Manifold Fitting: An Invitation to Statistics[1]

Zhigang Yao

National University of Singapore

April 5, 2024

[1] Manifold fitting with CycleGAN (*PNAS*), 2023

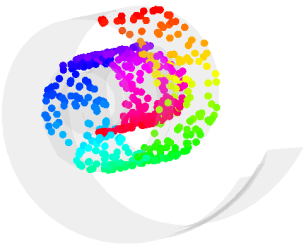
Relevant Collaborator, Students/RFs

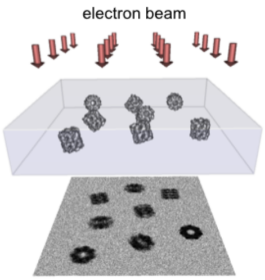
ST Yau (Harvard/Tsinghua)[1 3 4];
Y. Xia (Former RF, NUS)[2]; B. Li (RF, NUS) [3 4]; J. Su (Student, NUS)[1 3]; Y. Lu (Student, NUS)[4]

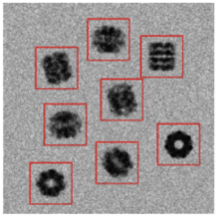
[1] Manifold fitting with CycleGAN (*PNAS*), 2024
[2] unbounded noise, 2019
[3] manifold fitting, 2022
[4] Single-cell via MF, 2024 (*PNAS*, revised)

Nonparametrics on Manifolds or Manifolds in Nonparametrics

"non-Abelian"; Picture says (almost) all:



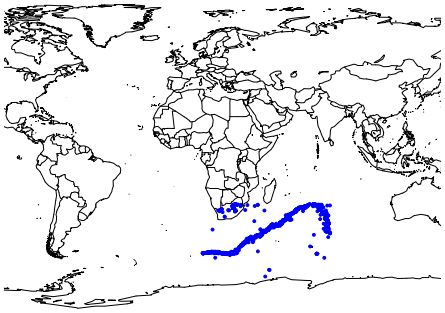


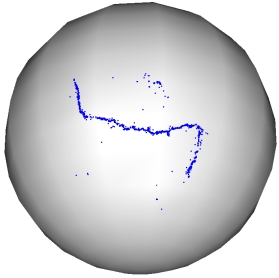


Data on Manifolds

Data on manifolds[†] may arise in (at least) two ways:

(1) Manifold is actual physical space where data reside
↪ Usually sphere; from geophysics to marine biology



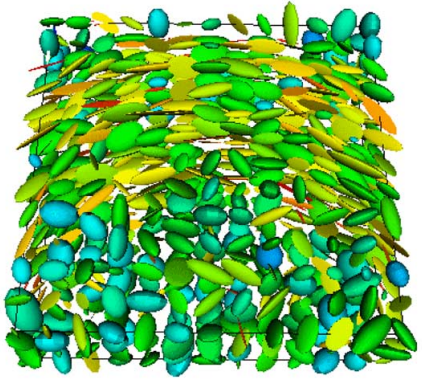


[†]contrast to manifolds in data analysis, or manifold-valued data

Data on Manifolds

Data on manifolds may arise in (at least) two ways:

(2) Multivariate data under non-linear constraints, thus being forced onto manifold
↪ e.g. cones for positive-def matrices or Stiefel manifolds for ordered bases



Data on Manifolds

$$D = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix} : v^\top D v > 0, v \in \mathbb{R}^3$$

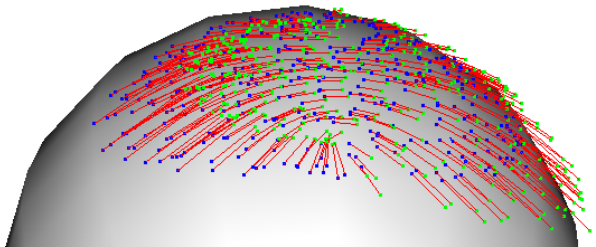
- $\mathcal{P}_3$: the space of all symmetric positive definite 3 by 3 matrices
- $\mathcal{P}_3 \subset \mathbb{R}^6$
- $\mathcal{P}_3$ is convex but not linear in $\mathbb{R}^6$: $\frac{3}{2}D_1 - \frac{1}{2}D_2$ might not in $\mathcal{P}_3$

A Principal Flow?

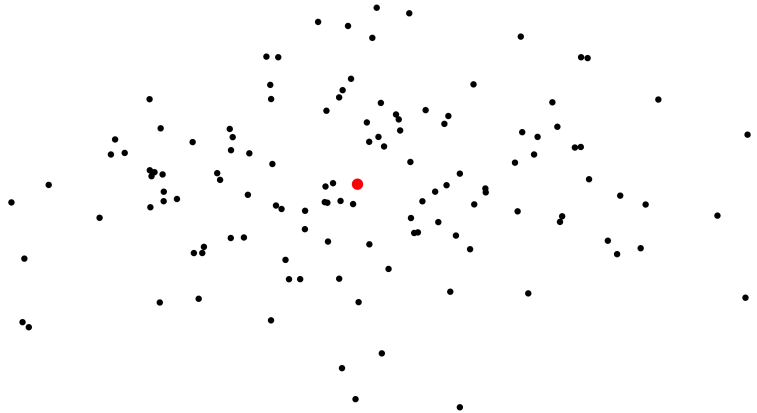
What does "*curve of maximal variation*" mean?

