Principal Decomposition with Nested Submanifolds

Su Jiaji

Department of Statistics and Data Science
National University of Singapore

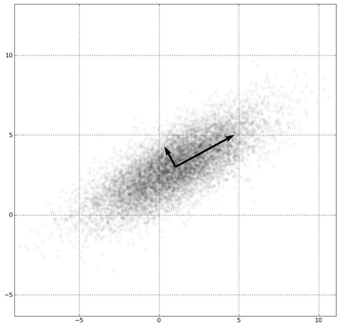
May 30, 2025
Yau Mathematical Sciences Center

Principal Component Analysis

- $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$: points of interest
- $\mathcal{A}_{D-1} = \{z \in \mathbb{R}^D : v_D^\top (z - p) = 0\}$ with $\| V_D \| = 1$
- minimize $\sum_i \mathrm{d}(x_i, \mathcal{A}_{D-1})^2 = \sum_i \left(v_D^\top (x_i - p) \right)^2$
- One solution: $p = \bar{x}$,

$$v_D = \underset{v \in \mathbb{R}^D,\, \|v\|=1}{\arg\min}\; v^\top \sum_i (x_i - \bar{x})(x_i - \bar{x})^\top v$$

- Remaining components: $x_i' = x_i - v_D v_D^\top (x_i - \bar{x})$.
- Repeat with $\{x_i'\} \subset \mathcal{A}_{D-1}$ ...



Principal Component Analysis

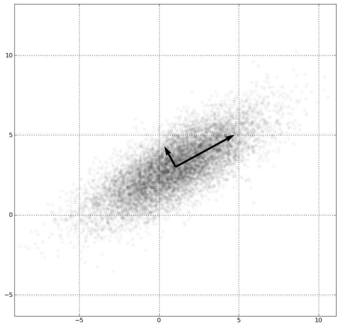
- $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$: points of interest
- $\mathcal{A}_{D-1} = \{z \in \mathbb{R}^D : v_D^\top (z - p) = 0\}$ with $\| V_D \| = 1$
- minimize $\sum_i \mathrm{d}(x_i, \mathcal{A}_{D-1})^2 = \sum_i \left(v_D^\top (x_i - p)\right)^2$
- One solution: $p = \bar{x}$,

$$v_D = \underset{v \in \mathbb{R}^D,\, \|v\|=1}{\arg\min} \; v^\top \sum_i (x_i - \bar{x})(x_i - \bar{x})^\top v$$

- Remaining components: $x_i' = x_i - v_D v_D^\top (x_i - \bar{x})$.
- Repeat with $\{x_i'\} \subset \mathcal{A}_{D-1}$ ...

$\hookrightarrow \mathcal{A}_d = \bar{x} + \mathrm{span}(v_1, \ldots, v_d)$

$$\{\bar{x}\} \subset \mathcal{A}_1 \subset \cdots \mathcal{A}_{D-1} \subset \mathbb{R}^D: \text{ a } \textit{flag}.$$



Linear Decomposition can be Inadequate

- Data can take values on non-Euclidean spaces:
  - resides in known non-linear spaces:
    $\hookrightarrow$ compositional data; semi-positive definite matrices ...
  - has non-Euclidean topology:
    $\hookrightarrow$ directions; molecular geometry ...
- or distributed around lower dimensional structures
  $\hookrightarrow$ strongly nonlinear covariance structures ...

Additive Principal Component*

For $X = (X_1, \ldots, X_D)^\top$, consider $\phi_i = \phi_i(X_i)$ s.t.

$$\phi_i \in H_i \subset \{\phi : \mathbb{E}[\phi(X)] = 0,\ \text{Var}[\phi(X)] < \infty\},$$

$$\Phi := (\phi_1, \ldots, \phi_D) \in H_1 \times \cdots \times H_D.$$

Target: minimize $\text{Var}(\sum_i \phi_i)$ subject to $\sum_i \text{Var}(\phi_i) = 1$.

- Use $\sum \phi_i(X_i)$ instead of $\sum a_i X_i$
- Focus on the smallest PC/low-variance relationships
- Find $\sum \phi_i(X_i) \approx 0$
- $\{x = (x_1, \ldots, x_D) \in \mathbb{R}^D : \sum \phi_i(x_i) = 0\}$: an *additive manifold with co-dimensional 1*

*Donnell, D. J., Buja, A., & Stuetzle, W. (1994). *Analysis of additive dependencies and concurvities using smallest additive principal components*. AoS

Principal Curves[†]

Find a smooth curve passing the 'middle' of the data set/minimizing variation orthogonal to the curve

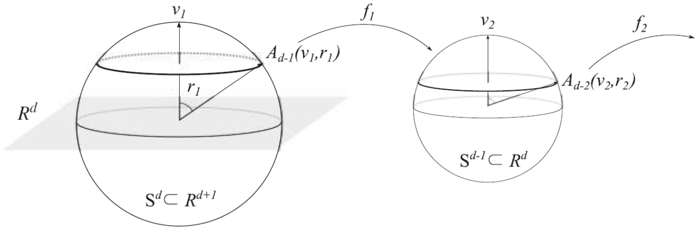
- $\gamma(t)$: a smooth curve with parameter t
- $\gamma^{-1}(x) = \arg\min_t \|x - \gamma(t)\|$
- Principal curve: $\mathbb{E}[X \mid \gamma^{-1}(X) = t] = \gamma(t)$
- Estimator:
  Let $\gamma_0(t) = \bar{x} + vt$, define
  $$\gamma_j(t) = \mathbb{E}[X \mid \gamma_{j-1}^{-1}(X) = t]$$

[†] Hastie, T., & Stuetzle, W. (1989). *Principal curves*. JASA.

