Manifold Fitting and Related Topics

Su Jiaji

College of Mathematics and Statistics
Chongqing University

October 16, 2025

Basic Information

Education

- 2018, *Zhejiang University*, Bachelor of Natural Science in Statistics
- 2024, *National University of Singapore*, PhD. in Statistics, Advisor: Zhigang Yao

Professional Experience

- 2022.08 – 2023.10, Student Reach Assistant, NUS
- 2023.10 – 2025.07, Research Fellow, NUS

Research

- Focus: Inference with singularity; Non-Euclidean data analysis;
- Outcome: Manifold fitting; principal nested submanifolds; their scientific applications *PNAS* $\times$ 2, *AoAS* $\times$ 1.

Origin of 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:

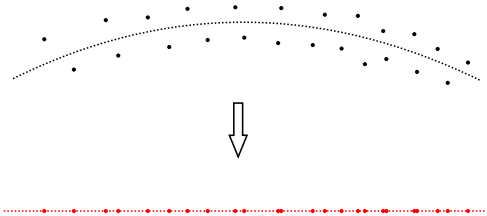
Let $\mathcal{M} \in \mathbb{R}^D$, $X \sim \mu(\mathcal{M})$, and

$$Y = X + \xi,$$

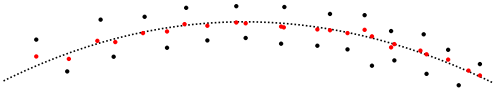
construct an estimator $\widehat{\mathcal{M}}$.

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

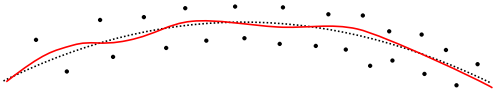
Target of manifold fitting



(a) Embedding



(b) Denoising



(c) Fitting

Illustrations for three types of manifold learning methods: embedding, denoising, and fitting.

Model setting

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

- $x_i \in \mathcal{M} \subset \mathbb{R}^D$, i.i.d sample from uniform distribution μ on $\mathcal{M}$
- $\mathcal{M}$: a d-dimensional, $\mathcal{C}^2$ boundaryless submanifold of $\mathbb{R}^D$
- The reach of $\mathcal{M}$ is at least τ
- $\xi_i \in \mathbb{R}^D$, $\xi_i \sim \phi_\sigma^{(D)}$
- $y_i \in \mathbb{R}^D$, $y_i \sim \mu \star \phi_\sigma^{(D)} =: \nu$

Reach: the largest τ such that $\forall a$ with $d(a, \mathcal{M}) < \tau$ has a unique nearest point in $\mathcal{M}$.

Nonparametric methods*

From the perspective of minimax risk:

$$\inf_{\widehat{\mathcal{M}}} \sup_{\nu} \mathbb{E}_{\nu} \left[d_H(\widehat{\mathcal{M}}, \mathcal{M}) \right].$$

Lower bound: Le Cam Method

Let $\nu_0 = \mu(\mathcal{M}_0) \star \phi$, $\nu_1 = \mu(\mathcal{M}_1) \star \phi$. For any $\widehat{\mathcal{M}}$,

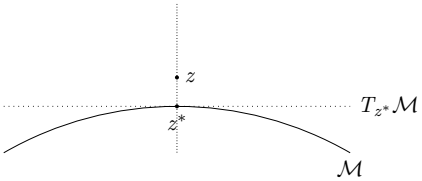
$$\sup_{\nu} \mathbb{E}_{\nu} \left[d_H(\widehat{\mathcal{M}}, \mathcal{M}) \right] \geq d_H(\mathcal{M}_0, \mathcal{M}_1) \frac{1}{8} (1 - \mathsf{TV}(\nu_0, \nu_1))^{2N}.$$

$\mathsf{TV}(P, Q) = \int |p - q|/2$: the total variation distance between P, Q.

Upper bound: find a proper $\widehat{\mathcal{M}}$

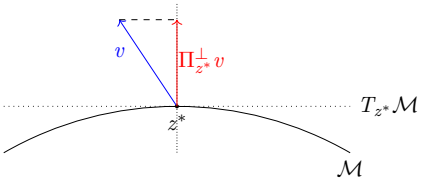
* Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., & Wasserman, L. *Minimax manifold estimation,* JMLR, 2012
Manifold estimation and singular deconvolution under Hausdorff loss. AoS, 2012; *Nonparametric ridge estimation,* AoS, 2014

Notations and definitions



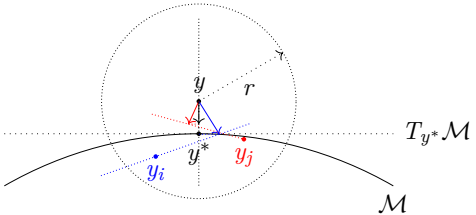
- z: a point of interest
- $z^* = \arg\min_{z' \in \mathcal{M}} \mathrm{d}(z, z')$: projection of z on $\mathcal{M}$
- $T_{z^*}\mathcal{M}$: tangent space of $\mathcal{M}$ at z^*