Would like a reasonably smooth curve $\gamma(x)$ whose derivative $\dot{\gamma}(x) \in T_x\mathcal{M}$ at any point $x \in \mathcal{M}$ is $\approx$ (parallel to) $\lambda_1(x)e_1(x)$

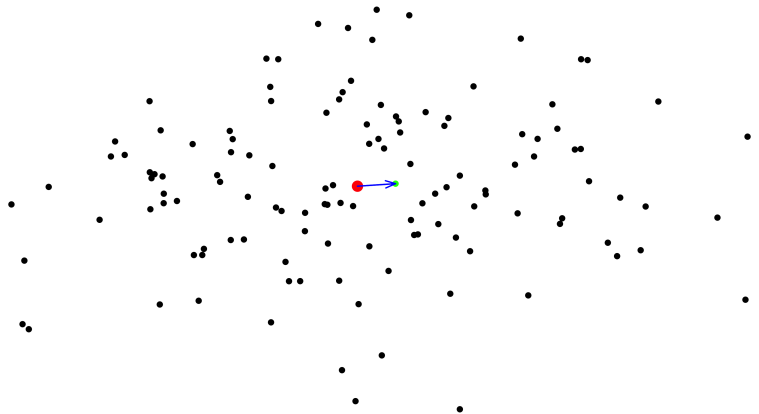
...AND maximizes the work done by the field on a particle traveling along its path



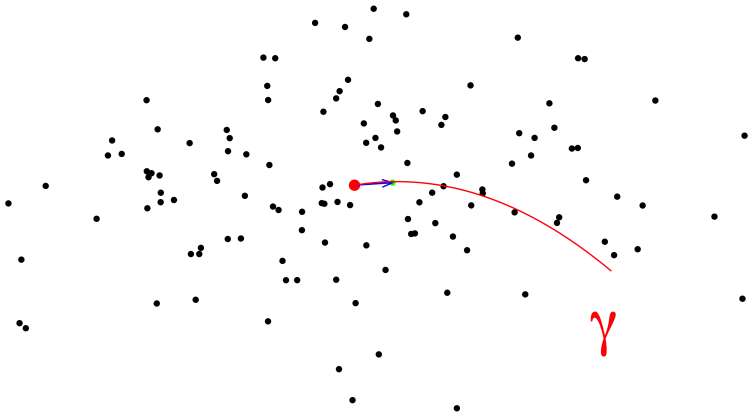
Illustration



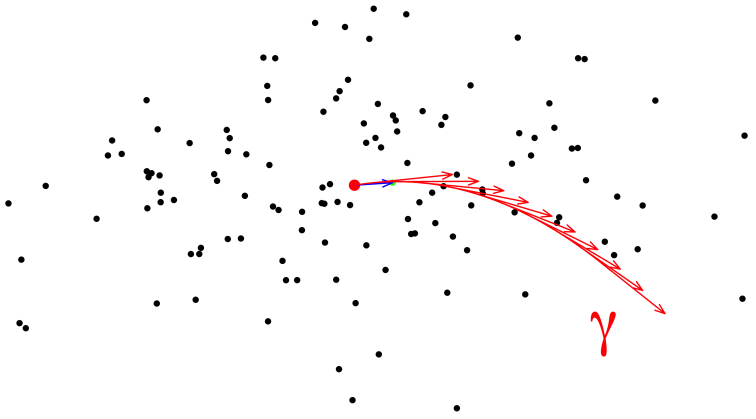
Illustration



Illustration



Illustration



A Principal Flow‡

(mod technicalities)[a] Curve with midpoint $\bar{x}$, maximizing

$$\int \left| \left\langle \dot{\gamma}(t), W(\gamma(t)) \right\rangle \right| dt$$

$$\text{SubM}(A, v, \mathcal{M}) = \Big\{ \gamma : [0, r] \to \mathcal{M}, \gamma \in C^2(\mathcal{M}), \gamma(s) \neq \gamma(s') \text{ for } s \neq s',$$
$$\gamma(0) = A, \dot{\gamma}(0) = v, \ell(\gamma[0, t]) = t \text{ for all } 0 \leq t \leq r \leq 1 \Big\}.$$

[a] (several technical issues will not be discussed)

- Answer: yes, reformulate to Euler-Lagrange equations
- $\exists$ unique solution under mild conditions on manifold+field
- Requires geodesics and second fundamental tensor
- Numerically Feasible for many "standard" manifolds
- Canonical: reduces to ordinary PCA in Euclidean spaces

‡ Panaretos, V. M., Pham, T., & Yao, Z. (2014). *Principal flows*. JASA.

Principal Curve§ and Examples¶

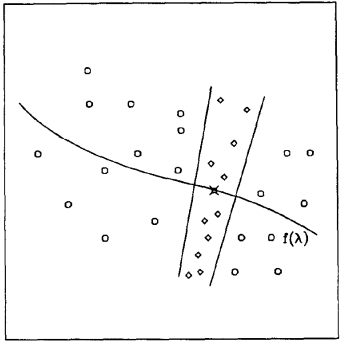
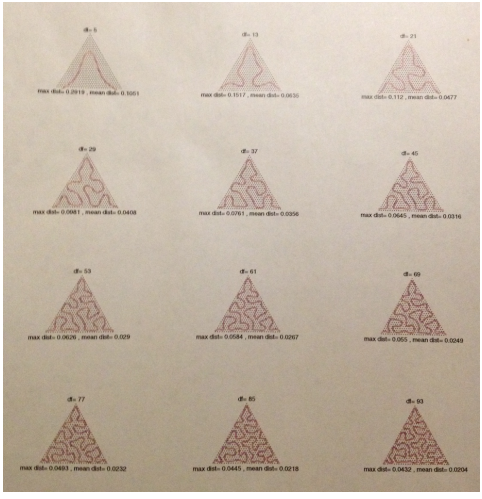


Figure (3.2) Each point on a principal curve is the average of the points that project there.



