independence conditions on dictionary $\mathcal{F}$ and subset $f_{j_1,\ldots j_s}$
- 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 Group Lasso recover S ?
  (Simple) Incoherence Conditions

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

  Theorem If, $\|X_{1:p}\| = 1$, $\mu\nu\sqrt{d} + \frac{\sigma\sqrt{nd}}{\lambda} < 1$ 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*

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

TangentSpaceLasso: ManifoldLasso without embedding

Simplification regress basis of $\mathcal{T}_\xi\mathcal{M}$ on gradients of f_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:*

$$\operatorname{rank}\begin{pmatrix}\operatorname{grad} f_S \\ \operatorname{grad} f_{S'}\end{pmatrix} = \operatorname{rank}\operatorname{grad} f_{S'} \,. \tag{4}$$

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

$$\mu_S = \sup_{\xi\in\mathcal{M}^\circ, j\in S, j'\notin S} |\mathbf{X}_{\{j\},\xi}^T \mathbf{X}_{\{j'\},\xi}| \tag{5}$$

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

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.*

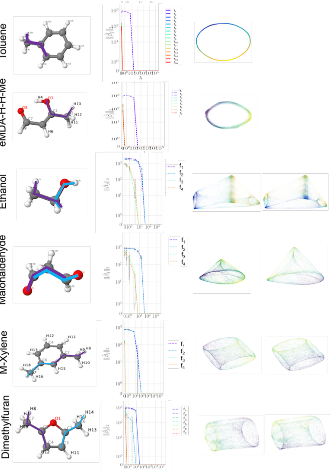
Let S be the 'true' support, $S(\widehat{\mathbf{B}})$ be the support selected by TSLasso, *μ_S and ν_S be defined by (5) and (6), and further assume*

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*

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

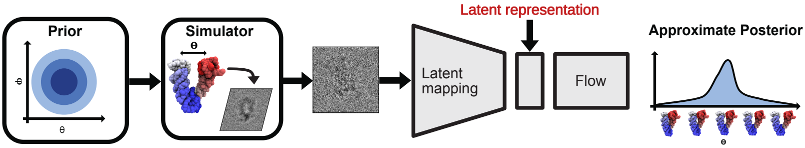
Experiments



Dataset	n	N_a	D	d	ϵ_N	m	n'	p
SwissRoll	10K	NA	51	2	.18	2	100	51
RigidEth	10K	9	50	2	3.5	3	100	12
Ethanol	50K	9	50	2	3.5	3	100	12
Malonald	50K	9	50	2	3.5	3	100	12
Toluene	50K	16	50	1	1.9	2	100	30
Ethanol	50K	9	50	2	3.5	3	100	756
Malonald	50K	9	50	2	3.5	3	100	756
	ϕ						MLasso	$\lvert\mathcal{G}\rvert$

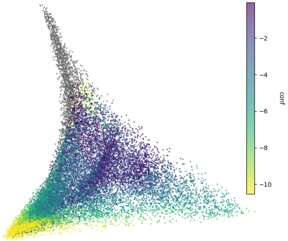
p = dictionary size, m = embedding dimension, n = sample size for manifold estimation, n' = sample size for ManifoldLasso

Understanding latent space representation of cryoEM images

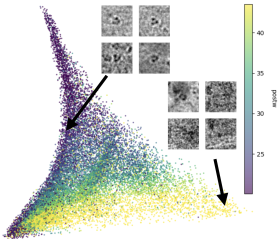


- Estimating conformation of Hemagluttinin molecules from cryoEM images
- Neural network trained on simulated images [Dingeldein et. al. biorXiv:2024]
- Unsupervised study of hidden layer representation: **low dimensional!**

conformation θ



SNR

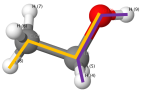
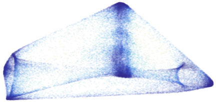
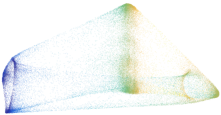


with Luke Evans, Vlad Murad, Lars Dingeldein, Pilar Cossio, Roberto Covino
[submitted NeurIPS 2024 MLSB Workshop]