Notations and definitions



- z: a point of interest
- $z^* = \arg\min_{z' \in \mathcal{M}} \mathsf{d}(z, z')$: projection of z on $\mathcal{M}$
- $T_{z^*}\mathcal{M}$: tangent space of $\mathcal{M}$ at z^*
- $\Pi^{\perp}_{z^*}$: projection matrix onto the normal space of $T_{z^*}\mathcal{M}$
- $\widehat{\Pi}^{\perp}_{z}$: estimator of $\Pi^{\perp}_{z^*}$

Fefferman et al. 2018†



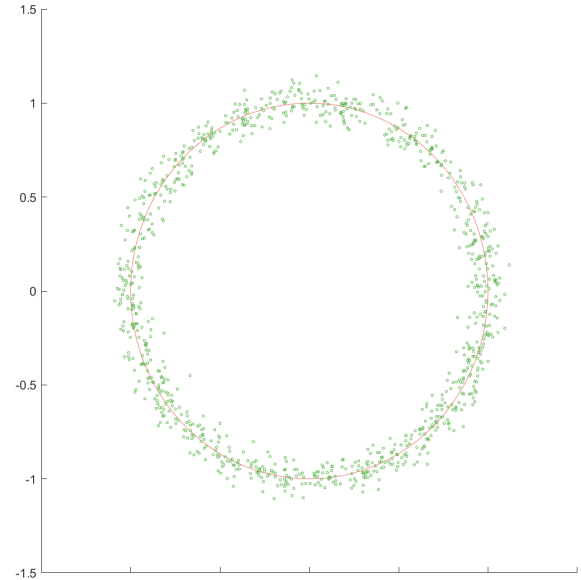
$$\mathcal{M}_o = \{y \in \mathbb{R}^D : d(y, \mathcal{M}) \le cr, \widehat{\Pi}_y^{\perp} F(y) = 0\}$$

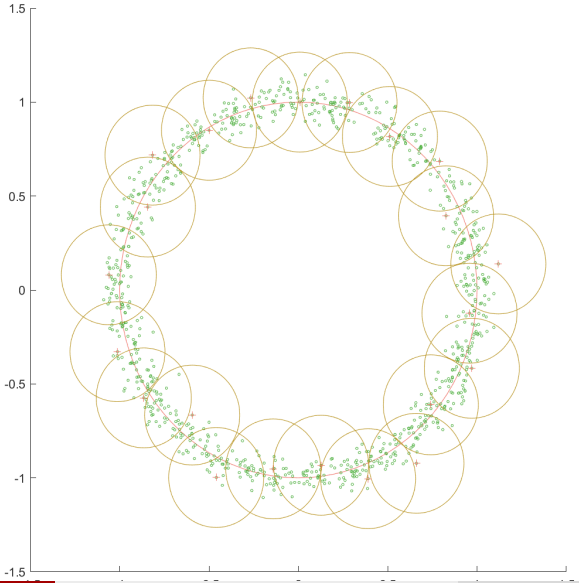
$$\Rightarrow d_H(\mathcal{M}, \mathcal{M}_o) \le \frac{Cdr^2}{\tau}$$

