Machine Learning Methods Based on Diffusion Processes

Final Defense Presentation

Chenchao Zhao

University of Illinois at Urbana-Champaign

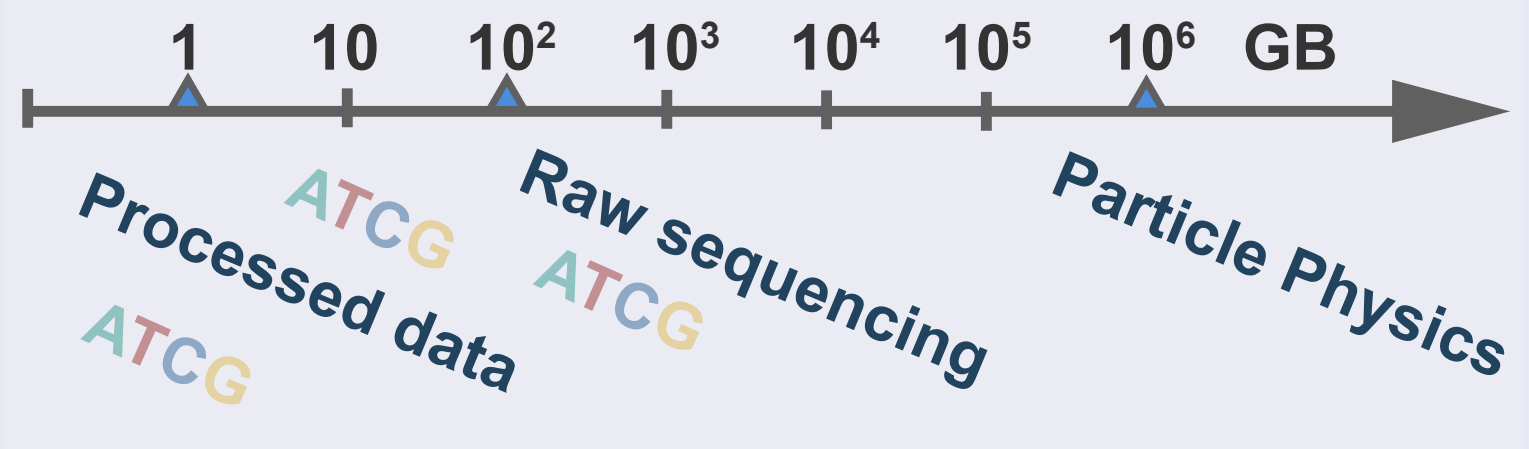
April 2, 2018

Outline

- Introduction
- Results
  1. Hyperspherical heat kernel and applications
     - heat diffusion on hypersphere
  2. Effective Dissimilarity Transformation (EDT)
     - discrete-time data drifting
  3. Quantum Transport Clustering (QTC)
     - quantum diffusion in networks
- Conclusion

Big data fuels the development of data science

How *big* are big data?



- Hard drive price (USD per GB) in 1981 was **$700,000**, while for today it is about **$0.03**.
- Data are thus collected in large quantities, which creates demand for automated data analysis techniques.

Machine learning is automated and data-driven

Origin of machine learning (Samuel 1959)

"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."

- The concept of "machine learning" (ML) predates big data era.
- ML is about designing programs that improve the performance via **data** instead of **explicit programming**.
- ML includes pattern or community detection, classification and regression, and etc.

Challenges in modern data analysis

- The n features are often stacked to form $\mathbb{R}\times\mathbb{R}\times\cdots\times\mathbb{R}=\mathbb{R}^n$, and m samples are usually organized as m vectors in $\mathbb{R}^n$.
- The **intrinsic geometry** of data generation model is often not $\mathbb{R}^n$.
  - Generalization of similarity or dissimilarity measures to curved spaces, e.g. hypersphere
- "**Curse of dimensionality**" (Bellman 1957) – number of feature n is often large.
  - Loss of contrast of Euclidean distance in high dimension
- Data distributions of **complex geometric shapes** are difficult to separate based on distance measures.
  - Network community detection problem