Summary of ManifoldLasso

- non-linear sparse regression in function spaces $\Rightarrow$ linear sparse regression (Group Lasso)
- ManifoldLasso= coordinate change from data driven coordinates $\phi_{1:m}$ to collective coordinates $\mathcal{F} = \{f_{1:p}\}$

scientific language | data driven coordinates | interpretable coordinates


$+$

$=$


- explains large scale structure with domain-relevant functions
- transmissible knowledge, compare embeddings from different experiments
- non-linear, non-parametric, basis-free, not symbolic regression [Brunton et al. 2016, Rudy et al. 2019] [Udrescu, Tegmark 2020]
- No manifold necessary immediate extensions to Principal Components, autoencoders (low dimensional!), sparse functional regression

Applications

- set of f's that covary (e.g. small protein folding), level sets (in progress)
- simultaneous explanation of multiple systems
- dynamical systems (future)

Summary: Towards knowledge that is transferable

Cluster validation without model assumptions [M NeurIPS 2018]

- A general method that can be applied to any clustering cost that has a convex relaxation / mixtures of gaussians
- A general framework for validation without model assumptions

Manifold coordinates with pysical meaning [arXiv:1811.11891]

- Interpretation in the language of the domain
- From non-parametric to parametric

Learning vector fields on manifolds [arXiv:2103.07626]
Python package `github.com/mmp2/megaman`

- tractable for millions of points
- manifold learning and clustering
- incorporates state of the art results

Towards unsupervised validation for unsupervised learning

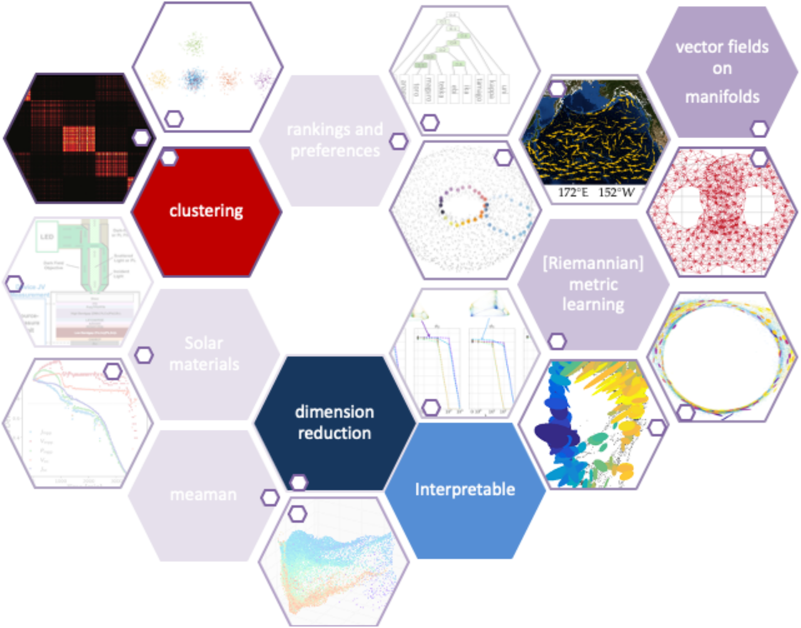
- In Machine Learning: Unsupervised Learning is the next big challenge
- In the sciences: Unsupervised Learning is about explanation and understanding
- Automated discoveries require automated validation
    - Combine data driven/machine learning methods with domain knowledge/concepts
    - On purely mathematical/statistical grounds
- Remove algorithmic artefacts
- Quantitative measures of "correctness" / robustness / uncertainty
- Is explanation unique?
- Statistical guarantees – with minimum od untestable assumptions
- Good community practices – all machine learning algorithms should come with validation procedures

Hanyu Zhang, Samson Koelle, Vlad Murad, Yu-Chia Chen, Weicheng Wu
Ioannis Kevrekidis (JHU)