§Hastie, T., & Stuetzle, W. (1989). *Principal curves*. JASA.

¶Thanks to Trevor Hastie for sharing the examples

An ideal principal sub-manifold‖

(Ideal principal sub-manifold)[a] k-th dimensional principal sub-manifold

$$\arg \sup_{\mathcal{N}\in \mathrm{SubM}\left(A,\epsilon,k,\mathcal{M}\right)} \int_{B\in\mathcal{N}} \left(\cos(\alpha_B) \times \sum_{j=1}^{k} \lambda_j(B,\mathcal{M}) \right) d\mu_{\mathcal{N}},$$

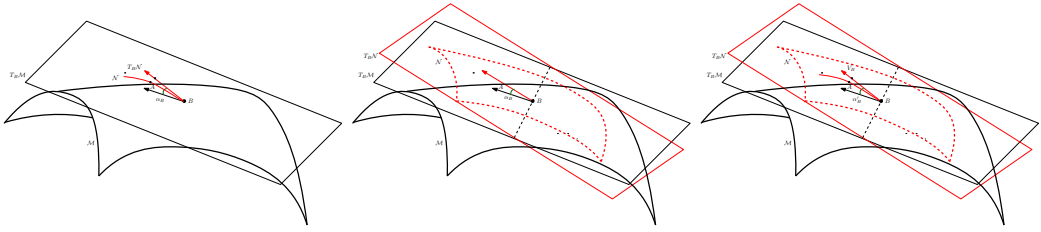
where $\mu_{\mathcal{N}}$ is the measure on $\mathcal{N}$

[a](subject to modification)

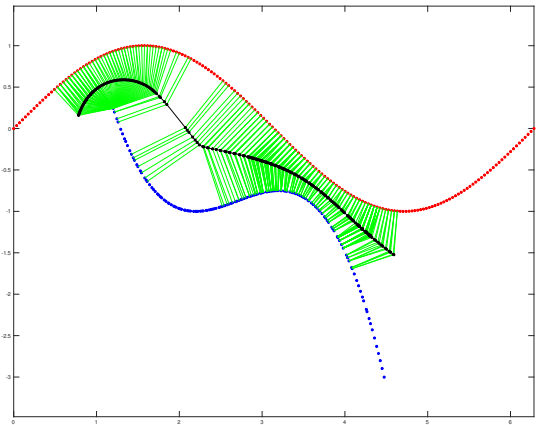
- To measure the degree of variation, we use the angle α_B between the hyperplane and tangent plane, $H_k(B,\mathcal{M})$ and $H_k(B,\mathcal{N})$.
- Theoretically, if $\alpha_B = 0$ for every B, then $H_k(B,\mathcal{M}) = H_k(B,\mathcal{N})$. For general cases, one would hope α_B is as small as possible.

‖Yao, Z., Eltzner, B., & Pham, T. (2016). *Principal sub-manifolds*. arXiv:1604.04318.

Principal flow and principal submanifold



Classification Boundary**



**Yao, Z., & Zhang, Z. (2019). *Principal boundary on Riemannian manifolds*. JASA.

A statistician's view on the connection to SYZ conjecture

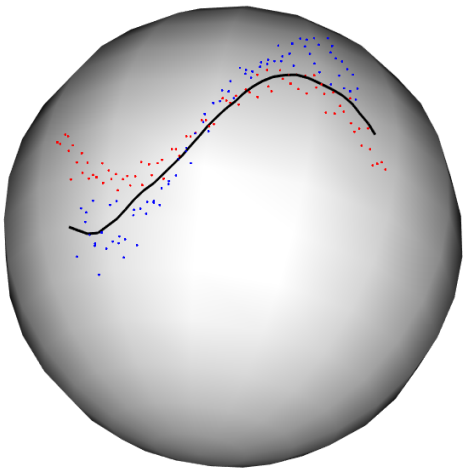
- The concept of finding a sub-manifold of the manifold data (data lying on manifold, e.g., a torus) is naturally rooted in a seemingly unrelated conjecture, namely, the SYZ conjecture[††]. Without diving into too many mathematical statements, the conjecture offers a geometrical way of breaking a complicated space (manifold) into its constituent parts.

- The problem is related to principal sub-manifold, which is an empirical calculation of such decomposition under some scenarios from the noisy data[‡‡].

[††] Strominger/Yau/Zaslow (1996)

[‡‡] Principal Sub-manifolds: New Theory and Methods (2023 manuscript)

A bit More Ambitious



An scRNA Clustering Project (On-going)