Chapter 1 Hyperspherical heat kernel

Question

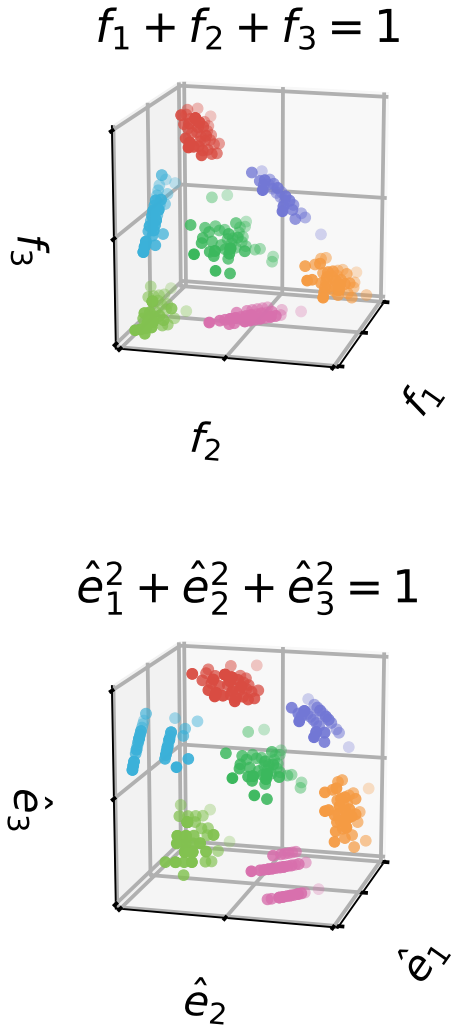
How should we measure similarity and dissimilarity on a high-dimensional manifold $\mathcal{M}$?

Multinomial distribution and hyperspherical geometry

- In natural language processing, we often assume the word counts

$$\{x_i\}_{i=1}^n \sim \text{Multinomial}(N, \{p_i\}_{i=1}^n)$$

  where $\sum_{i=1}^n p_i = 1$, and $\sum_{i=1}^n x_i = N$.
- The word frequency $f_i = \frac{x_i}{N}$ is ℓ_1-normalized $\sum_{i=1}^n f_i = 1$.
- Let $\hat{e}_i \equiv \sqrt{f_i}$, then $\sum_{i=1}^n \hat{e}_i^2 = \|\hat{\mathbf{e}}\|_2^2 = 1$ and $\hat{\mathbf{e}} \in S^{n-1}$.
- Thus, a **document** is represented as a point on a hypersphere (Lafferty & Lebanon 2005).



Euclidean heat kernel is a Gaussian function

- Gaussian function is widely applied in statistics and machine learning.
- What is the counter part on a manifold, e.g. S^{n-1}?
- The Euclidean heat kernel

$$G(\mathbf{x},\mathbf{y};t)=\left(\frac{1}{4\pi t}\right)^{\frac{n}{2}}\mathrm{e}^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{4t}}$$

  is a Gaussian radial basis function $K^{\mathrm{rbf}}(\mathbf{x},\mathbf{y};\gamma)=\mathrm{e}^{-\gamma|\mathbf{x}-\mathbf{y}|^2}$.
- $(\partial_t-\Delta_x)G(\mathbf{x},\mathbf{y};t)=0$ with initial data $G(\mathbf{x},\mathbf{y};0^+)=\delta(\mathbf{x}-\mathbf{y})$.
- Thus, we can generate a Gaussian-like kernel function on a manifold using heat diffused from a point source.

Parametrix heat kernel on S^{n-1} (Lafferty & Lebanon 2005)

- In short time limit $t \downarrow 0$, the heat kernel should resemble the Euclidean kernel

$$G^{\text{prx}}(\theta; t) = \left(\frac{1}{4\pi t}\right)^{\frac{n-1}{2}} \mathrm{e}^{-\frac{\theta^2}{4t}} \left[u_0(\theta) + u_1(\theta)t + u_2(\theta)t^2 + \cdots\right]$$

- All u_p are singular at $\theta = \pi$, u_2 is even singular at $\theta = 0$.
- With corrections from u_0 and u_1, $G^{\text{prx}}(\theta; t)$ increases with θ when n is large.
- Heuristically, $K^{\text{prx}}(\hat{x}, \hat{y}; t) = \mathrm{e}^{-\arccos^2 \hat{x}\cdot\hat{y}/4t}$ (Lafferty & Lebanon 2005).
- The "kernel" is singular at $\hat{x} \cdot \hat{y} = -1$ and not positive-definite for large t.

Exact heat kernel on S^{n-1}

- The Euclidean Laplacian can be interpreted as kinetic energy operator $\hat{\mathbf{P}}^2$ which generates diffusion in $\mathbb{R}^n$: $G(\mathbf{x}, \mathbf{y}; t) = \mathrm{e}^{-t\hat{\mathbf{P}}^2}\delta(\mathbf{x}-\mathbf{y})$
- The momentum $\hat{\mathbf{P}}$ can be decomposed into radial and angular parts.
- S^{n-1} is embedded in $\mathbb{R}^n$, then diffusion in S^{n-1} is generated by **angular momentum** operator $\hat{L}^2$:

$$G^{\mathrm{ext}}(\hat{x}, \hat{y}; t) = \mathrm{e}^{-t\hat{L}^2}\delta(1-\hat{x}\cdot\hat{y}).$$

- Spherical harmonics $Y_{\ell,\mathbf{m}}(\hat{x})$ satisfy $\hat{L}^2 Y_{\ell,\mathbf{m}} = \ell(\ell+n-2)Y_{\ell,\mathbf{m}}$.