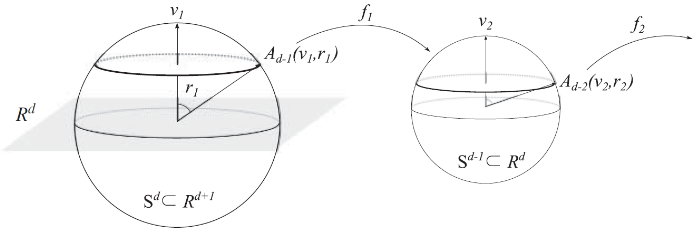
Principal Nested Spheres[‡]

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

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



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

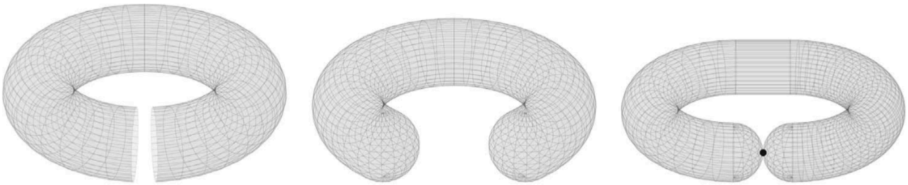


- Let $\mathcal{L}_j(v,r) = \sum_{i=1}^n \left(\cos^{-1}(v^\top x_i^{(j-1)}) - r\right)^2$, and

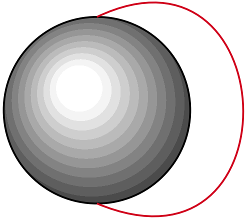
$$(\widehat{v}_j, \widehat{r}_j) = \underset{v \in \mathcal{S}^{D-j+1}, r \in (0,2\pi]}{\arg\min} \mathcal{L}_{j-1}(v,r).$$

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

Torus PCA[§]

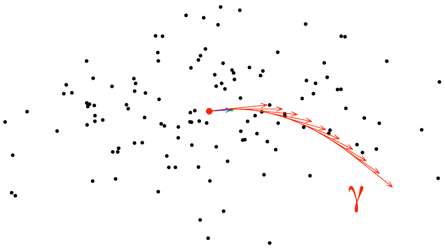


- Data on torus: $T^D = \mathcal{S}^1 \times \cdots \times \mathcal{S}^1$
- Have gaps on most dimensions
- Cut into special spheres and perform PNS



[§]Eltzner, B., Huckemann, S., & Mardia, K. V. (2018). *Torus principal component analysis with applications to RNA structure*. AoAS

Principal Flows[¶]

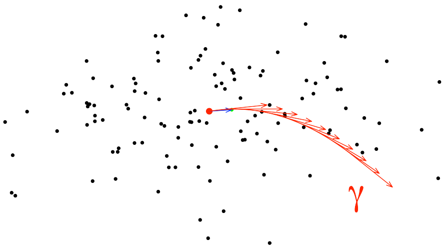


Find a smooth curve $\gamma(t)$ s.t.

- residing on known Riemannian manifold $(\mathcal{M}, g)$
- passing the 'middle' of data

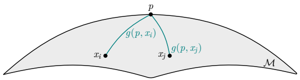
[¶] Panaretos, V. M., Pham, T., & Yao, Z. (2014). *Principal flows*. JASA.

Principal Flows[¶]



Find a smooth curve $\gamma(t)$ s.t.

- residing on known Riemannian manifold $(\mathcal{M}, g)$
- passing the 'middle' of data
- start from $\gamma(0)$, the *Fréchet mean*



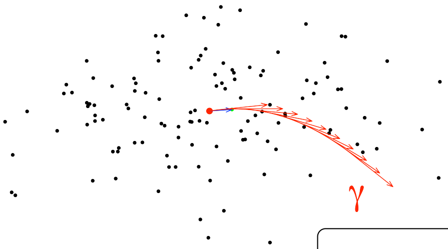
$$\Psi(p) = \sum g^2(p, x_i)/N$$

Fréchet mean: $\arg\min_{p\in\mathcal{M}} \Psi(p)$

Fréchet variance: $\min_{p\in\mathcal{M}} \Psi(p)$

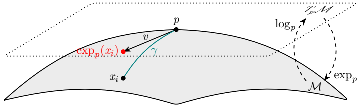
[¶] Panaretos, V. M., Pham, T., & Yao, Z. (2014). *Principal flows*. JASA.

Principal Flows[¶]



Find a smooth curve $\gamma(t)$ s.t.

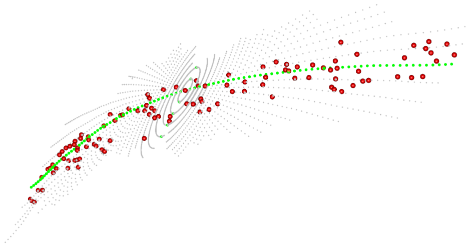
- residing on known Riemannian manifold $(\mathcal{M}, g)$
- passing the 'middle' of data
- start from $\gamma(0)$, the *Fréchet mean*
- $\dot{\gamma}$ fits the first principal direction of the *tangent space PCA* at $\gamma(t)$.



Project x_i's from $\mathcal{M}$ to $T_{\gamma(t)}\mathcal{M}$;
Perform local PCA at $\gamma(t)$ in $T_{\gamma(t)}\mathcal{M}$.

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

Principal Sub-manifold[‖]



Find a smooth sub-manifold s.t.

- residing on known Riemannian manifolds; passing the 'middle' of data
- start from the Fréchet mean
- the ray $\dot{\gamma}$ fits the principal direction derived from the tangent space PCA at $\gamma(t)$
- the collection of all such rays is the principal sub-manifold.

[‖] Yao, Z., Eltzner, B., & Pham, T. (2016–2024). *Principal sub-manifolds*. Statistica Sinica, to appear.