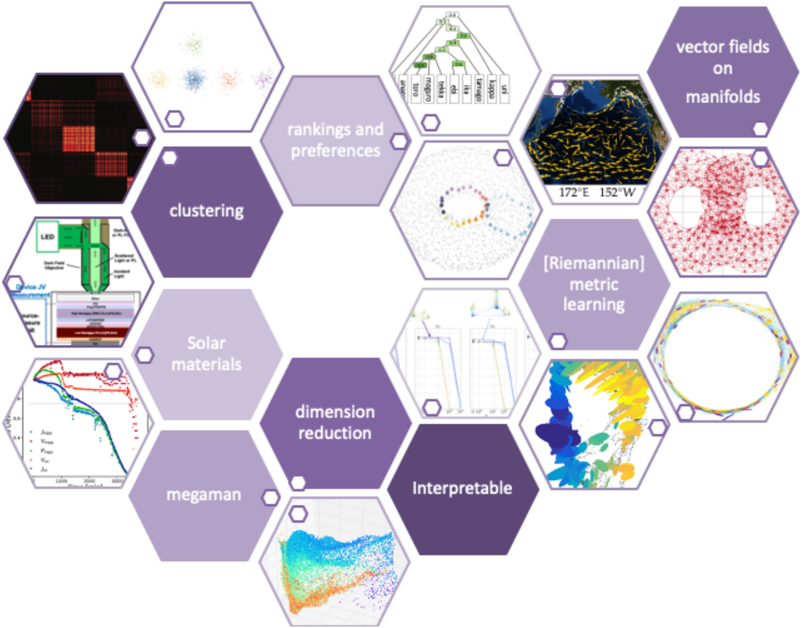
Alexandre Tkatchenko (Luxembourg), Stefan Chmiela (TU Berlin)
Pilar Cossio (Flatiron), Luke Evans (Flatiron)
Lars Dingeldein (Frankfurt), Roberto Covino (Frankfurt)

Thank you

NUS
National University of Singapore
NSF
Inria
ipam
iMSi

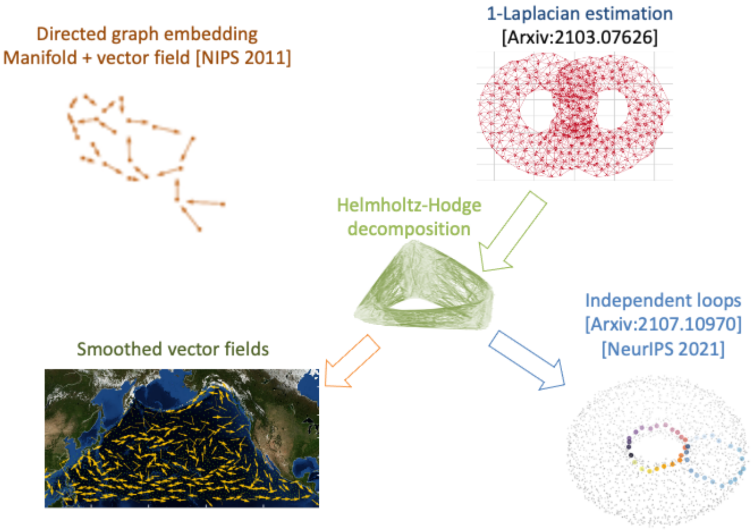


vector fields on manifolds
rankings and preferences
172°E 152°W
clustering
[Riemannian] metric learning
Solar materials
dimension reduction
Interpretable
meaman



clustering
rankings and preferences
vector fields on manifolds
172°E 152°W
LED
[Riemannian] metric learning
Solar materials
dimension reduction
megaman
Interpretable

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



References I

[1] Meila , *"How to tell when a clustering is approximately correct using convex relaxations"*, , Advances in Neural Information Processing Systems (NeurIPS), 2018.