Exact hyperspherical heat kernel

Eigenfunction expansion on S^{n-1}

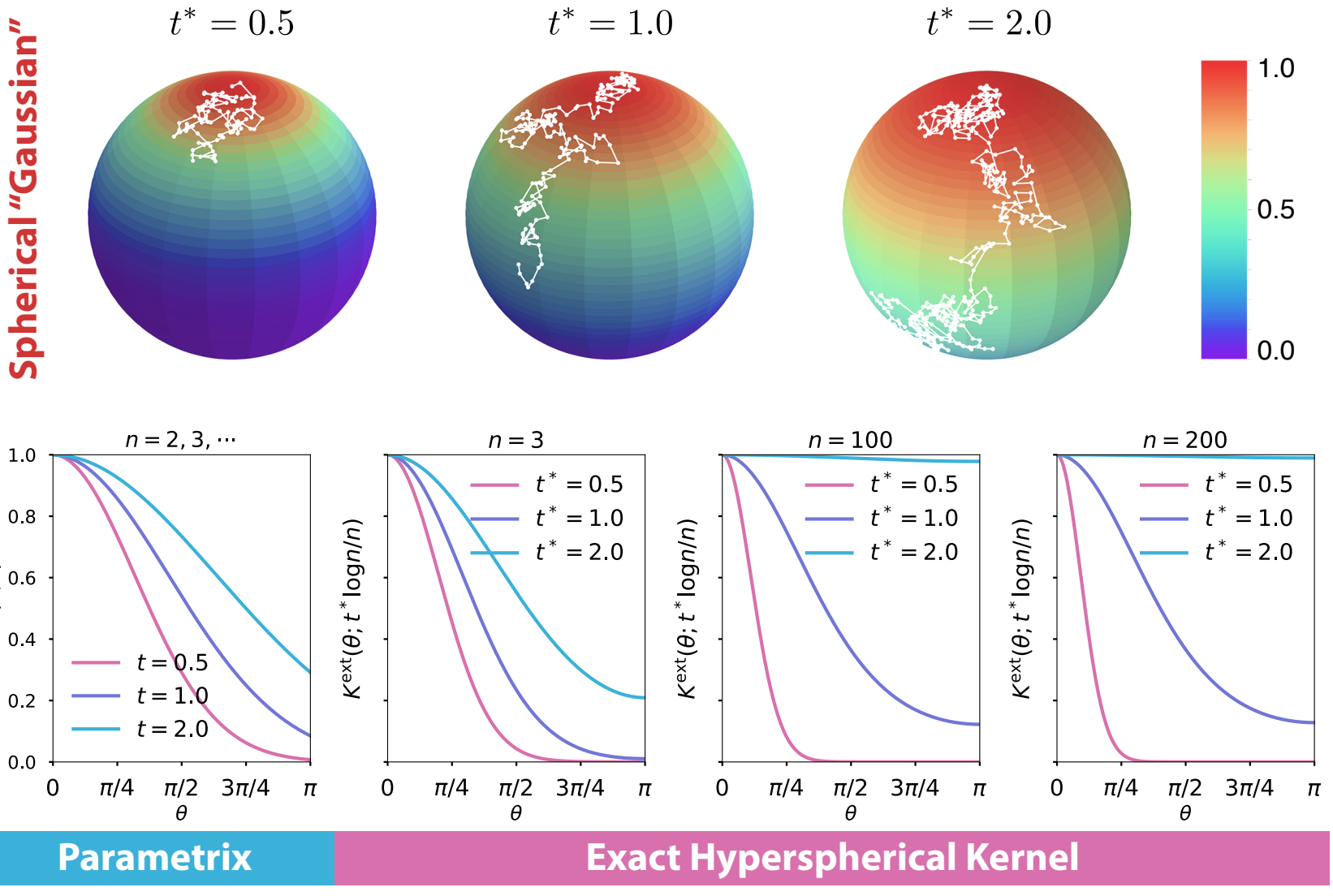
The exact hyperspherical heat kernel is an absolutely and uniformly convergent series

$$G^{\text{ext}}(\hat{x},\hat{y};t) = \sum_{\ell=0}^{\infty} \mathrm{e}^{-\ell(\ell+n-2)t} \frac{2\ell+n-2}{n-2} \frac{1}{A_{S^{n-1}}} C_{\ell}^{\frac{n}{2}-1}(\hat{x}\cdot\hat{y})$$

where $C_{\ell}^{\alpha}(w)$ with $w \in [-1,1]$ are the *Gegenbauer* polynomials and $A_{S^{n-1}} = 2\pi^{\frac{n}{2}}/\Gamma\left(\frac{n}{2}\right)$.