Some Issues

The mean curves and surfaces:

- Most of them are with dimension 1 or co-dimension 1
- Hard to extend, in terms of dimensionality
- Lack of ‘flag’ structure

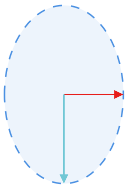
The PCA on spheres, tori or *simplex*:

- Restricted by the form of data
- Inherit another form of ‘linear constraint’

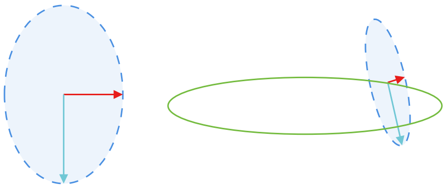
Inspirations:

- *Backward* is more natural
- *Principal directions* are more flexible

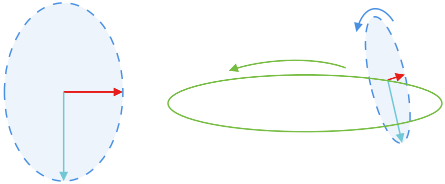
Intuitions: Decompose based on Smooth Covariance Structures



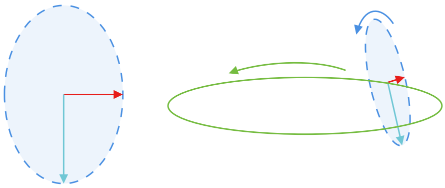
Intuitions: Decompose based on Smooth Covariance Structures

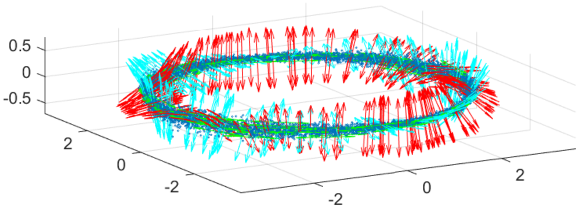


Intuitions: Decompose based on Smooth Covariance Structures

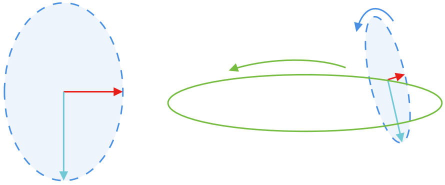


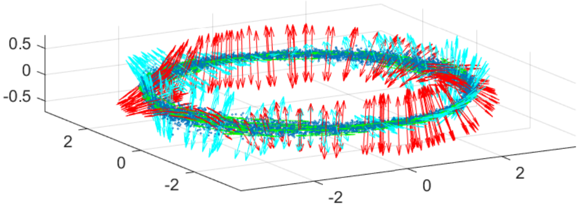
Intuitions: Decompose based on Smooth Covariance Structures





Intuitions: Decompose based on Smooth Covariance Structures

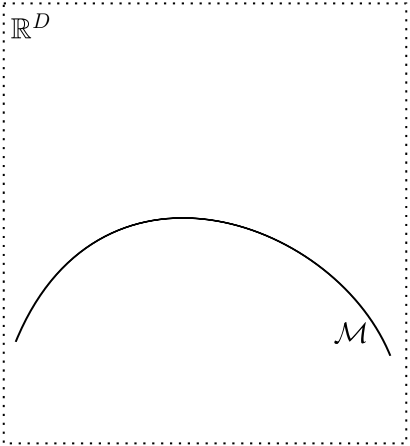




Fitting submanifolds such that

- 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

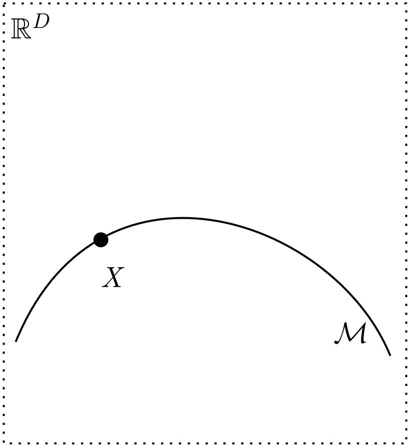
Manifold Fitting Setup*



- $\mathcal{M} \subset \mathbb{R}^D$: smooth latent manifold
- $\dim(\mathcal{M}) = d$, $d < D$, reach$(\mathcal{M}) \geq \tau$

*Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.
Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.
Yao, Z., Su, J., Li, B., & Yau, S.T. (2023). *Manifold fitting.* arXiv preprint arXiv:2304.07680. ...

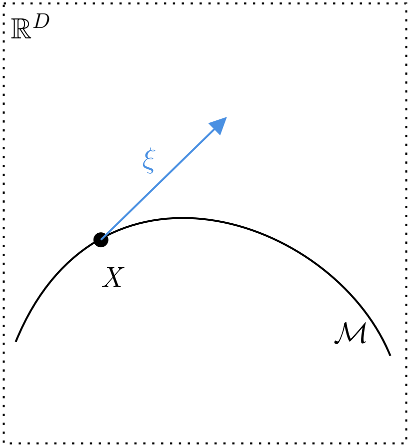
Manifold Fitting Setup*



- $\mathcal{M} \subset \mathbb{R}^D$: smooth latent manifold
- $\dim(\mathcal{M}) = d$, $d < D$, reach$(\mathcal{M}) \geq \tau$
- $X \in \mathcal{M} \subset \mathbb{R}^D$: unobserved sample from $\mu(\mathcal{M})$

*Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.
Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.
Yao, Z., Su, J., Li, B., & Yau, S.T. (2023). *Manifold fitting.* arXiv preprint arXiv:2304.07680. ...

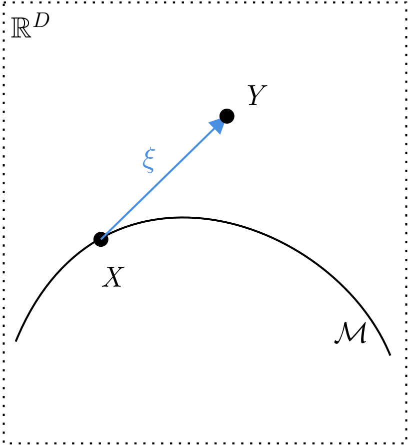
Manifold Fitting Setup*



- $\mathcal{M} \subset \mathbb{R}^D$: smooth latent manifold
- $\dim(\mathcal{M}) = d$, $d < D$, reach$(\mathcal{M}) \geq \tau$
- $X \in \mathcal{M} \subset \mathbb{R}^D$: unobserved sample from $\mu(\mathcal{M})$
- $\xi \in \mathbb{R}^D$, $\xi \sim \phi_\sigma^{(D)}$: ambient space noise

*Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.
Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.
Yao, Z., Su, J., Li, B., & Yau, S.T. (2023). *Manifold fitting.* arXiv preprint arXiv:2304.07680. ...