Accession Code	Author	Organism	Tissue	Number of Cell Types	Number of Cells	Number of Genes
E-MTAB-2600	Kolodziejczyk et.al [1]	Mus musculus	Embryonic stem cells	3	704	38658
E-MTAB-3321	Goolam et.al [2]	Mus musculus	Different cell stage embryos	5	124	41428
GSE36552	Yan et.al [3]	Homo sapiens	Embryonic stem cells	6	90	20214
GSE59739	Usoskin et.al [4]	Mus musculus	Lumbar dorsal root ganglion	4	622	25334
GSE60361	Zeisel et.al [5]	Mus musculus	Cerebral cortex	7	3005	19972
GSE67835	Darmanis et.al [6]	Homo sapiens	Adult and fetal human brain	9	466	22088
GSE81252	Camp et.al [7]	Homo sapiens	Liver bud	11	465	19020
GSE81547	Enge et.al [8]	Homo sapiens	Pancreas	7	2476	22256
GSE81608	Xin et.al [9]	Homo sapiens	Islet cells	8	1600	39851
GSE83139	Wang et.al [10]	Homo sapiens	Pancreatic endocrine cells	7	457	19950
GSE84133-Mouse	Baron et.al [11]	Mus musculus	Pancreas	13	1886	14878
GSE85241	Muraro et.al [12]	Homo sapiens	Pancreas	10	2126	19127
GSE103322	Puram et.al [13]	Homo sapiens	Oral cavity tumors	10	5902	23686
GSE108097	Han et.al [14]	Mus musculus	Major mouse organ types	11	6954	15006
EGAD00001010074	Nowicki-Osuch et.al [15]	Homo sapiens	Esophagus and stomach	5	3282	33234
GSE202352	Wiedemann et.al [16]	Homo sapiens	Hip, palm, and sole skin	4	2303	30933
16-WM8C	Joseph et.al [17]	Mus musculus	Prostate and urethra	2	1647	19492
GSE132042-Intestine	Schaum et.al [18]	Mus musculus	Intestine	5	1887	17985
GSE132042-Heart	Schaum et.al [18]	Mus musculus	Heart	6	906	21069
GSE132042-Liver	Schaum et.al [18]	Mus musculus	Liver	11	2859	21069
E-MTAB-11265	He et.al [19]	Homo sapiens	Embryonic and fetal lungs	5	649	16122
MAC-Bladder	Han et.al [20]	Mus musculus	Bladder	16	2746	20670
MAC-Brain	Han et.al [20]	Mus musculus	Brain	15	4038	16906
Midbrain	Siletti et.al [21]	Homo sapiens	Midbrain	11	4714	59357
SRP041736	Pollen et.al [22]	Homo sapiens	Cerebral cortex	11	249	14805

- 25 scRNA datasets
- Recent (10-15y) *CNS*
- $D \sim 15k$-$40k$

What is manifold fitting?

Geometric Whitney problem:

Given $\mathcal{A} \subset \mathbb{R}^D$, $d < D$, construct

$$\widehat{\mathcal{M}} \subset \mathbb{R}^D$$

to approximate $\mathcal{A}$, with $\dim(\widehat{\mathcal{M}}) = d$.

How well can $\widehat{\mathcal{M}}$ estimate $\mathcal{A}$ in terms of distance and smoothness?

Statistics and Data Science:

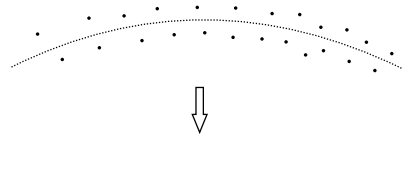
Let $\mathcal{M} \in \mathbb{R}^D$, $X \sim \mu(\mathcal{M})$, and

$$Y = X + \xi,$$

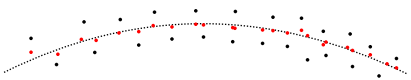
construct an estimator $\widehat{\mathcal{M}}$.

What are the bias/asymptotic properties of $\widehat{\mathcal{M}}$?

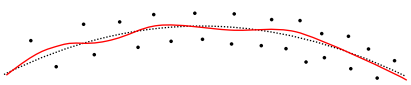
Benefit of manifold fitting



(a) Embedding



(b) Denoising



(c) Fitting

- Known manifold
- Unknown manifold → Manifold fitting:

Genovese et al (2012 a,b), Fefferman et al (2016), Mohammed/Narayanan (2017), Yao/Xia (2019), Fefferman et al (2021), Yao/Su/Li/Yau (2023)[a], Yao/Su/Yau (2024)[b], Yao/Li/Lu/Yau (2024)[c].

[a] *Manifold fitting*, arXiv:2304.07680.

[b] *Manifold fitting with CycleGAN*, PNAS.

[c] *Single-Cell Analysis via Manifold Fitting: A New Framework for RNA Clustering and Beyond*, revision at PNAS.

Manifold Distribution Principle

The fundamental principle of data science:

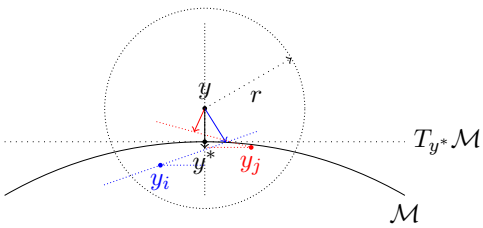
Each natural concept corresponds to a *dataset*, where each *sample* is a point in the dataset. The dataset is distributed near a low-dimensional manifold, which is called the *data manifold* $\mathcal{M}$. The data manifold $\mathcal{M}$ is embedded in a high-dimensional *ambient space* $\mathbb{R}^D$. The dataset can be abstracted as a probability distribution μ on the data manifold $\mathcal{M}$.

Namely,

$$y_i = x_i + \xi_i \quad \text{for} \quad i = 1, 2, ..., N$$

- $x_i \in \mathcal{M} \subset \mathbb{R}^D$: unobserved sample from $\mu(\mathcal{M})$
- $\xi_i \in \mathbb{R}^D$, $\xi_i \sim \phi_\sigma^{(D)}$: ambient space noise
- $y_i \in \mathbb{R}^D$, $y_i \sim \mu \star \phi_\sigma^{(D)}$: observation

Yao 2019* improves Fefferman 2018†



$$\widehat{\mathcal{M}} = \{y \in \mathbb{R}^D : d(y, \mathcal{M}) \le cr,$$
$$c < 1, \widehat{\Pi}_y^{\perp}(y - \widetilde{y}) = 0\}$$
$$\Rightarrow d(y, \mathcal{M}) \le Cr^2 \text{ for any } y \in \widehat{\mathcal{M}}$$

with probability

$$1 - d\exp\{-cNr^{d+2}\}.$$

- $r = \mathcal{O}(\sqrt{\sigma})$, $N \ge Cr^{-(d+2)}$
- $\widetilde{y} = \sum_i \alpha_i(y) y_i$: weighted mean of y_i
- $\widehat{\Pi}_y^{\perp} = \Pi_{hi}(\sum_i \alpha_i(y)\widehat{\Pi}_{y_i})$: estimator of $\Pi_{y^*}^{\perp}$

*Yao, Z., & Xia, Y. (2019). *Manifold fitting under unbounded noise.* arXiv:1909.10228.
†Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.