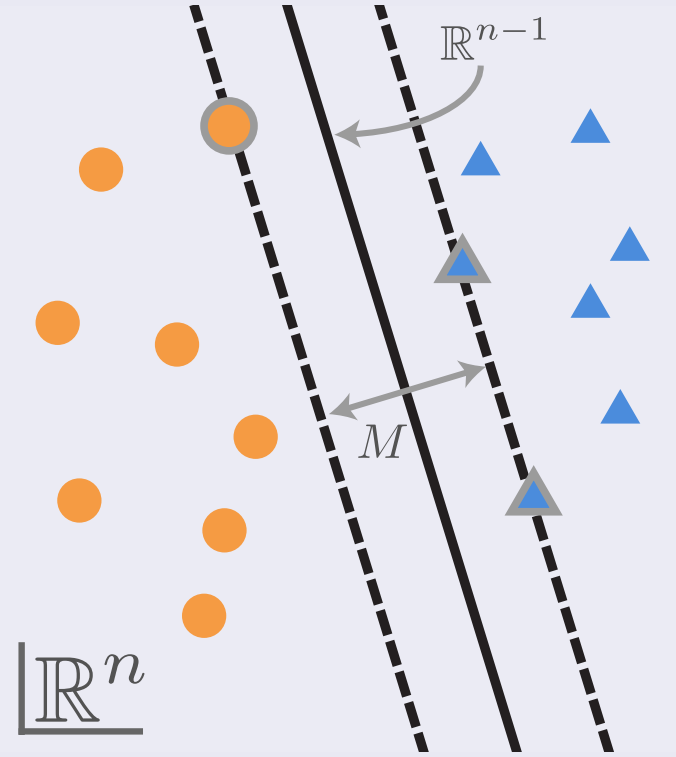
- Exact kernel $K^{\text{ext}} = G^{\text{ext}}(\hat{x}\cdot\hat{y};t)/G^{\text{ext}}(1;t)$.
- When $n \gg 1$, rescale time $t = t^* \log n/n$.

K^{prx} is an approximation to K^{ext} at $t \downarrow 0$

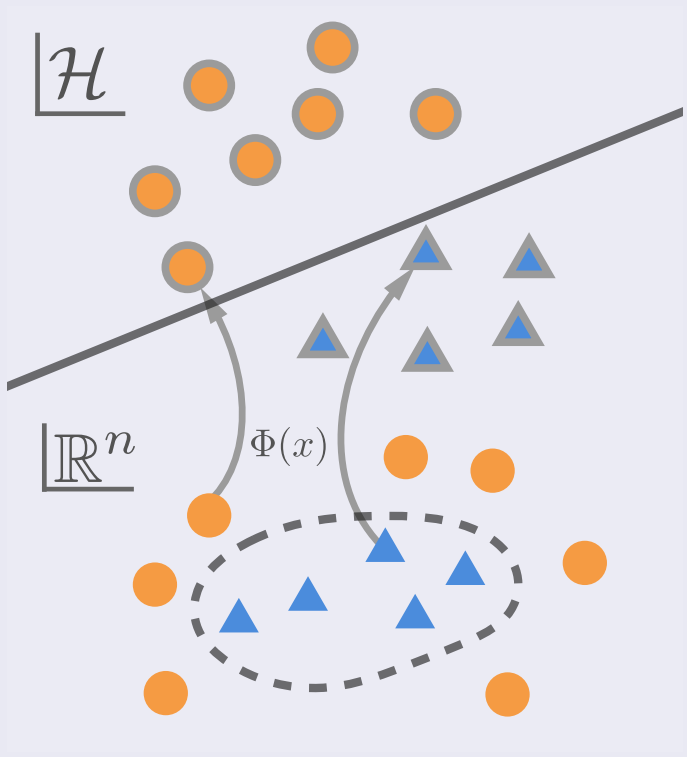


Kernel Support Vector Machine (SVM)

Optimal hyperplane



Feature map $\Phi : \mathbb{R}^n \to \mathcal{H}$



Kernel SVM classification of count data sets

Table: WebKB-4-University Document Classification.

m_{r}	lin	rbf	cos	prx	ext
100	74.2%	75.1%	84.4%	85.4%	**85.6%**
200	80.9%	82.0%	89.2%	89.6%	**89.9%**
300	83.2%	84.1%	89.9%	90.5%	**91.1%**
400	86.7%	86.1%	91.3%	91.7%	**92.3%**

- $K^{\mathrm{lin}} = \mathbf{x} \cdot \mathbf{y}$, $K^{\cos} = \hat{x} \cdot \hat{y}$, $K^{\mathrm{rbf}} = \mathrm{e}^{-\gamma |\mathbf{x}-\mathbf{y}|^2}$, $K^{\mathrm{prx}} = \mathrm{e}^{-\arccos^2 \hat{x} \cdot \hat{y}/4t}$, and K^{ext}.
- In document classifications based on word counts,
  - K^{ext} reduced the error of K^{rbf} by 41% $\sim$ 45%
  - K^{ext} reduced the error of K^{prx} by 1% $\sim$ 7%.