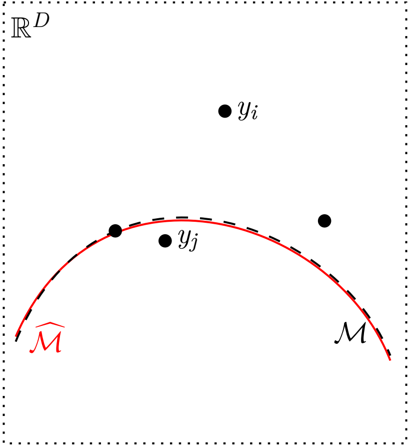
Manifold Fitting Setup*



- $\mathcal{M} \subset \mathbb{R}^D$: smooth latent manifold
- $\dim(\mathcal{M}) = d$, $d < D$, reach$(\mathcal{M}) \geq \tau$
- $X \in \mathcal{M} \subset \mathbb{R}^D$: unobserved sample from $\mu(\mathcal{M})$
- $\xi \in \mathbb{R}^D$, $\xi \sim \phi_\sigma^{(D)}$: ambient space noise
- $Y = X + \xi \sim \nu = \mu \star \phi_\sigma^{(D)}$: observation

*Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.
Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.
Yao, Z., Su, J., Li, B., & Yau, S.T. (2023). *Manifold fitting.* arXiv preprint arXiv:2304.07680. ...

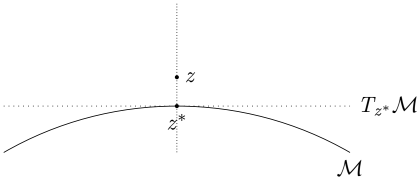
Manifold Fitting Setup*



- $\mathcal{M} \subset \mathbb{R}^D$: smooth latent manifold
- $\dim(\mathcal{M}) = d$, $d < D$, reach$(\mathcal{M}) \geq \tau$
- $X \in \mathcal{M} \subset \mathbb{R}^D$: unobserved sample from $\mu(\mathcal{M})$
- $\xi \in \mathbb{R}^D$, $\xi \sim \phi_\sigma^{(D)}$: ambient space noise
- $Y = X + \xi \sim \nu = \mu \star \phi_\sigma^{(D)}$: observation
- Estimate $\mathcal{M}$ with $\widehat{\mathcal{M}}$ based on $\{y_i\}_{i=1}^N$

*Fefferman, C., et al. (2018). *Fitting a putative manifold to noisy data.* PMLR.
Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.
Yao, Z., Su, J., Li, B., & Yau, S.T. (2023). *Manifold fitting.* arXiv preprint arXiv:2304.07680. ...

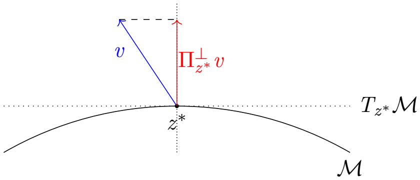
Fit the Latent Manifold[†]



- 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^*

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

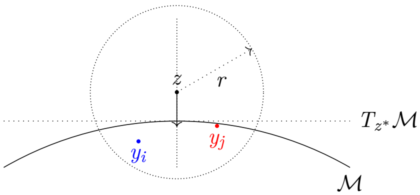
Fit the Latent Manifold[†]



- 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^*
- $\Pi_{z^*}^{\perp}$: projection matrix onto the normal space of $T_{z^*}\mathcal{M}$
- $\widehat{\Pi}_{z}^{\perp}$: estimator of $\Pi_{z^*}^{\perp}$

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

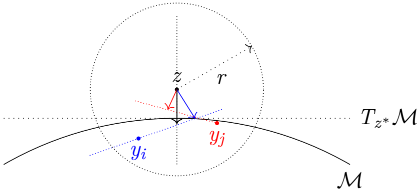
Fit the Latent Manifold[†]



- $r = \mathcal{O}(\sqrt{\sigma})$

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

Fit the Latent Manifold[†]



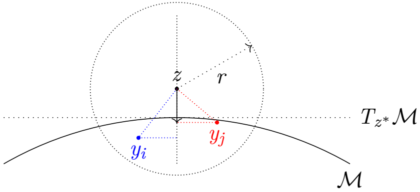
$$\tilde{\alpha}_i(z) = \left(1 - \frac{\|z - y_i\|^2}{r^2}\right)^\beta \mathbb{I}(\|z - y_i\| \le r)$$

$$\alpha_i(z) = \frac{\tilde{\alpha}_i(z)}{\sum_{i=1}^n \tilde{\alpha}_i(z)}, \quad \beta > 2,$$

- $r = \mathcal{O}(\sqrt{\sigma})$
- $\widehat{\Pi}^\perp_{y_i}$: estimator of $\Pi^\perp_{y_i^*}$
- $\Psi_z = \mathbb{P}_{D-d}\left(\sum_i \alpha_i(z)\widehat{\Pi}^\perp_{y_i}\right)$: estimator of $\Pi^\perp_{z^*}$
- $\mathbb{P}_d(A)$: projection of matrix A on its leading d eigenvectors

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

Fit the Latent Manifold[†]



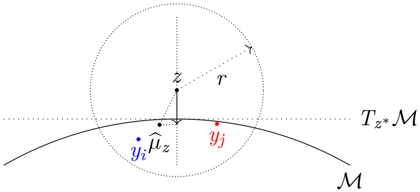
$$\tilde{\alpha}_i(z) = \left(1 - \frac{\|z - y_i\|^2}{r^2}\right)^{\beta} \mathbb{I}(\|z - y_i\| \le r)$$

$$\alpha_i(z) = \frac{\tilde{\alpha}_i(z)}{\sum_{i=1}^{n} \tilde{\alpha}_i(z)}, \quad \beta > 2,$$

- $\Psi_z = \mathbb{P}_{D-d}\left(\sum_i \alpha_i(z)\widehat{\Pi}^{\perp}_{y_i}\right)$
- $f(z) = \sum_i \alpha_i(z)\Psi_z(z - y_i)$

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

Fit the Latent Manifold[†]



$$\widehat{\mathcal{M}} = \{z \in \mathbb{R}^D : \mathsf{d}(z, \mathcal{M}) \le cr, c < 1, f(z) = 0\}$$

$$\Rightarrow \mathsf{d}(z, \mathcal{M}) \le Cr^2 \text{ for any } z \in \widehat{\mathcal{M}}$$

with probability

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

- $\Psi_z = \mathbb{P}_{D-d}\left(\sum_i \alpha_i(z)\widehat{\Pi}^{\perp}_{y_i}\right)$
- $\widehat{\mu}_z = \sum_i \alpha_i(z) y_i$: weighted mean of y_i
- $f(z) = \sum_i \alpha_i(z)\Psi_z(z - y_i) = \Psi_z(z - \widehat{\mu}_z)$

[†] Yao, Z., & Xia, Y. (2019 – 2025). *Manifold fitting under unbounded noise.* JMLR, to appear.