[2] Meila , *"The local equivalence of several distances between clusterings – A geometric perspective"*, Machine Learning Journal, **86**(3), pp 369-389, 2012.

[3] Wan, Y*, Meila , *"Graph clustering: block-models and model-free results"*, Advances in Neural Information Processing Systems (NIPS), 2016.

[4] Wan, Y*, Meila , *"Benchmarking recovery theorems for the DC-SBM"*, International Symposium on Artificial Intelligence and Mathematics (ISAIM), 2016.

[5] Meila, M. and Zhang, H. (2021). *"Distribution free optimality intervals for clustering"*. *arXiv*, 2107.14442.

[6] Zhang, H*, Meila , *"The Parametric Stability of Well-separated Spherical Gaussian Mixtures"* *arXiv*,2302.00242.
Manifold learning. Interpretable manifold coordinates

[7] Meila , Zhang, H.*, *"Manifold learning: what, how, and why"*, Annual Reviews in Statistics and its Applications, (accepted) 2024.

References II

[8] Koelle, S.*, Zhang, H.*, Meila , , Chen, Y-C.$*$, *"Manifold coordinates with physical meaning"*, Journal of Machine Learning Research, 2022.

[9] Koelle, S.*, Zhang, H.*, Meila , *"Parametrizing manifolds by dictionaries"*, (submitted)

Learning flows and vector fields with higher order Laplacians

[10] Chen, Y.*, Wu, W.*, Meila, M., and Kevrekidis, I. G. (2021). *"Helmholtzian eigenmap: Topological feature discovery & edge flow learning from point cloud data"*. *arXiv*, 2103.07626.

[11] Chen, Y.-C.* and Meila, M. *"The decomposition of the higher-order homology embedding constructed from the k-laplacian"*. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15695–15709 (Oral presentation). Curran Associates, Inc., 2021.

"Testing" population stability (K-means loss)

A1. $\mathcal{D} = \{x_1, \cdots, x_n\}$ is sampled i.i.d. from $\mathcal{P}$, supported on a subset of $\mathbb{R}^d$. $\mathcal{P}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.
A2. [Uniform Convergence of Loss_{Km}] There exists a function $\Psi(n, \delta)$ such that, for any n sufficiently large and $\delta \in (0, 1]$, with probability $1 - \delta$

$$\sup_{\mathcal{C} \in \mathbf{C}_K(\mathcal{D})} |\text{Loss}_{\text{Km}}(\mathcal{P}; \mathcal{C}) - \text{Loss}_{\text{Km}}(\mathcal{D}, \mathcal{C})| \leq \Psi(n, \delta)$$

Theorem

Suppose $\mathcal{P}$ satisfies Assumptions 1 and 2, and let $\delta \in (0, 1]$. If any optimal clustering $\mathcal{C}^{opt}$ on $\mathcal{P}$ is (α, ε) unstable for some $\alpha > 0$, then with probability $1 - \delta$ over samples $\mathcal{D}$, with $|\mathcal{D}| = n$, any optimal clustering $\widehat{\mathcal{C}}^{opt}$ of $\mathcal{D}$ is $(\alpha + 2\Psi(n, \delta/2), \varepsilon/2 - \sqrt{\log(4/\delta)/2n})$ unstable.

Theorem 4. *Let $P \in \mathcal{M}(K, \pi_{\min}, \pi_{\max}, c)$. Suppose P' is any model in $\mathcal{M}(K', \pi_{\min}, \pi_{\max}, c)$ such that $TV(P, P') \leq 2\epsilon$ where $\max\{K, K'\} \leq 1/\pi_{\min}$, $\pi_{\max} \leq 1 - (\min\{K, K'\} - 1)\pi_{\min}$. Let c_0, η_0 be defined as in (7) and (8). Then, if $c \geq c_0\eta_0$ and $\pi_{\min} > 2\epsilon$, we have $K = K'$ and further, there exists a permutation $\mathrm{perm} \in \mathbb{S}_K$ and constants $c^* \in [0, c_0], \eta^* \in [1, \eta_0]$ satisfying (9) and (10), such that for each $i \in [K]$,*