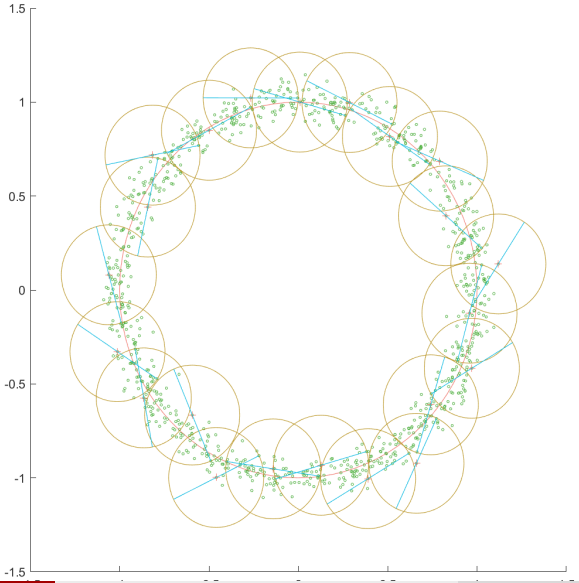
with probability $1 - N^{-C}$

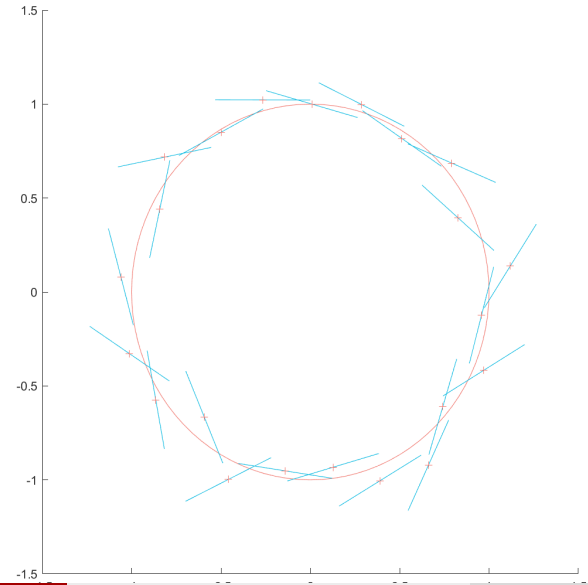
- y_i, y_j: points in the cr/d-net only
- $r = C\sqrt{\sigma}$; $N/\log(N) \ge Cr^{-2d}$;
- $F(y) = \sum_i \alpha_i(y)\widehat{\Pi}_{y_i}^{\perp}(y - y_i)$
- $\widehat{\Pi}_y^{\perp} = \Pi_{hi}(\sum_i \alpha_i(y)\widehat{\Pi}_{y_i})$: estimator of $\Pi_{y^*}^{\perp}$
- Π_{hi}: projection onto the span of the eigenvectors corresponding to the largest $D - d$ eigenvalues

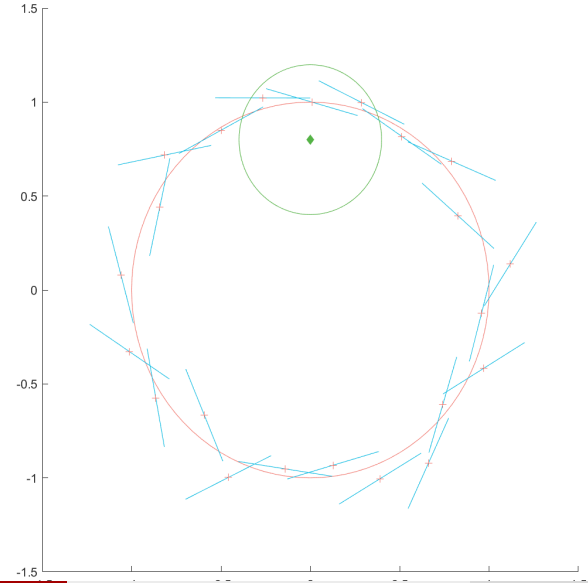
† Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* COLT.