How Does This Work?

With properly designed weighting functions:

$$\tilde{\alpha}_i(z) = \left(1 - \frac{\|z - y_i\|^2}{r^2}\right)^{\beta} \mathbb{I}(\|z - y_i\| \leq r), \quad \alpha_i(z) = \frac{\tilde{\alpha}_i(z)}{\sum_{i=1}^{n} \tilde{\alpha}_i(z)}, \quad \beta > 2,$$

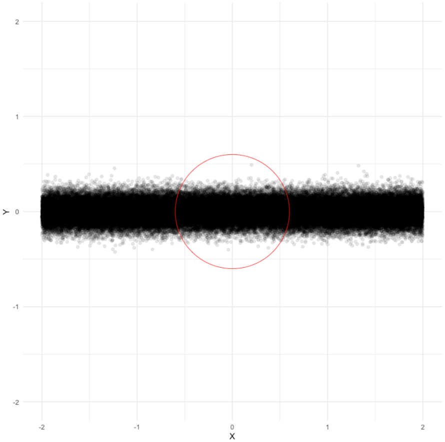
- Ψ_z is second order smooth, and estimates $\Pi_{z^*}^{\perp}$ well;
- the Jacobian matrix $J_f(x)$ satisfies $\|J_f(x) - \Psi_z\|_F \leq C\sigma/r$

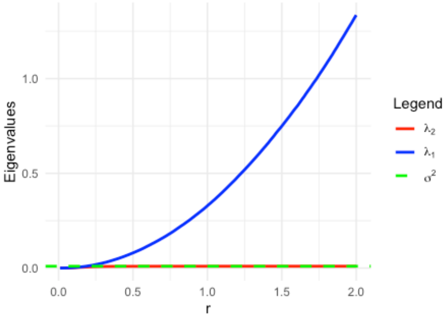
Together with the smoothness of $\mathcal{M}$, for $x \in \widehat{\mathcal{M}}$, in its neighbourhood

- construct $h(z) = V_x^{\top} f(z)$, which is rank $D - d$
- $f(z) = 0$ iff $h(z) = 0$

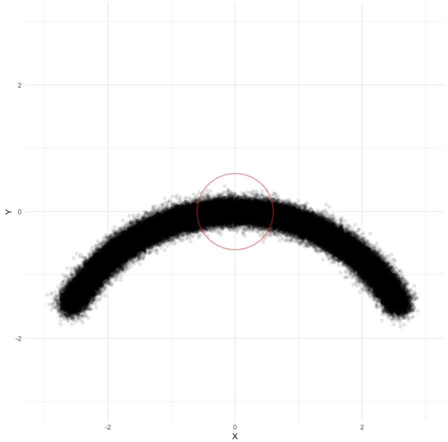
$\hookrightarrow \widehat{\mathcal{M}}$ is d-dimensional locally. Reach can be derived from Hessian of f.

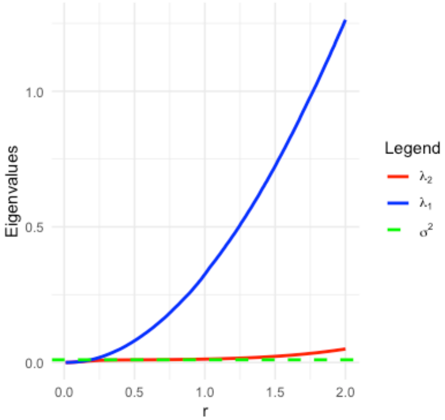
Why not Kernel Methods?



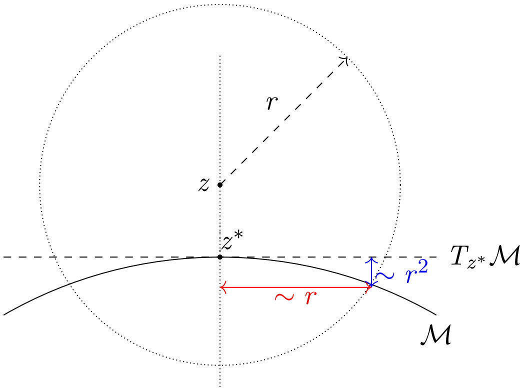


Why not Kernel Methods?





Why not Kernel Methods?



By choosing a proper $r \gg \sigma$:

- $P(Y \in \mathcal{B}_D(z,r)) \propto r^d/\mathrm{vol}(\mathcal{M})$
- Given $Y \in \mathcal{B}_D(z,r)$, the variance of Y:

  $\leftrightarrow \ \sim r^2 + \sigma^2$

  $\updownarrow \ \sim r^4 + \sigma^2$

  $\hookrightarrow$ allows estimating $\Pi^{\perp}_{z^*}$

Principal Nested Submanifolds – Population

- $X \in \mathbb{R}^D$: random vector of interest
- $\mathcal{X}$: support of X
- r: a radius parameter
- z: point of interest, with $\mathrm{d}(z, \mathcal{X}) < cr$
- Local average: $\mu_z = E\{X \mid X \in \mathcal{B}(z, r)\}$
- Local covariance matrix centered at z:

$$\Sigma_z = E\{(X - \mu_z)(X - \mu_z)^\top \mid X \in \mathcal{B}(z, r)\}$$

- SVD:
  - $\lambda_{z,1} < \cdots < \lambda_{z,D}$
  - $v_{z,1}, \ldots, v_{z,D}$, with $\|v_{z,j}\| = 1$