Yao 2019 – with more details

For a point y such that $d(y, \mathcal{M}) \leq cr$, $c < 1$, let

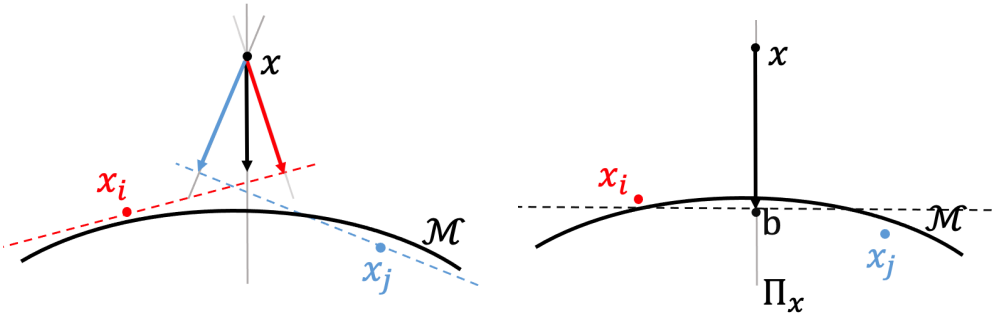
$$y - \widetilde{y} = \sum_i \alpha_i(y)(y - y_i),$$

with

$$\tilde{\alpha}_i(y) = \begin{cases} (1 - \frac{\|y-y_i\|_2^2}{r^2})^k, & \|y - y_i\|_2 \leq r \\ 0, & \|y - y_i\|_2 > r \end{cases},$$

$$\alpha_i(y) = \tilde{\alpha}_i(y) / \sum \tilde{\alpha}_i(y)$$

and $\widehat{\Pi}_{y_i} = I - VV^\top$, where V is the $D \times d$ matrix whose columns are the eigenvectors corresponding to the largest d eigenvalues of $\sum_{j \in I_{y_i, r'}} (y_j - y_i)(y_j - y_i)^\top$, $r' \geq 2r$.

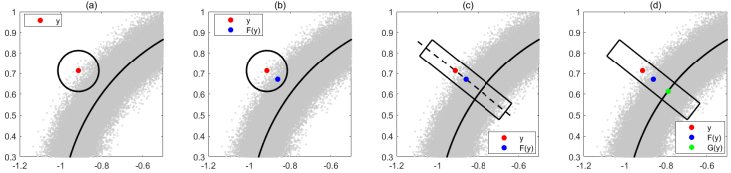


Difference: Π_x is used to estimate the orthogonal projection onto the normal space of $\mathcal{M}$ at x^*, the black dot $\mathbf{b}$ is used to estimate a point in $T_{x^*}\mathcal{M}$. Then the space $\{x' : \Pi_x(x' - \mathbf{b}) = \mathbf{0}\}$, illustrated as the black dashed line, approximates $T_{x^*}\mathcal{M}$, and the bias from x to the black dashed line is the estimated bias from x to $\mathcal{M}$, geometrically illustrated as the black arrow.

Local contraction‡: the fancy σ^2 bound

By setting

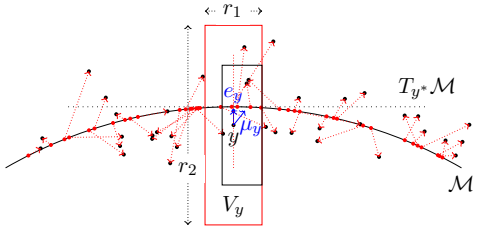
- $r_0 = C_1\sigma$
- $N = C_2 D r_0^{-d}\sigma^{-3}$
- $r_1 = c_1\sigma$
- $r_2 = C_3\sigma\sqrt{\log(1/\sigma)}$



Local contraction in two steps:

(1): estimate contraction direction;

(2): estimate local average.



‡Yao, Z., Su, J., Li, B. and Yau, S.T. *Manifold Fitting*. arXiv:2304.07680.

For a point y such that $d(y, \mathcal{M}) = \mathcal{O}(\sigma)$, let

$$F(y) = \sum \alpha_i(y) y_i,$$

with

$$\tilde{\alpha}_i(y) = \begin{cases} (1 - \frac{\|y - y_i\|_2^2}{r_0^2})^k, & \|y - y_i\|_2 \le r_0 \\ 0, & \|y - y_i\|_2 > r_0 \end{cases}, \quad \alpha_i(y) = \frac{\tilde{\alpha}_i(y)}{\sum \tilde{\alpha}_i(y)}$$

with $k \ge 2$ being a constant.

Theorem 1

For a point y such that $d(y, \mathcal{M}) = \mathcal{O}(\sigma)$,

$$\sin\{\Theta\left(F(y) - y,\ y^* - y\right)\} \le C\sigma\sqrt{\log(1/\sigma)},$$

for some constant C, with probability no less than $1 - C_1 \exp\{-C_2\sigma^c\}$.