$$||\mu_i - \mu'_{\mathrm{perm}(i)}|| \leq c^*\eta^*\sigma_i \tag{11}$$

$$\max\{\sigma_i/\sigma'_{\mathrm{perm}(i)}, \sigma'_{\mathrm{perm}(i)}\sigma_i\} \leq \eta^* \tag{12}$$

$$|\pi_i - \pi'_{\mathrm{perm}(i)}| \leq 2\epsilon + (1 - \pi_{\min} + \pi_{\max})\Phi(-C(c, c^*, \eta^*)) \;, \tag{13}$$

where $C(c, c^*, \eta^*)$ is defined by

$$C(c, c^*, \eta^*) := \sqrt{\frac{c^2}{2(\eta^*)^2} + \frac{1}{2\eta^*}(c - \frac{c^*}{2})^2 - \frac{(c^*)^2(1+\eta^*)^2}{16(\eta^*)^2}} - \frac{c^*}{2} \;. \tag{14}$$

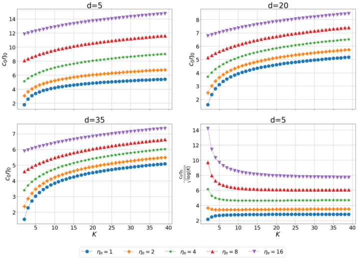
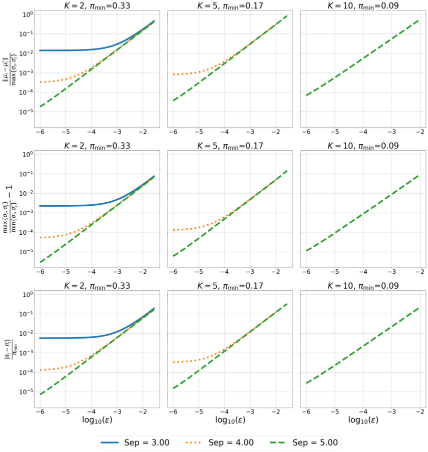


Figure 2: **Sufficient minimal separation** $c_0\eta_0$ in Theorem 4 under different settings. **Top**
Top right, Bottom Left show the dependence of $c_0\eta_0$ on K and $\eta_\pi = \pi_{\max}/\pi_{\min}$ in dimen
$d = 5, 20, 35$, respectively. **Bottom right** shows that the dependence of $c_0\eta_0$ on K asympto
$\sqrt{\log K}$.



K-means Sublevel Set problem

$$\text{Loss}(\mathcal{C}) \quad = \quad \langle D, X(\mathcal{C})\rangle, \quad D = \text{squared distance matrix} \in \mathbb{R}^{n\times n}$$

$$(\text{SS}_{\text{Km}}) \quad \delta \;=\; \min_{X' \in \mathcal{X}} \langle X(\mathcal{C}), X' \rangle \quad \text{s.t.} \langle D, X' \rangle \;\leq\; \text{Loss}(\mathcal{C})$$

a Semi-Definite Program (SDP).

Algorithm

Input Matrix of squared distances D, clustering $\mathcal{C}$

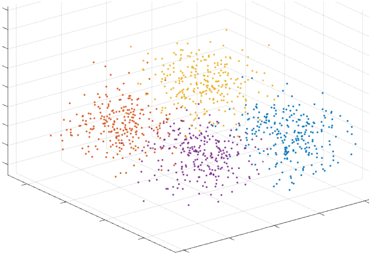
1. Solve (SS_{Km}), get optimal value δ.
2. **If** $\epsilon = (K - \delta)p_{\max} \leq p_{\min}$ **then** $\mathcal{C}$ is stable

   else no guarantee.