Kernel SVM classification of time series

- Hyperspherical projection allows the data points to explore the whole S^{n-1}.
- The log-returns of one stock in n days form a time-series and lie in $\mathbb{R}^n \backslash \{0\}$.
- S&P500 stocks: 91 *Financial* v. 64 *Information Technology*
- If $n > m$, the K^{ext} outperformed all other kernels
  - reduced the error of K^{rbf} by $29 \sim 51\%$,
  - reduced the error of K^{prx} by $17 \sim 51\%$.
- Thus, K^{prx} and K^{ext} perform *differently* if the data points explore the *whole* hypersphere.

Spherical geometry improved classification accuracy

- With assumption of the data generating distributions, the feature space $\mathbb{R}^n$ can be transformed into a manifold $\mathcal{M}$.
  - e.g., frequencies of n distinct words in a document, and n instances of a time series can be mapped to S^{n-1}.
  - S^{n-1} geometry improves classification accuracy in document and time series data.
- The distances between samples are then measured with geodesic distance of $\mathcal{M}$.

Chapter 2 Effective Dissimilarity Transformation

Question

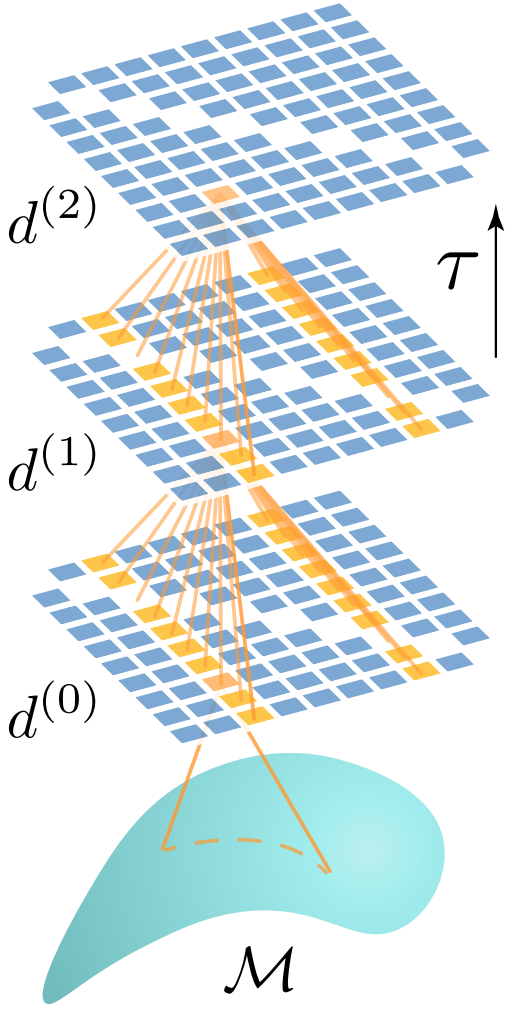
Without knowledge of the feature space $\mathcal{M}$, is it possible to reveal the hidden structures encrypted in pairwise distances d_{ij}?