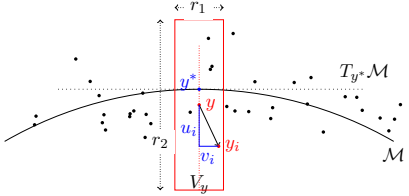
For the same point y, let $F_{\mathcal{M}}(y) = \sum \beta_i(y) y_i$, with

$$\tilde{\beta}_i(y) = \begin{cases} (1 - \frac{\|u_i\|_2^2}{r_2^2})^k (1 - \frac{\|v_i\|_2^2}{r_1^2})^k, & y_i \in \widehat{V}_y, \\ 0, & y_i \notin \widehat{V}_y, \end{cases}$$

$$\beta_i(y) = \tilde{\beta}_i(y) / \sum \tilde{\beta}_i(y),$$

where

$$u_i = \frac{(y - F(y))(y - F(y))^\top}{\|y - F(y)\|_2^2}(y - y_i), \quad v_i = y - y_i - u_i.$$



Theorem 2

For a point y such that $d(y, \mathcal{M}) = \mathcal{O}(\sigma)$,

$$\|F_{\mathcal{M}}(y) - y^*\|_2 \leq C\sigma^2 \log(1/\sigma),$$

for some constant C, with probability no less than $1 - C_1 \exp\{-C_2 \sigma^c\}$.

Construct manifold estimators

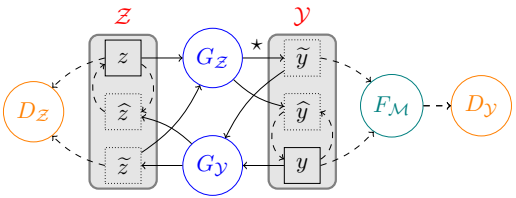
Theorem 3 (with initialization)

Suppose that $\widetilde{\mathcal{M}}$ *is a* d*-dimensional manifold with a positive reach* $\tau_0 \geq \tau$ *and* $d_H(\widetilde{\mathcal{M}}, \mathcal{M}) = O(\sigma)$*. Then, with high probability,* $\widehat{\mathcal{M}} = F_{\mathcal{M}}(\widetilde{\mathcal{M}})$ *is also a* d*-dimensional manifold that satisfies*

1. *For any point* $y \in \widehat{\mathcal{M}}$, $d(y, \mathcal{M}) \leq C\sigma^2 \log(1/\sigma)$.
2. *For any point* $x \in \mathcal{M}$, $d(x, \widehat{\mathcal{M}}) \leq C\sigma^2 \log(1/\sigma)$.
3. *For any two point* $y_1 \neq y_2 \in \widehat{\mathcal{M}}$, $\|y_1 - y_2\|_2^2 / d(y_2, T_{y_1}\widehat{\mathcal{M}}) \geq cr\tau$.

$\widetilde{\mathcal{M}} = \{y : d(y, \mathcal{M}) \leq C\sigma,\ \Pi^*(F(y) - y) = 0\}$.
Π^*: a pre-defined projection matrix with rank $D - d$.

CycleGAN/Manifold fitting framework¶



- $\mathcal{Z} \subset \mathbb{R}^d$: feature space
- $\mathcal{Y} \subset \mathbb{R}^D$: ambient space
- $G_{\mathcal{Z}}$, $G_{\mathcal{Y}}$: generators
- $D_{\mathcal{Z}}$, $D_{\mathcal{Y}}$: discriminators
- $F_{\mathcal{M}}$: manifold fitting sub-module

Main objective§: Let $Z \sim \mathrm{Unif}(0,1)^d$,

$$G^*_{\mathcal{Z}}(Z) := \min_{G_{\mathcal{Z}} \in \mathcal{C}(G_{\mathcal{Z}})} \mathrm{Div}(G_{\mathcal{Z}}(Z) \star \phi_\sigma, \nu),$$

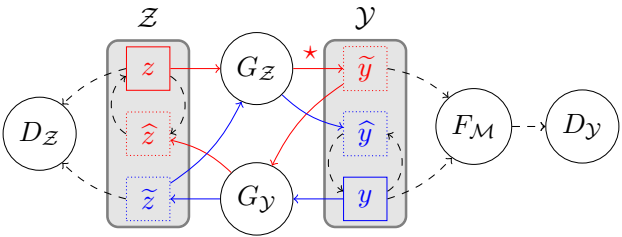
and estimate the latent manifold with

$$\widetilde{\mathcal{M}} := \widehat{G^*_{\mathcal{Z}}}(\mathcal{Z}), \quad \text{or} \quad \widehat{\mathcal{M}} := F_{\mathcal{M}} \circ \widehat{G^*_{\mathcal{Z}}}(\mathcal{Z}).$$

§ ν is the probability density of $Y \in \mathcal{Y}$ in the ambient space

¶Yao Z., Su J., and Yau S.T., *Manifold fitting with CycleGAN*, PNAS, Jan, 2024.

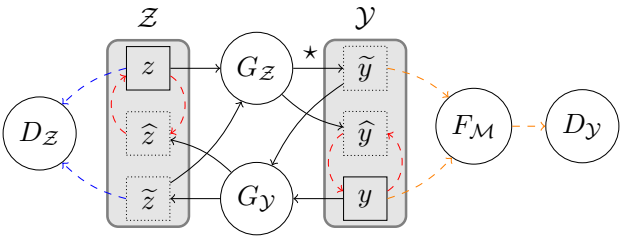
Forward step



- $\widetilde{y} = G_{\mathcal{Z}}(z) + \xi$
- $\widehat{z} = G_{\mathcal{Y}}(\widetilde{y})$
- $\xi \sim N(0, \sigma^2 I_D)$

- $\widetilde{z} = G_{\mathcal{Y}}(y)$
- $\widehat{y} = G_{\mathcal{Z}}(\widetilde{z})$