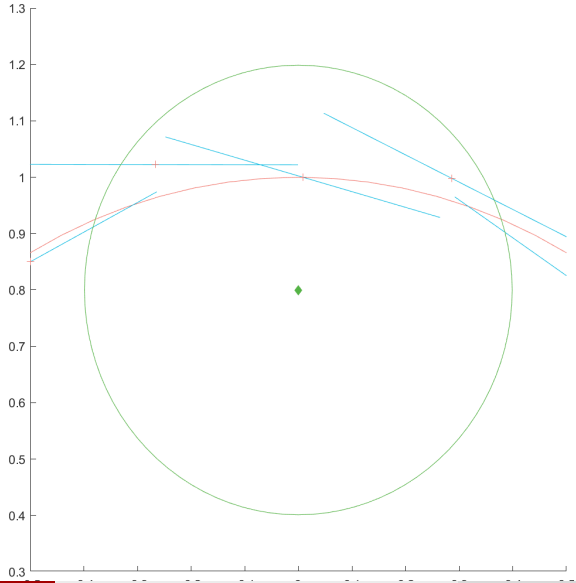

1.5
1
0.5
0
-0.5
-1
-1.5

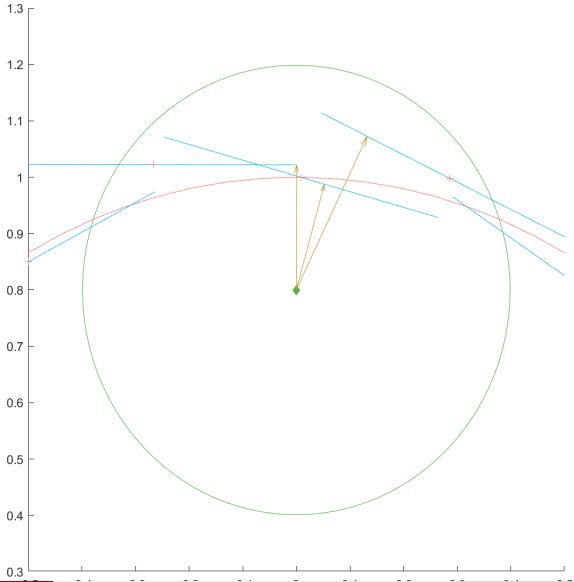

1.5
1
0.5
0
-0.5
-1
-1.5

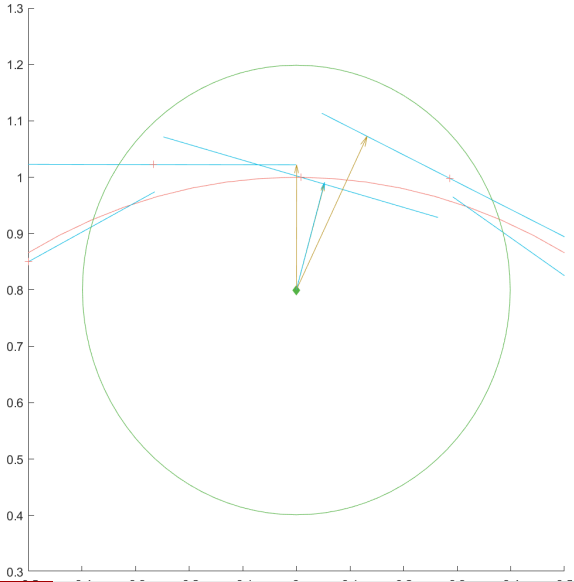

1.5
1
0.5
0
-0.5
-1
-1.5

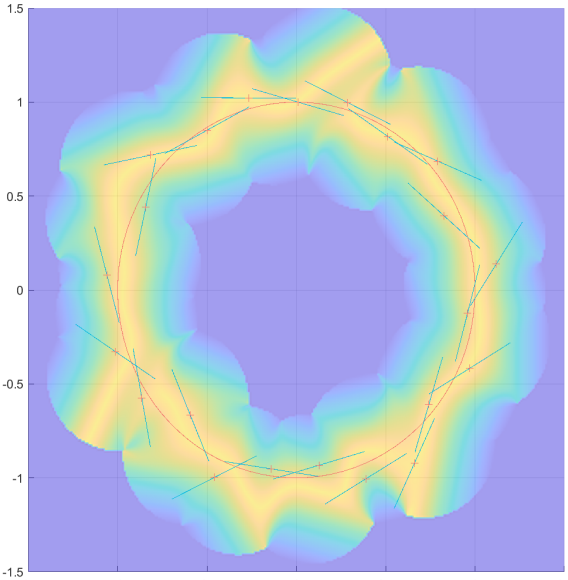

1.5
1
0.5
0
-0.5
-1
-1.5


1.5
1
0.5
0
-0.5
-1
-1.5

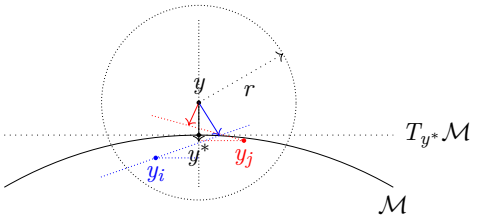

1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3


1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3


1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3


1.5
1
0.5
0
-0.5
-1
-1.5

Yao 2019‡ improves Fefferman 2018



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

$$\widehat{\mathcal{M}} = \{y \in \mathbb{R}^D : d(y, \mathcal{M}) \leq cr, \widehat{\Pi}_y^{\perp}(y - \widetilde{y}) = 0\}$$

$$\Rightarrow d(y, \mathcal{M}) \leq Cr^2 \text{ for any } y \in \widehat{\mathcal{M}}$$

with probability

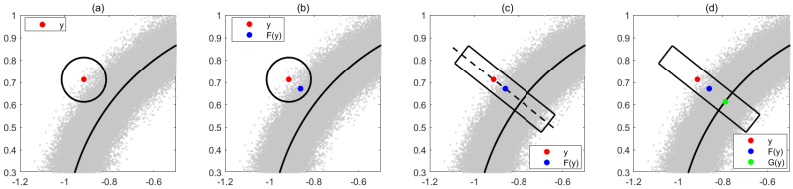
$$1 - d\exp\{-cNr^{d+2}\}.$$

‡ Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR.

Two-step local contraction§

By setting

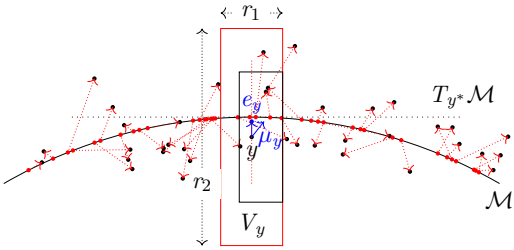
- $r_0 = C_1\sigma$
- $N = C_2 D r_0^{-d}\sigma^{-3}$
- $r_1 = c_1\sigma$
- $r_2 = C_3\sigma\sqrt{\log(1/\sigma)}$



Local contraction in two steps:

(i): estimate contraction direction;

(ii): estimate local average.



§ Yao, Z., Su, J., Li, B., & Yau, S. T. (2023). *Manifold fitting*. arXiv preprint.

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

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

with

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

with $k \ge 2$ being a constant.

Theorem

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 $1 - C_1 \exp\{-C_2\sigma^c\}$.