The effective dissimilarity transformation

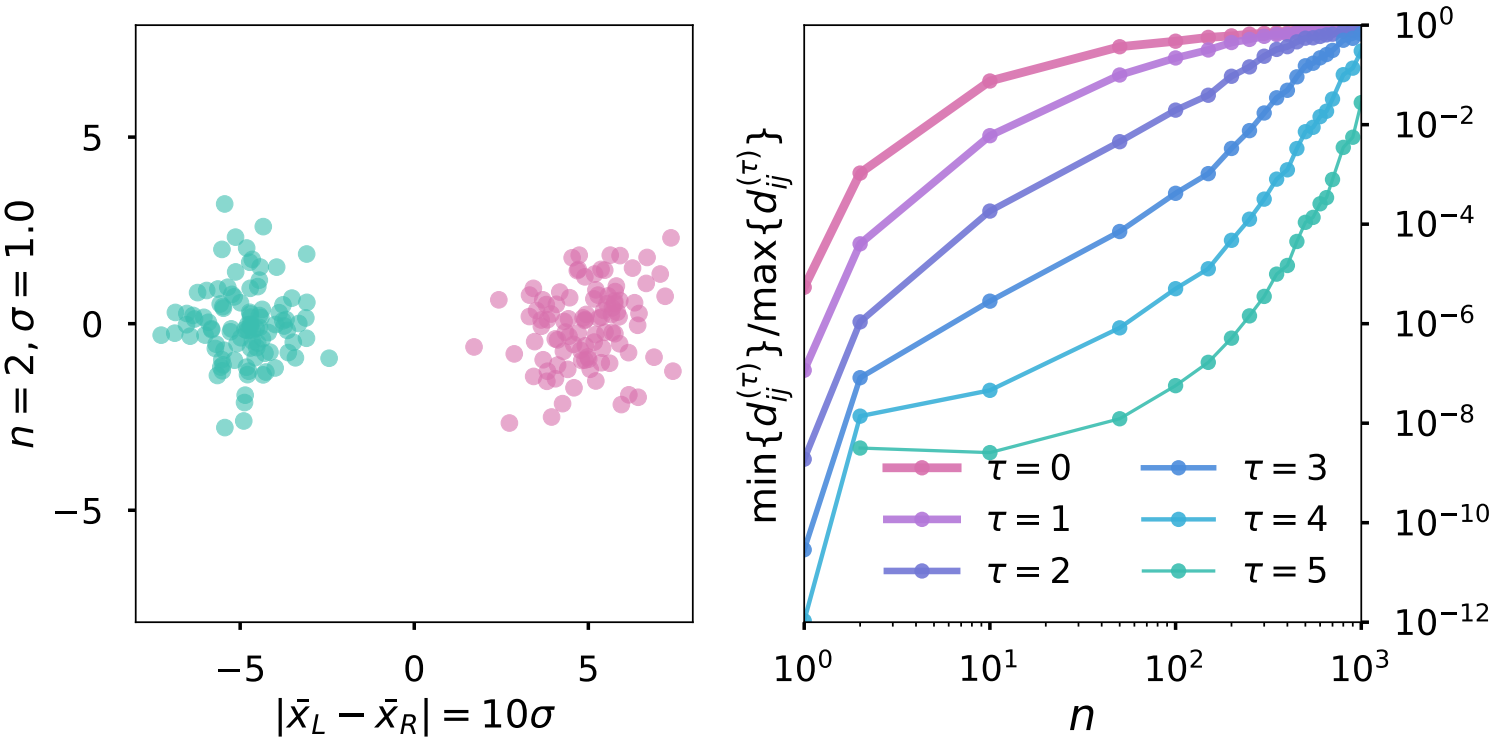
EDT: $d^{(\tau)} \mapsto d^{(\tau+1)}$

1. Input: d_{ij}, and let $\mathbf{d}_j = d_{\cdot j}$
2. $\tilde{\mathbf{d}}_j = \mathbf{d}_j / \|\mathbf{d}_j\|_1 = \mathbf{d}_j / \sum_{i=1}^{m} d_{ij}$
3. $(\hat{\mathbf{p}}_j)_i \equiv \sqrt{(\tilde{\mathbf{d}}_j)_i}$ for $i = 1, 2, \cdots, m$
4. Output: $d_{ij}^{(\text{new})} = 1 - \hat{\mathbf{p}}_i \cdot \hat{\mathbf{p}}_j$

- EDT can be iterated, i.e. $d^{(\tau)} \mapsto d^{(\tau+1)}$ for $\tau = 0, 1, 2 \cdots$
- The original $d^{(0)}$ can be any dissimilarity measure.