Loss functions (2 adversarial loss and 1 cycle loss)‖



- $\mathcal{L}_{\text{cycle}}(G_{\mathcal{Z}}, G_{\mathcal{Y}}) = \text{av.}\left(\|z_i - \widehat{z}_i\|_1\right) + \text{av.}\left(\|y_i - \widehat{y}_i\|_1\right),$
- $\mathcal{L}_{\mathcal{Y}\to\mathcal{Z}}(G_{\mathcal{Y}}, D_{\mathcal{Z}}) = \text{av.}\left([D_{\mathcal{Z}}(z_i) - 1]^2\right) + \text{av.}\left([D_{\mathcal{Z}}(\widetilde{z}_i) - 0]^2\right),$
- $\mathcal{L}_{\mathcal{Z}\to\mathcal{Y}}(G_{\mathcal{Z}}, D_{\mathcal{Y}}, F_{\mathcal{M}}) = \text{av.}\left([D_{\mathcal{Y}}(F_{\mathcal{M}}(y_i)) - 1]^2\right) + \text{av.}\left([D_{\mathcal{Y}}(F_{\mathcal{M}}(\widetilde{y}_i)) - 0]^2\right),$
- $\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{Y}\to\mathcal{Z}}(G_{\mathcal{Y}}, D_{\mathcal{Z}}) + \mathcal{L}_{\mathcal{Z}\to\mathcal{Y}}(G_{\mathcal{Z}}, D_{\mathcal{Y}}, F_{\mathcal{M}}) + \lambda\mathcal{L}_{\text{cycle}}(G_{\mathcal{Z}}, G_{\mathcal{Y}}).$

‖Given m batched sample z_i and n batched sample y_i; λ is negative parameter

Role of Generators

Solve (non-sample version):

$$G_{\mathcal{Z}}^*, G_{\mathcal{Y}}^* = \arg\max_{G_{\mathcal{Z}}, G_{\mathcal{Y}}} \min_{F_{\mathcal{M}}, D_{\mathcal{Z}}, D_{\mathcal{Y}}} \mathcal{L}(G_{\mathcal{Z}}, G_{\mathcal{Y}}, F_{\mathcal{M}}, D_{\mathcal{Z}}, D_{\mathcal{Y}}).$$

- Manifold estimators (sample-based):

  $\widetilde{\mathcal{M}} = \widehat{G_{\mathcal{Z}}^*}(\mathcal{Z})$ or $\widehat{\mathcal{M}} = F_{\mathcal{M}}(\widetilde{\mathcal{M}})$ estimates $\mathcal{M}$.

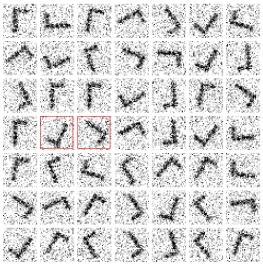
- Noise canceling:

  $\widehat{G_{\mathcal{Z}}^*} \circ \widehat{G_{\mathcal{Y}}^*}: \ y_i \mapsto \widehat{y}_i \in \widetilde{\mathcal{M}}$.

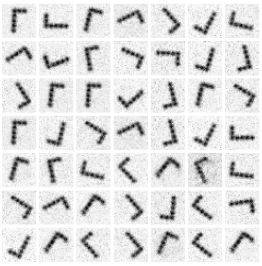
- Nonlinear interpolating:

  $\widehat{G_{\mathcal{Z}}^*}\left(t\widehat{G_{\mathcal{Y}}^*}(y_i) + (1-t)\widehat{G_{\mathcal{Y}}^*}(y_j)\right)$ nonlinear interpolates between $\widehat{y}_i$ and $\widehat{y}_j$.

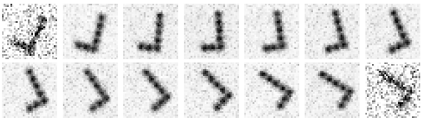
Fitting with 1D rotation group



(a)



(b)



(c)

(a) Images of a rotating simple shape, with ambient space noise.

(b) Denoised version of (a) with CycleGAN/Manifold Fitting model.

(c) Nonlinear interpolation of two examples with red boxes in (a).

Fitting in scRNA space – A real example

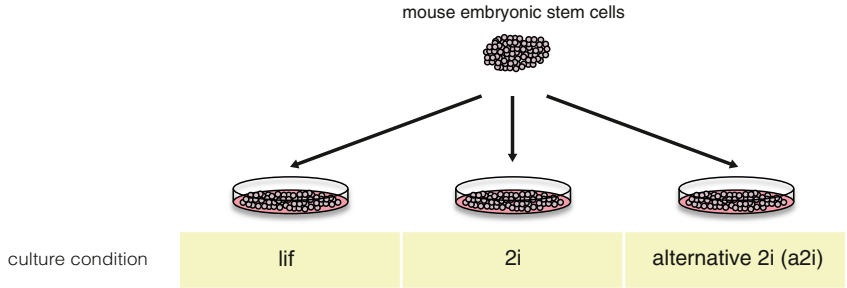
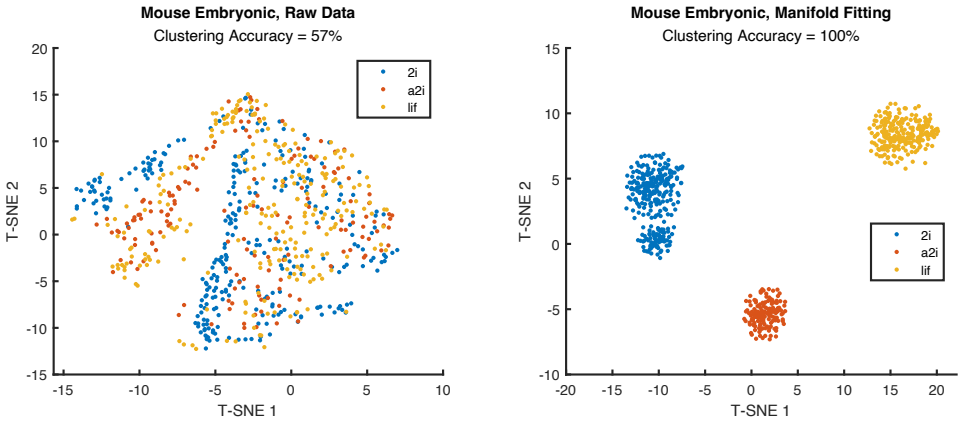


Figure: Mouse embryo stem cells (Kolodziejczyk et al. 2017) contain 704 cells in 3 classes (lif, 2i, a2i).