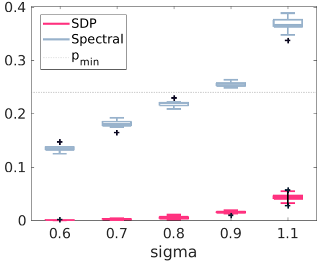
Experiments with K-means clusterings

$K = 4$ equal Gaussian clusters, $n = 1024$, $||\mu_k - \mu_l|| = 4\sqrt{2} \approx 5.67$

data for $\sigma = 0.9$



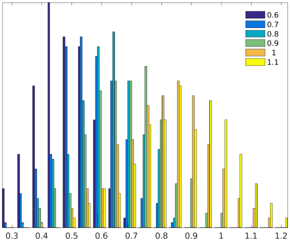
Values of ϵ vs cluster spread σ



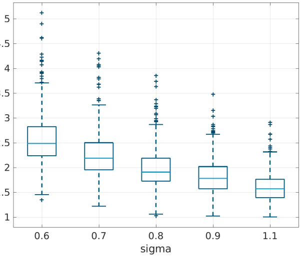
Spectral=[M ICML06], SDP=[M NeurIPS 2018]

Separation statistics

distance to own center over min center separation, colored by σ.



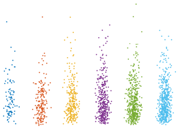
distance to second closest center over distance to own center, versus σ



Results for unequal clusters

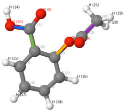
K=4	Unequal normal clusters			Unequal non-normal clusters		
σ	$n = 200$	$n = 400$	$n = 800$	$n = 200$	$n = 400$	$n = 80$
0.6	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.001(0.001)	0.001(0.000)	0.002(0.0
0.8	0.01(0.01)	0.01(0.01)	0.01(0.01)	0.006(0.004)	0.004(0.002)	0.007(0.0
1.0	0.09 (0.05)	0.06 (0.01)	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.03 (0.
1.2	0.28 (0.08)	0.21 (0.05)	0.21 (0.03)	0.16 (0.06)	0.14 (0.03)	0.13 (0.

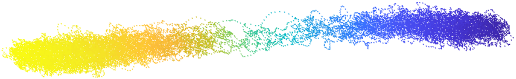
$K = 6$	normal	non-normal
σ	$n = 525$	$n = 525$
0.06	0.00(0.00)	0.005(0.001)
0.08	0.01(0.00)	0.006(0.001)
0.1	0.01(0.00)	0.009(0.003)



Outlier removal: before clustering, 0.2–0.5% fraction of points i with largest $\sum_j D_{ij}$ were removed; j ranges over $p_{\min}/2$ nearest neighbors of i.

Aspirin ($C_9O_4H_8$) molecular simulation data [Chmiela et al. 2017]





$K = 2$
$p_{\min} = .26$
$p_{\max} = .74$

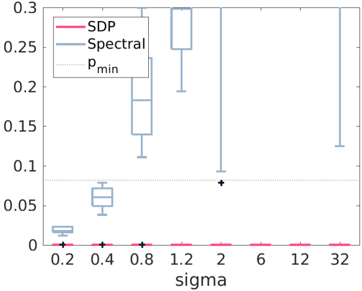
all data	$n = 2118$	$\varepsilon = 0.065$	computing time 17h
1271 inliers removed	$n = 847$	$\varepsilon = 0.047$	computing time 42min

b

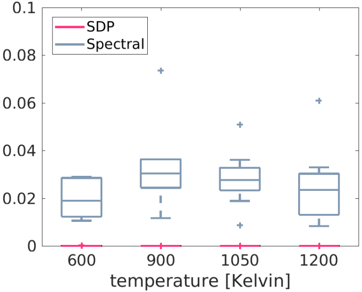
Results for Spectral Clustering by Normalized Cut

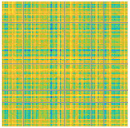
Spectral=[M AISTATS05], SDP=[M NeurIPS 2018]