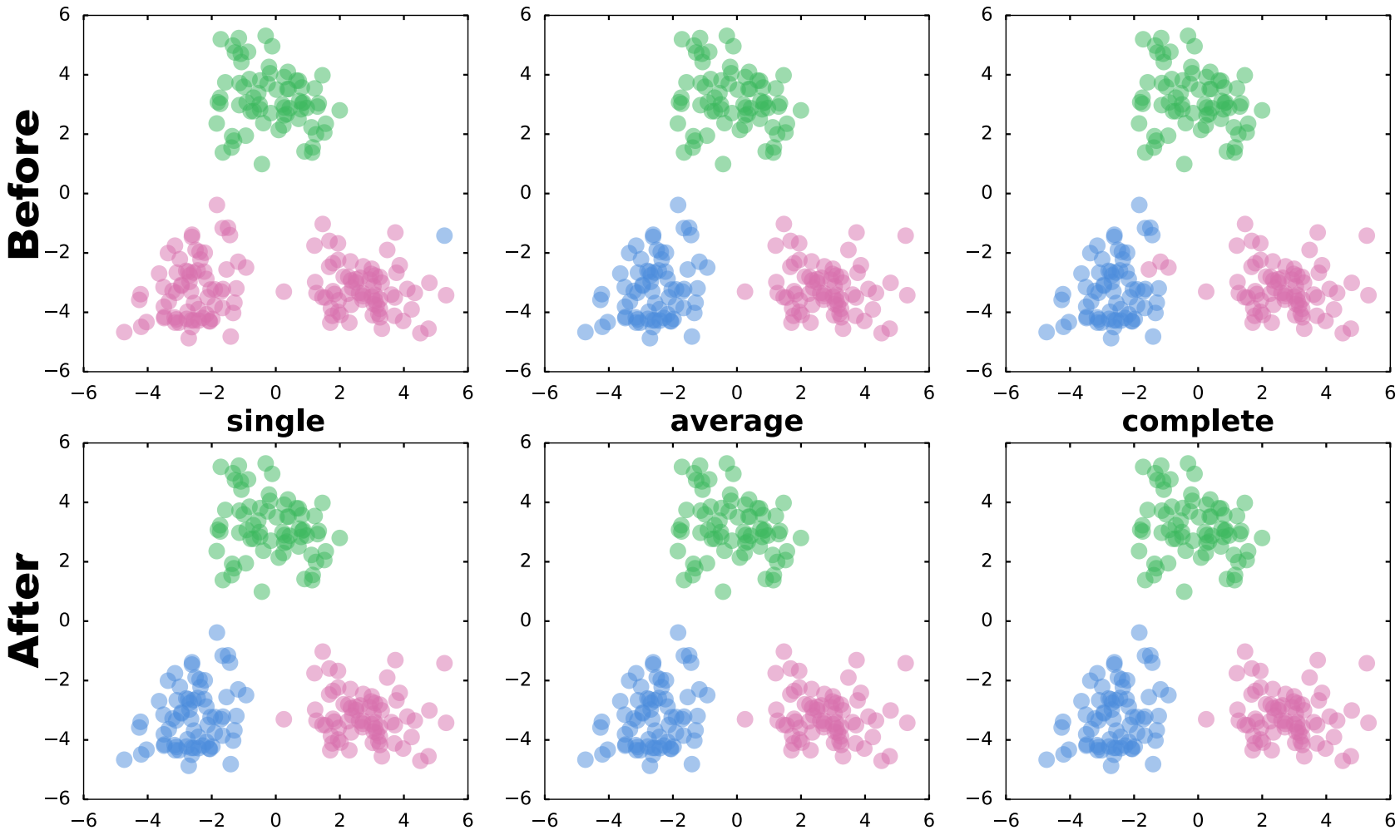


Curse of dimensionality for $\mathbb{R}^n$

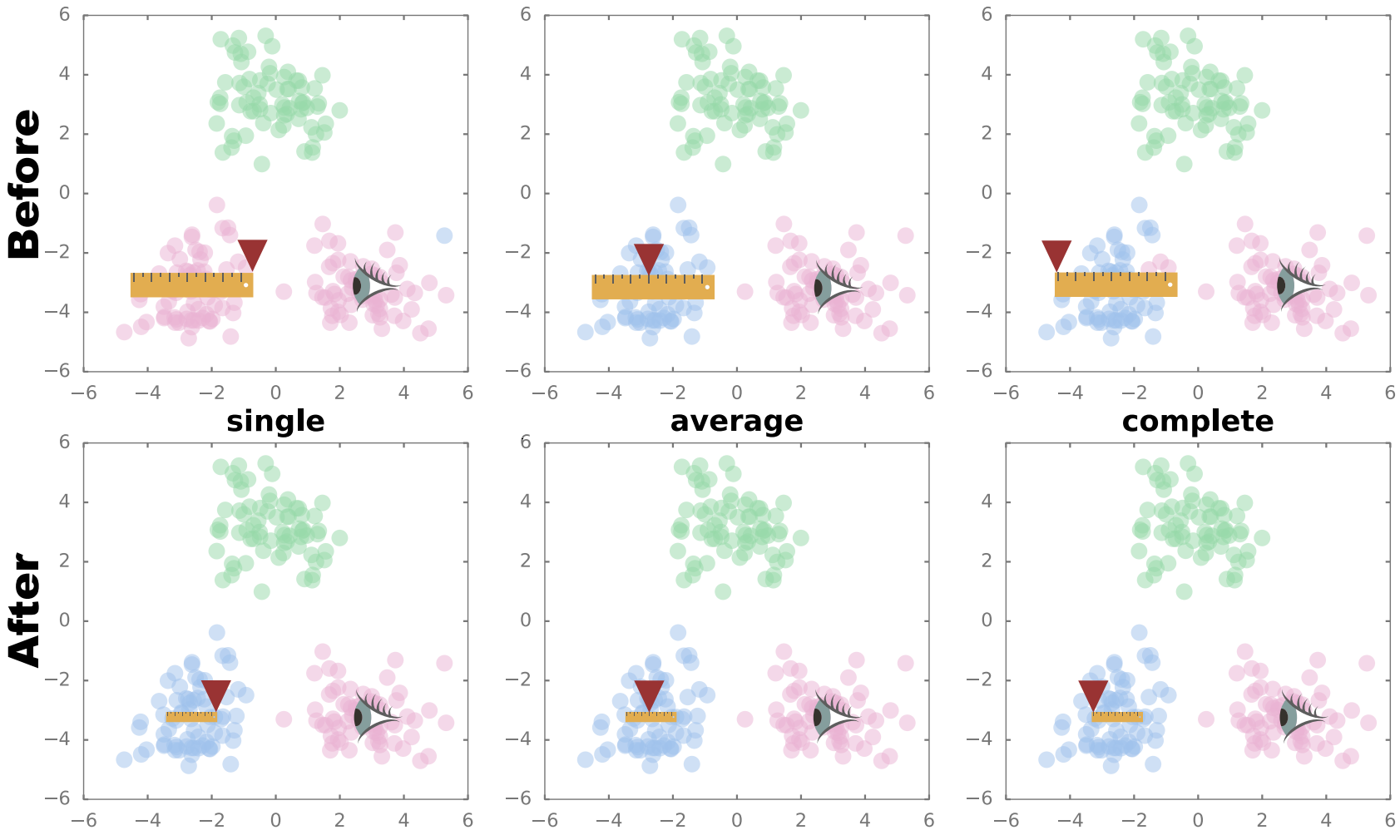


- $\tau = 0$ corresponds to Euclidean distance, where $\min\{d_{ij}^{(0)}\} \approx \max\{d_{ij}^{(0)}\}$ as $n \uparrow \infty$.

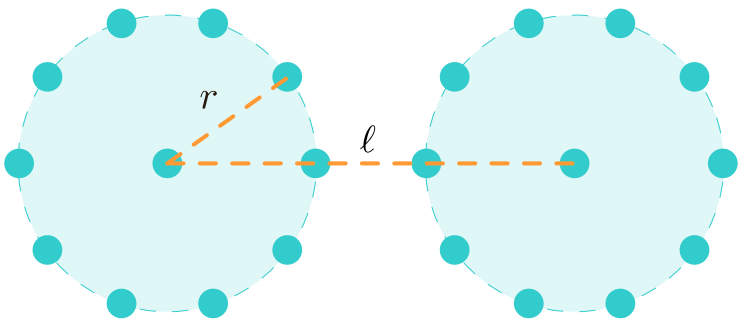
Euclidean metric hierarchical clustering varied with linkages