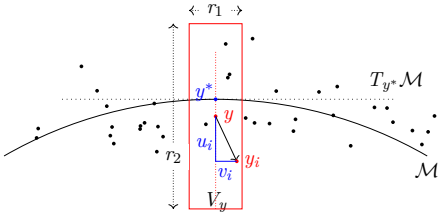
For the same point y, let $F_{\mathcal{M}}(y) = \sum \beta_i(y) y_i$, with

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

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

where

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



Theorem

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 (with initialization)

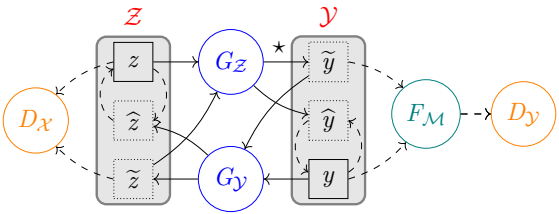
Suppose that $\widetilde{\mathcal{M}}$ is a d-dimensional manifold with a positive reach $\tau_0 \geq \tau$ and $d_H(\widetilde{\mathcal{M}}, \mathcal{M}) = O(\sigma)$. Then, with high probability, $\widehat{\mathcal{M}} = F_{\mathcal{M}}(\widetilde{\mathcal{M}})$ is also a d-dimensional manifold that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

Combine with generative model[¶]



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

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

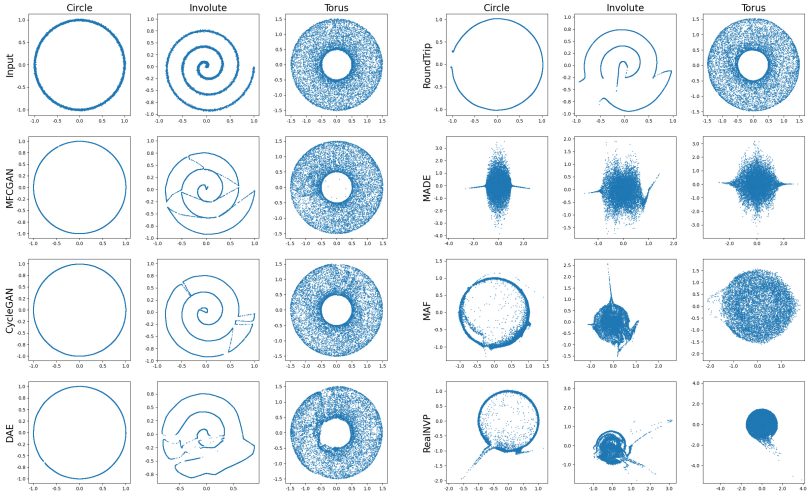
$$\min_{G_{\mathcal{Z}} \in \mathcal{C}(G_{\mathcal{Z}})} \mathrm{Div}(G_{\mathcal{Z}}(Z) \star \phi_\sigma, \nu),$$

and estimate the latent manifold with

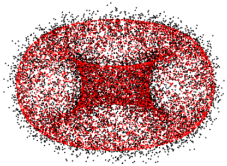
$$\widehat{\mathcal{M}} := \widehat{G^*_{\mathcal{Z}}}(\mathcal{Z}), \quad \text{or} \quad \widehat{\mathcal{M}} := F_{\mathcal{M}} \circ \widehat{G^*_{\mathcal{Z}}}(\mathcal{Z}).$$

[¶] Yao, Z., Su, J., & Yau, S. T. (2024). *Manifold fitting with CycleGAN*. PNAS.

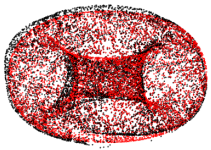
Comparison with neural network-based methods



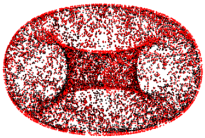
Comparison with established manifold fitting methods