Principal Nested Submanifolds – Population

- $\Pi_{z,j} = v_{z,j} v_{z,j}^\top$: projection matrix onto the j-th principal direction at z

Consider the jth bias vector

$$f_j(z) = \Pi_{z,j}(\mu_z - z),$$

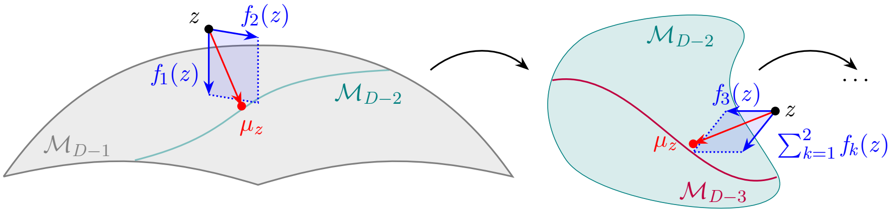
and let

$$\mathcal{M}_d = \{z \in \mathbb{R}^D : \; \mathsf{d}(z, \mathcal{X}) \leq cr, \; \sum_{j=1}^{D-d} f_j(z) = 0\}.$$

Then, if X satisfies concentration, identifiability, and smoothness assumptions, there is

- $\dim(\mathcal{M}_d) = d$
- $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_{D-1} \subset \mathbb{R}^D$

Intuition of Principal Nested Submanifolds



Empirical

- $\mathcal{X}_D = \{x_i\}_{i=1}^N \subset \mathbb{R}^D$
- Local covariance matrices:

$$\widehat{\Sigma}_i = \frac{\sum_j (x_i - x_j)(x_i - x_j)^\top \mathbb{I}(\|x_i - x_j\| < r)}{\sum_j \mathbb{I}(\|x_i - x_j\| < r)}$$

- $\widehat{\Pi}_{i,j}$: projection matrix onto the j-th smallest principal direction of $\widehat{\Sigma}_i$
- Weighting functions:

$$\tilde{\alpha}_i(z) = \left(1 - \frac{\|z - x_i\|^2}{r^2}\right)^\beta \mathbb{I}(\|z - x_i\| \leq r), \quad \alpha_i(z) = \frac{\tilde{\alpha}_i(z)}{\sum_{i=1}^n \tilde{\alpha}_i(z)}$$

with $\beta > 2$.

Empirical

Estimate

- $\widehat{\Pi}_{z,j} = \mathbb{P}_1\left(\sum_i \alpha_i(z)\widehat{\Pi}_{i,j}\right)$
- $\widehat{\mu}_z = \sum_i \alpha_i(z)x_i$
- $\widehat{f}_j(z) = \widehat{\Pi}_{z,j}\left(\widehat{\mu}_z - z\right)$

and define the estimator for the nested d-dimensional submanifold as

$$\widehat{\mathcal{M}}_d = \{z \in \mathbb{R}^D : \ \mathsf{d}(z, \mathcal{X}_D) \le cr, \ \sum_{j=1}^{D-d} \widehat{f}_j(z) = 0\}.$$

Then, $\dim(\widehat{\mathcal{M}}_d) = d$, $\widehat{\mathcal{M}}_1 \subset \cdots \subset \widehat{\mathcal{M}}_{D-1} \subset \mathbb{R}^D$, and $\widehat{\mathcal{M}}_d$ is close to $\mathcal{M}_d$ with high probability.

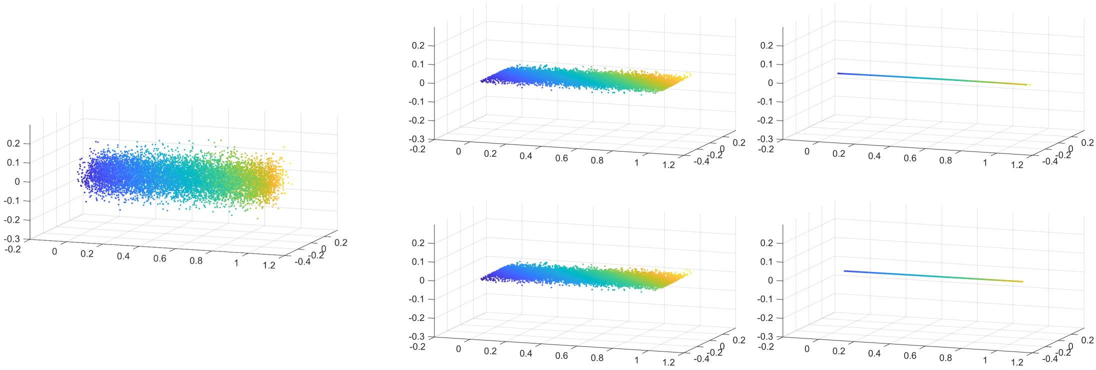
Estimation & Projection

- Embed data on $\mathcal{M}$ in $\mathbb{R}^D$ if needed
- Start with $\mathcal{X}_D$ and calculate all $\widehat{\Pi}_{i,j}$'s
- for $d = \dim(\mathcal{M}) - 1 : 1$
  - for each element in $\mathcal{X}_{d+1}$, update with rule

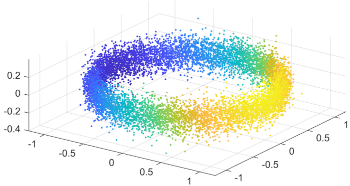
$$z' = z + \sum_{j=1}^{D-d} \widehat{f}_j(z)$$

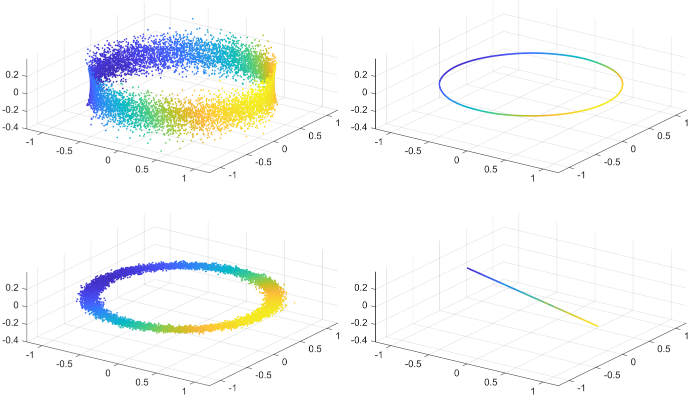
  - stop when $\| \sum_{k=1}^{j} \widehat{f}_j(z) \| < \epsilon$
  - set the converge points as $\mathcal{X}_d$
  - project $\mathcal{X}_d$ on $\mathcal{M}$ if needed.
- Return $\mathcal{X}_d$ for $d = 1, \ldots, \dim(\mathcal{M}) - 1$.