Focuses:

- Utilizing the potential molecular mechanisms governing cell differentiation and maintenance.
- Keeping the three classes of Mouse embryo stem cells.
- Improving other unsupervised clustering methods with the help of fitting.

Unsupervised clustering with tSNE



Both yx19 and ysl23 improve the spatial distribution of the data and the unsupervised clustering score for this data after fitting (→), significantly higher than the other methods without using fitting (←).

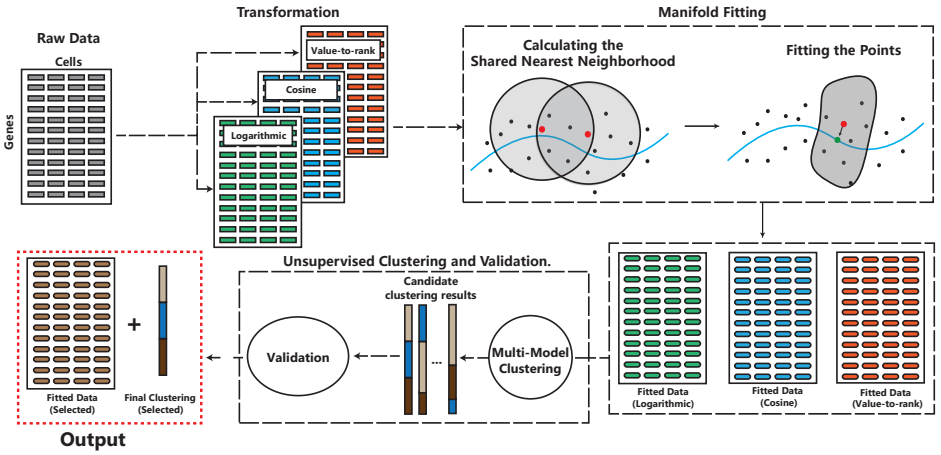
Summary (ARI) of 25 scRNA datasets ($D \sim 15k$-$40k$)

Data	scDHA Raw	Manifold Fitting
EMTAB2600	1.00	1.00
EMTAB3321	0.86	0.91
GSE36552	0.78	0.86
GSE59739	0.64	0.88
GSE60361	0.82	0.87
GSE67835	0.72	0.75
GSE81252	0.37	0.41
GSE81608	0.53	0.82
GSE83139	0.70	0.83
GSE84133-M	0.47	0.67
GSE85241	0.92	0.87
GSE103322	0.59	0.59
GSE108097	0.26	0.39
SRP041736	0.85	0.91
EGAD00001010074	0.46	0.55
GSE202352	0.42	0.73
16-WM8C	0.09	0.80
GSE132042	0.60	0.84
GSE132042-Liver	0.46	0.67
Midbrain	0.51	0.93
GSE81547	0.42	0.46
E-MTAB-11265	0.65	0.75
MAC-Bladder	0.51	0.57
Local11	0.47	0.82
MAC-Brain	0.13	0.83
Average	0.57	0.75

scDHA[a]: A leading scRNA clustering method.

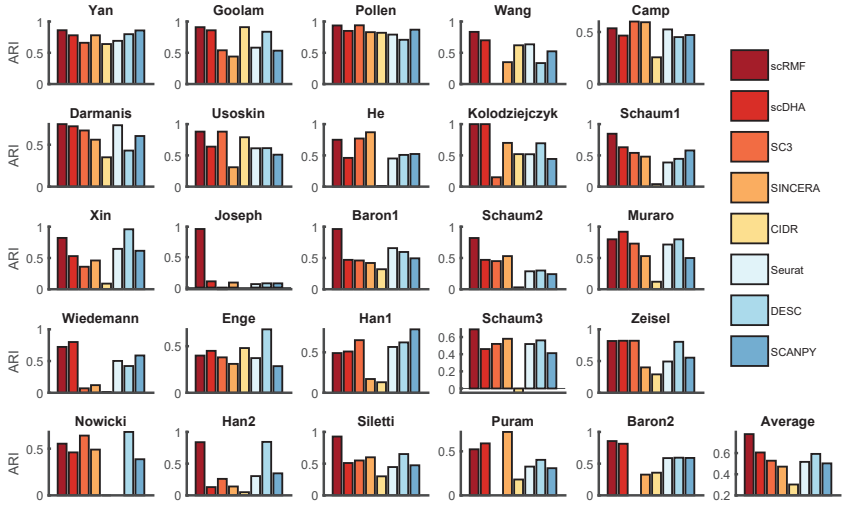
[a] Tran, Duc, et al. (2021). *Fast and precise single-cell data analysis using a hierarchical autoencoder*. Nature communications.

Overview of the scAMF pipeline**



**Yao Z., Li B., Lu Y., and Yau S.T.,*Single-Cell Analysis via Manifold Fitting: A New Framework for RNA Clustering and Beyond*, revision at PNAS.

Clustering performance of scAMF and other methods, measured by ARI.



Clustering with or without manifold fitting.