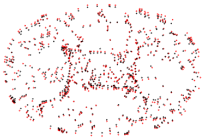
(d) Input



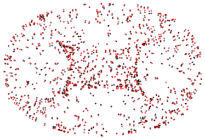
(e) w/o $F_{\mathcal{M}}$



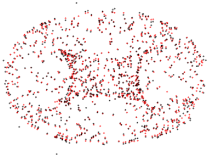
(f) w/ $F_{\mathcal{M}}$



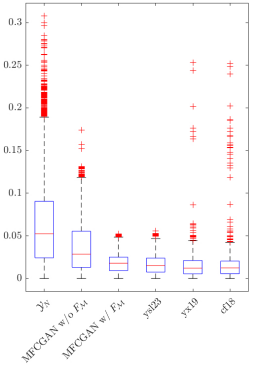
(g) ysl23



(h) yx19



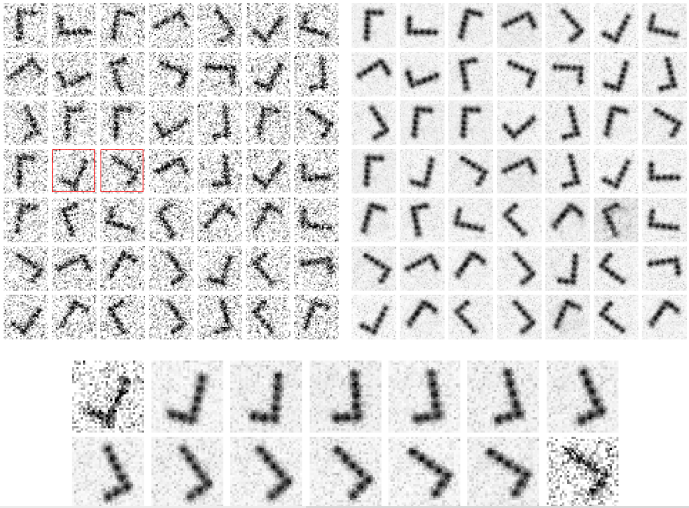
(i) cf18



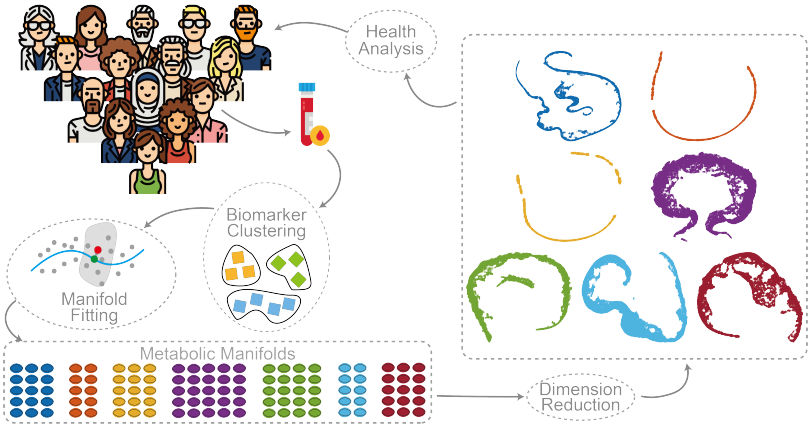
(j) Distances

Fitting with 1D rotation group

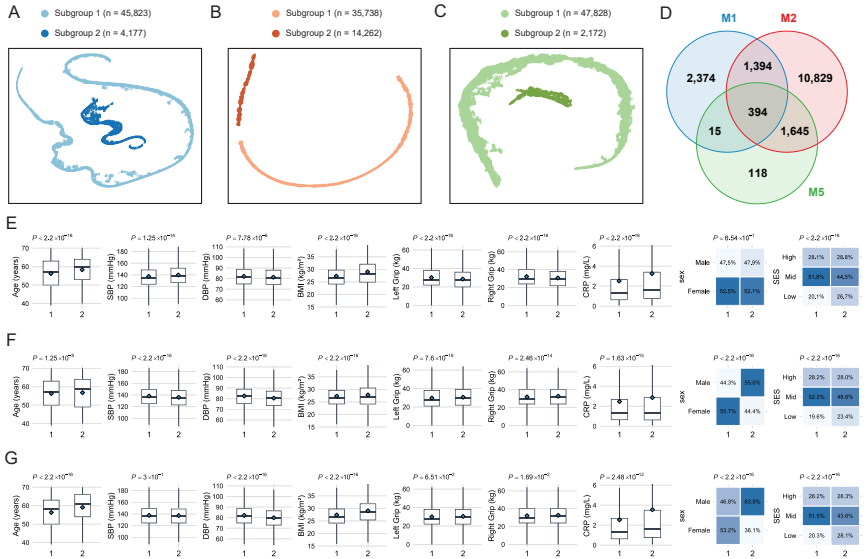
- ↖: Images of a rotating simple shape, with ambient space noise.
- ↗: Denoising with $\widehat{G^*_{\mathcal{Z}}} \circ \widehat{G^*_{\mathcal{Y}}}$.
- ↓: Nonlinear interpolation of two examples with red boxes.

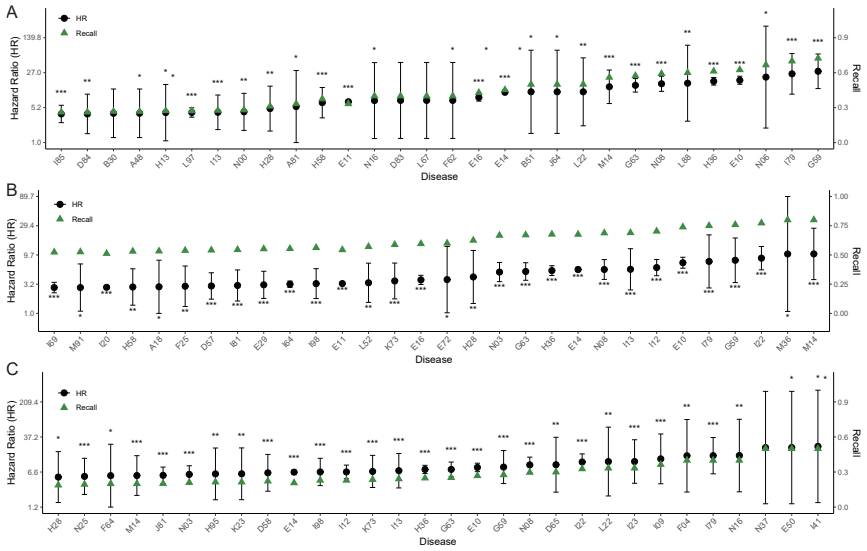


Manifold Fitting in NMR Metabolic Biomarkers ‖



‖ Li, B., Su, J., Lin, R., Yau, S. T., & Yao, Z. (2025). *Manifold fitting reveals metabolomic heterogeneity and disease associations in UK Biobank populations.* PNAS.