Simulation with Euclidean Space I

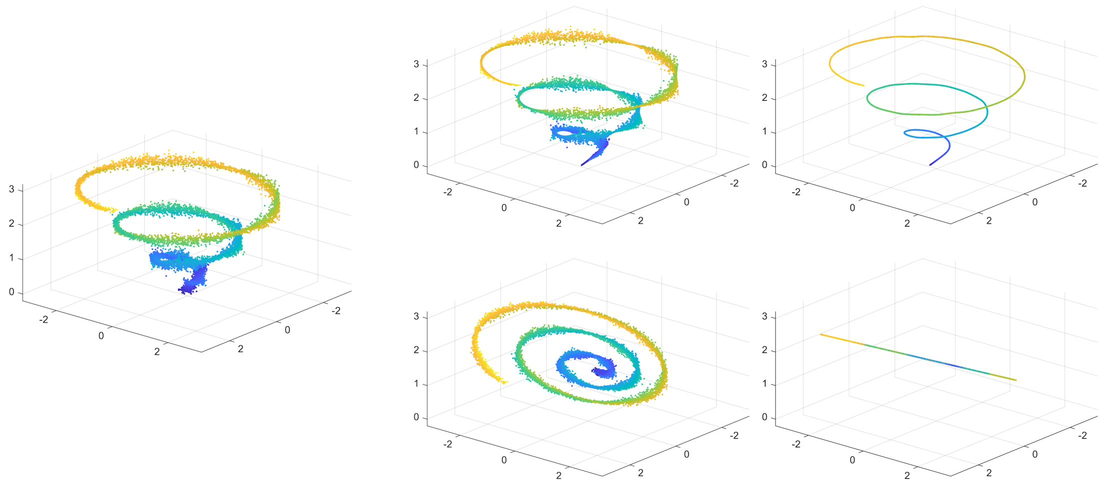


Simulation with Euclidean Space II

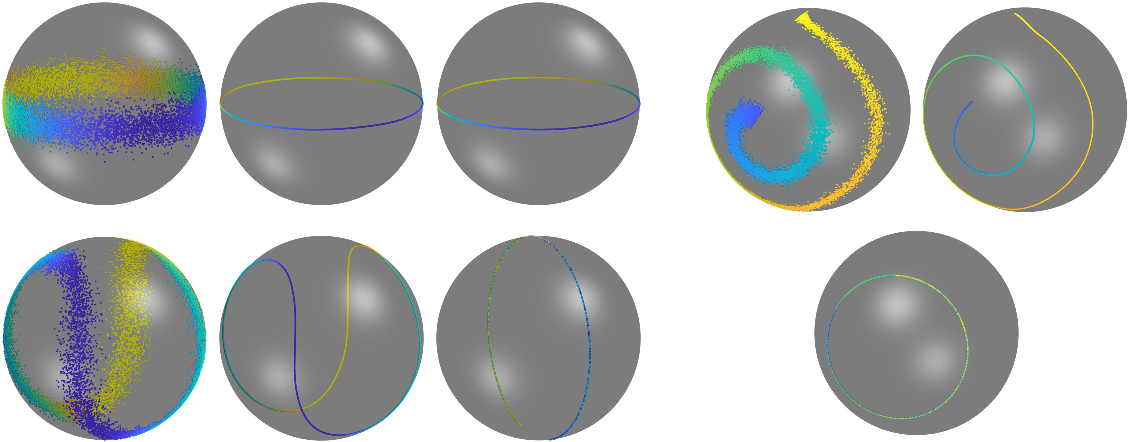




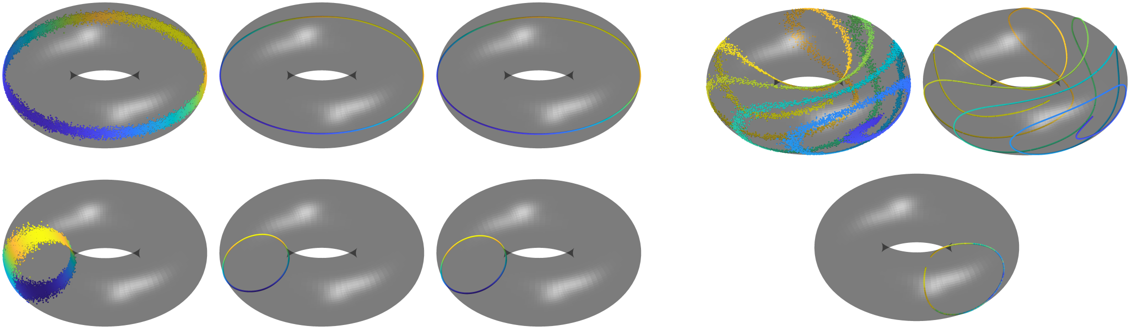
Simulation with Euclidean Space III



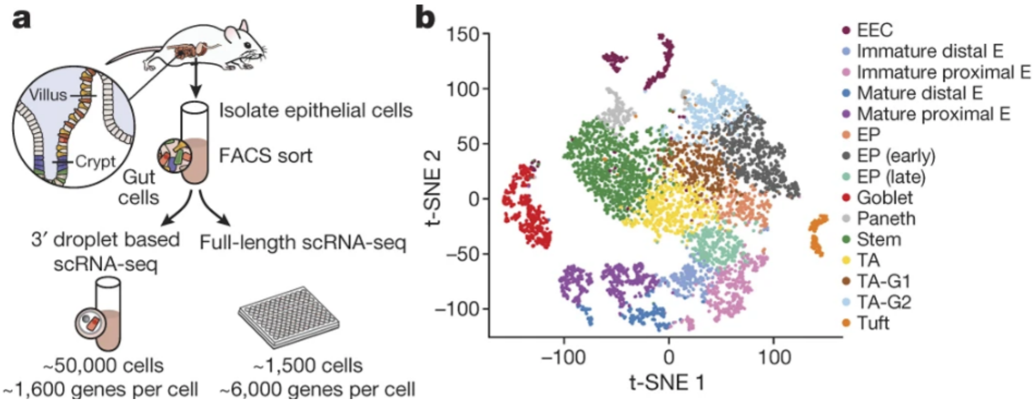
Simulation with Spheres



Simulation with Tori



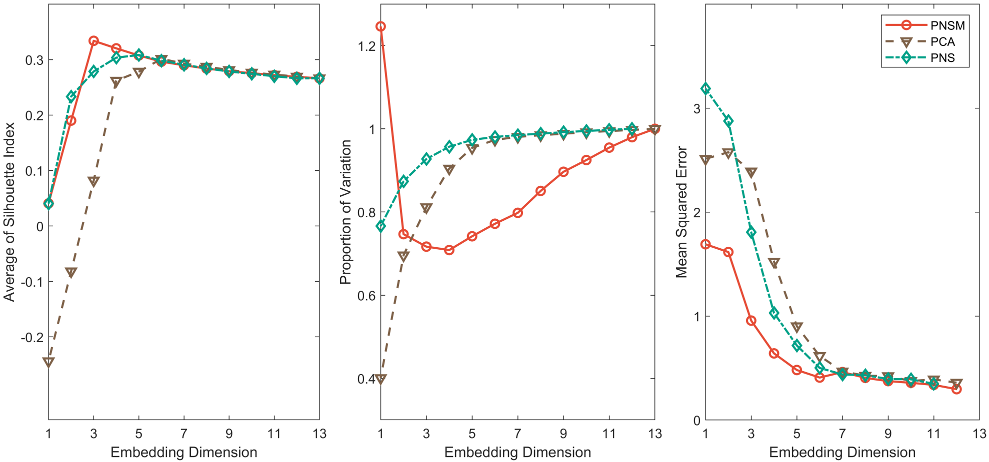
scRNA Data Set[‡]



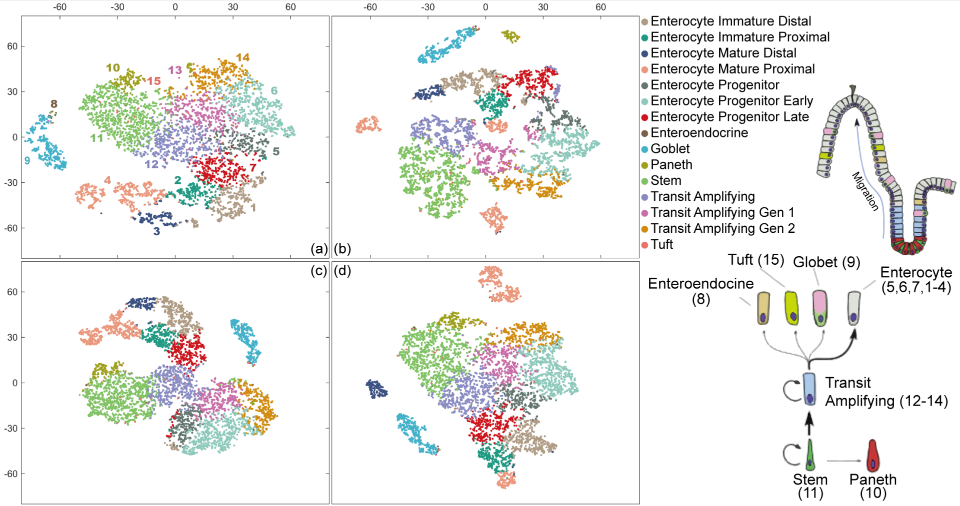
Totally 7,216 small intestinal epithelium cells from 6 mice, with the first 13 PCs.

[‡] Haber, Adam L., et al. (2017). *A single-cell survey of the small intestinal epithelium.* Nature.

Compare

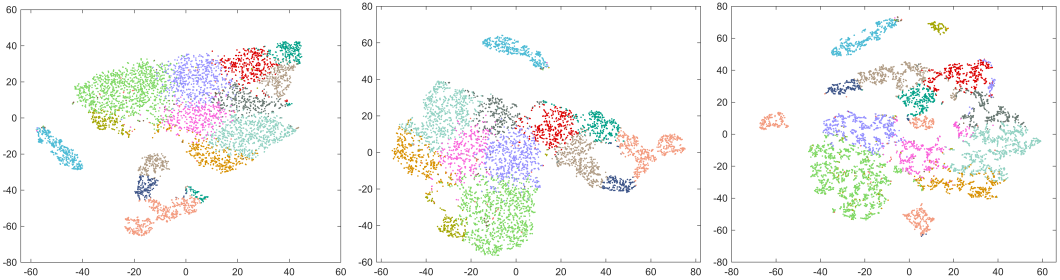

Average of Silhouette Index
Embedding Dimension
Proportion of Variation
Embedding Dimension
Mean Squared Error
Embedding Dimension
PNSM
PCA
PNS

tSNE Visualization§



§(a)Orignial Data; (b) PNSM with $d = 3$; (c) PCA with $d = 6$; (d) PNS with $d = 5$

tSNE Visualizations: $d \in \{13, 8, 3\}$



Conclusion & Outlook

- **Principal Nested Submanifolds**: a nonlinear, data-driven generalization of PCA.
  - Recovers nested low-dimensional structures from data
  - Applies to general spaces
  - Achieves superior performance in simulation and real single-cell data
- Future directions:
  - Scalable versions for large-scale omics data.
  - Integration into neural networks like generative models.

Thank you!
Questions are welcome.

Slides and preprint:
sujiaji.cn/nav