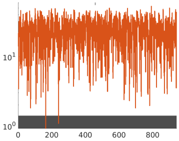
Synthetic S, $n = 100$

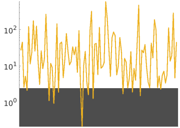


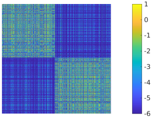
Chemical reaction data, $n \approx 1000$









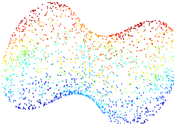


Brief intro to manifold learning algorithms

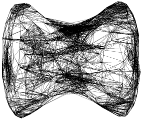
ALL ML Algorithms

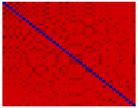
- **Input** Data $p_1, \ldots p_n$, embedding dimension m, neighborhood scale parameter ϵ
- Construct neighborhood graph p, p' neighbors iff $||p - p'||^2 \leq \sqrt{\epsilon}$
- Construct a $n \times n$ sparse distance matrix

$$D = [||p - p'||]_{p,p' \text{neighbors}}$$



$p_1, \ldots p_n \subset \mathbb{R}^D$





Isomap ML algorithm

ISOMAP [Tennenbaum, deSilva & Langford 00]

1. Find all shortest path distances in neighborhood graph
2. Construct matrix of distances

$$M = [\text{distance}^2_{pp'}]$$

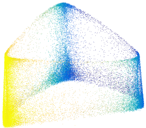
3. use M and Multi-Dimensional Scaling (MDS) to obtain d dimensional coordinates for $p \in \mathcal{D}$

Diffusion Maps Algorithm

Input coordinates $U \in \mathbb{R}^{n\times D}$, bandwidth $\sqrt{\epsilon}$, embedding dimension s

1. Compute Laplacian $L \in \mathbb{R}^{n\times n}$
2. Compute eigenvectors of L for smallest $s+1$ eigenvalues $[\phi_0\, \phi_1\, \ldots \phi_s] \in \mathbb{R}^{n\times s}$
   - ϕ_0 is constant and not informative
   - These are the slow modes of the system

The embedding coordinates of $p_{i:}$ are $(\phi_{i1}, \ldots \phi_{is})$



- Embedding dimension s = number of eigenvectors
- Intrinsic dimension $d \leq s$ effective number of degrees of freedeom

UMAP: Uniform Manifold Approximation and Projection [McInnes, Healy, Melville,2018]

Input k number nearest neighbors, d,

1. Find k-nearest neighbors
2. Construct (asymmetric) similarities w_{ij}, so that $\sum_j w_{ij} = \log_2 k$. $W = [w_{ij}]$.
3. Symmetrize $S = W + W^T - W. * W^T$ is similarity matrix.
4. Initialize embedding ϕ by LaplacianEigenmaps.
5. Optimize embedding.

   Iteratively for n_{iter} steps

   5.1 Sample an edge ij with probability $\propto \exp -d_{ij}$
   5.2 Move ϕ_i towards ϕ_j
   5.3 Sample a random j' uniformly
   5.4 Move ϕ_i away from $\phi_{j'}$
   Stochastic approximate logistic regression of $||\phi_i - \phi_j||$ on d_{ij}.

Output ϕ

The Laplacian

Laplacian

Input coordinates $U \in \mathbb{R}^{n \times D}$, bandwidth $\sqrt{\epsilon}$

1. Compute similarity matrix $S_{ij} = \exp\left[-\frac{||U_{i:} - U_{j:}||^2}{\epsilon}\right]$
2. First normalization $d_i = \sum_{j=1}^{n} S_{ij}$, $\tilde{L}_{ij} = L_{ij}/d_i d_j$
3. Second normalization $d_i' = \sum_{j=1}^{n} \tilde{L}_{ij}$, $L_{ij} = \tilde{L}_{ij}/d_i'$
   removes the biases due to sampling density
4. Output L, d_i'

- Laplacian L central to understanding the manifold geometry
- $\lim_{n\to\infty} L = \Delta_{\mathcal{M}}$ [Coifman,Lafon 2006]
- $\sqrt{\epsilon}$ represents the scale of the local neighborhood