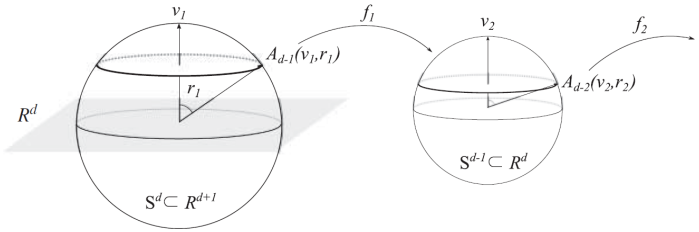

A
Subgroup 1 (n = 45,823)
Subgroup 2 (n = 4,177)
B
Subgroup 1 (n = 35,738)
Subgroup 2 (n = 14,262)
C
Subgroup 1 (n = 47,828)
Subgroup 2 (n = 2,172)
D
M1
M2
M5
2,374
1,394
10,829
394
15
1,645
118
E
F
G
Age (years)
SBP (mmHg)
DBP (mmHg)
BMI (kg/m²)
Left Grip (kg)
Right Grip (kg)
CRP (mg/L)
sex
SES
Male
Female
High
Mid
Low


A
HR
Recall
Hazard Ratio (HR)
139.8
27.0
5.2
1.0
Recall
0.9
0.6
0.3
0.0
I85 D84 B30 A46 H13 L97 I13 N00 H28 A81 H35 E11 N16 D83 L97 R62 E16 E14 B51 J64 L22 M14 G63 N06 L88 H36 E10 N06 I79 G59
Disease
B
HR
Recall
Hazard Ratio (HR)
89.7
29.4
9.7
3.2
1.0
Recall
1.00
0.75
0.50
0.25
0.00
I69 M61 I20 H58 A18 F25 D57 I81 E29 I64 I88 E11 L52 K73 E16 E72 H28 N63 G63 H36 E14 N08 I15 I12 E10 I79 G59 I22 M36 M14
Disease
C
HR
Recall
Hazard Ratio (HR)
209.4
37.2
6.6
1.2
Recall
0.9
0.6
0.3
0.0
H28 N25 F64 M14 J81 N03 H85 K23 D38 E14 I88 I12 K73 I13 H36 G63 E10 G59 N08 D65 I22 L22 I23 I09 F04 I79 N16 N37 E50 I41
Disease

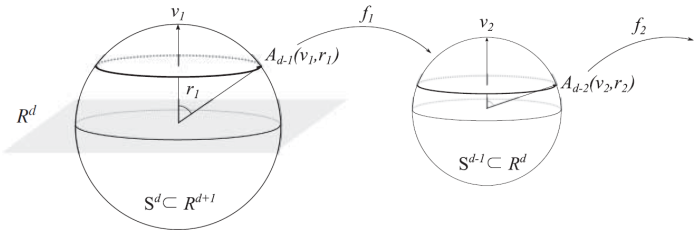
Principal Nested Spheres*

- $\{x_i^{(0)}\}_{i=1}^n \subset \mathcal{S}^D \subset \mathbb{R}^{D+1}$
- For $p, q \in \mathcal{S}^D$, $\mathrm{d}(p, q) = \arccos(p^\top q)$
- With $v \in \mathcal{S}^D$ and $r \in (0, \pi/2]$, a sub-sphere of $\mathcal{S}^D$:

$$A_{D-1}(v, r) = \{x \in \mathcal{S}^D : \mathrm{d}(v, x) = r\} = \mathcal{S}^D \cap \{x \in \mathbb{R}^{D+1} : v^\top x = \cos(r)\}$$



* Jung, S., Dryden, I. L., & Marron, J. S. (2012). *Analysis of principal nested spheres.* Biometrika.

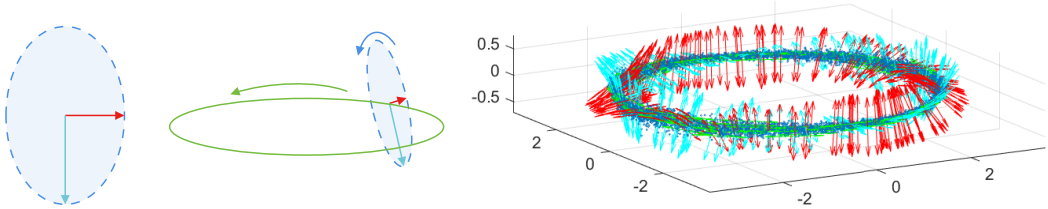


- Let $\mathcal{L}_j(v,r) = \sum_{i=1}^n \left(\cos^{-1}(v^\top x_i^{(j-1)}) - r\right)^2$, and

$$(\widehat{v}_j, \widehat{r}_j) = \underset{v \in \mathcal{S}^{D-j+1}, r \in (0,2\pi]}{\arg\min} \mathcal{L}_{j-1}(v,r).$$

- $\widehat{A}_{D-j} = A_{D-j}(\widehat{v}_j, \widehat{r}_j)$
- project $x_i^{(j-1)}$ along the geodesic to $\widehat{A}_{D-j}$
- re-scale and rewrite coordinates to make $\{x_i^{(j)}\} \subset \mathcal{S}^{D-j}$

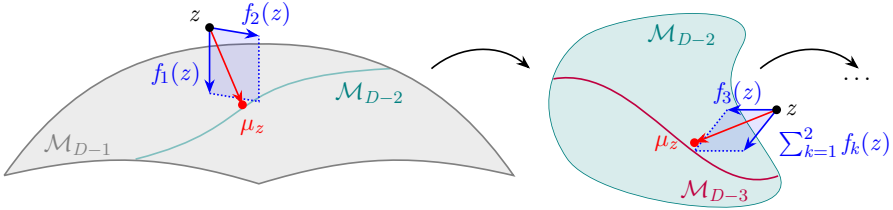
Intuitions: Decompose based on Smooth Covariance Structures



Fitting submanifolds such that

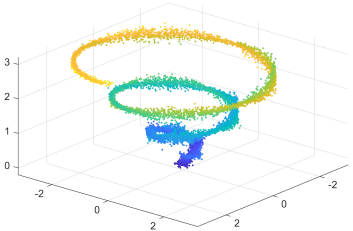
- tangent spaces approximate the linear space generated by principal directions
- passing through the 'middle' of the data clouds
- with different dimensionalities and a nested structure

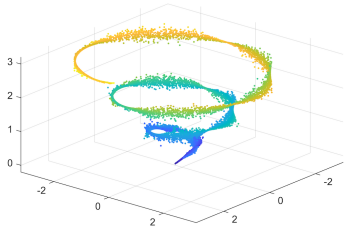
Intuition of Principal Nested Submanifolds†

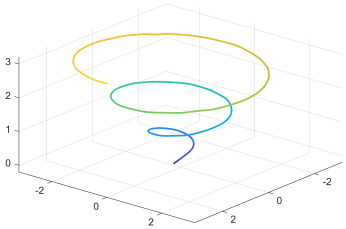


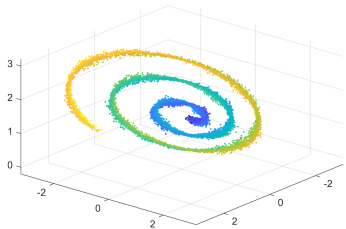
† Su, J., & Yao, Z. (2025). *Principal Decomposition with Nested Submanifolds*. arXiv preprint.

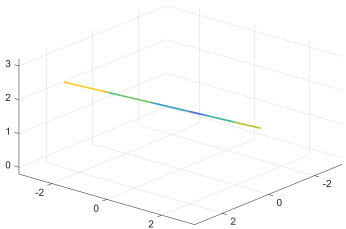
Simulation with Euclidean Space



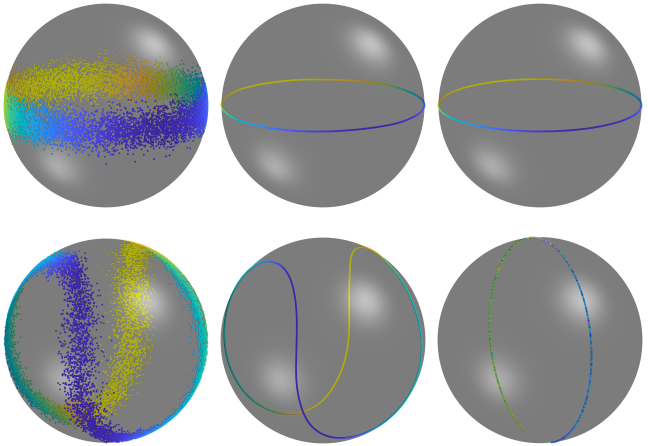


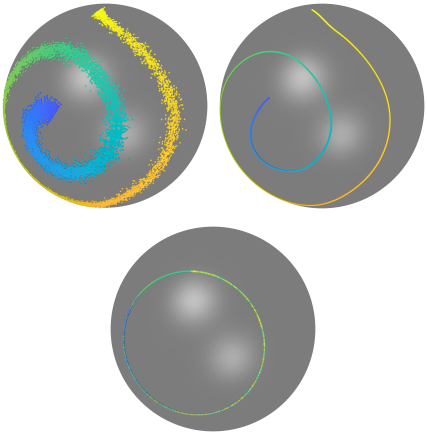






Simulation with Spheres





Conclusion & Outlook

- **Manifold Fitting**: a nonlinear, data-driven generalization of PCA.
  - Recovers low-dimensional structures from data
  - Easy to be integrated into other pipelines
- Future directions:
  - Refinement of manifold fitting theory
  - Scalable versions for large-scale omics data
  - Integration into neural networks to solve scientific problems

Thank you!
Questions are welcome.

Slides and preprint:

sujiaji.cn/nav