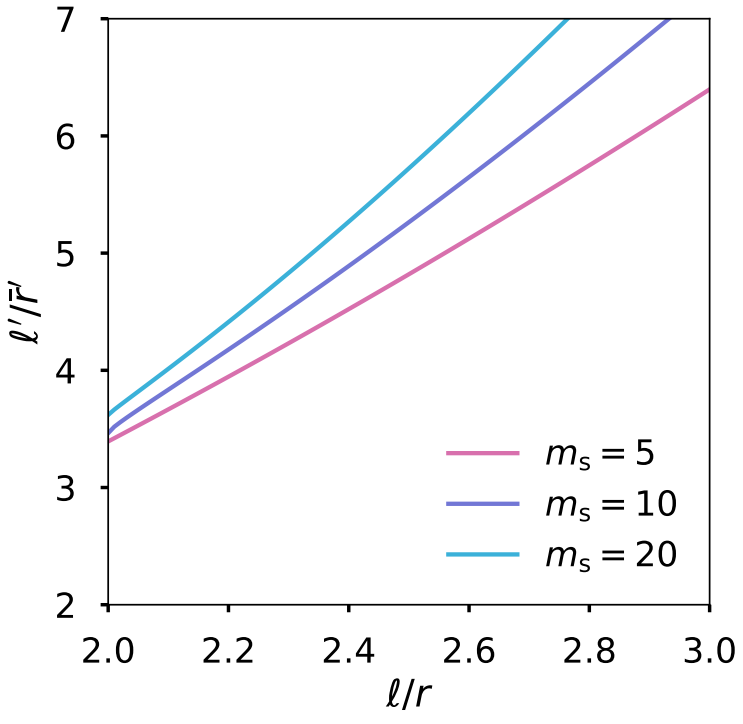
Perspective contraction induced by EDT



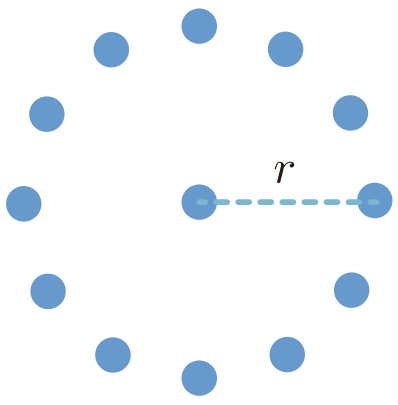
Cluster condensation – sample points gravitate under EDT



- $\ell'/\bar{r}'$ increased with cluster size m_s and original ℓ/r.



Local deformation measured with a probe



- After EDT,

$$\{r_i = r\}_{i=1}^{m_s} \mapsto \{r_i'\}_{i=1}^{m_s}$$

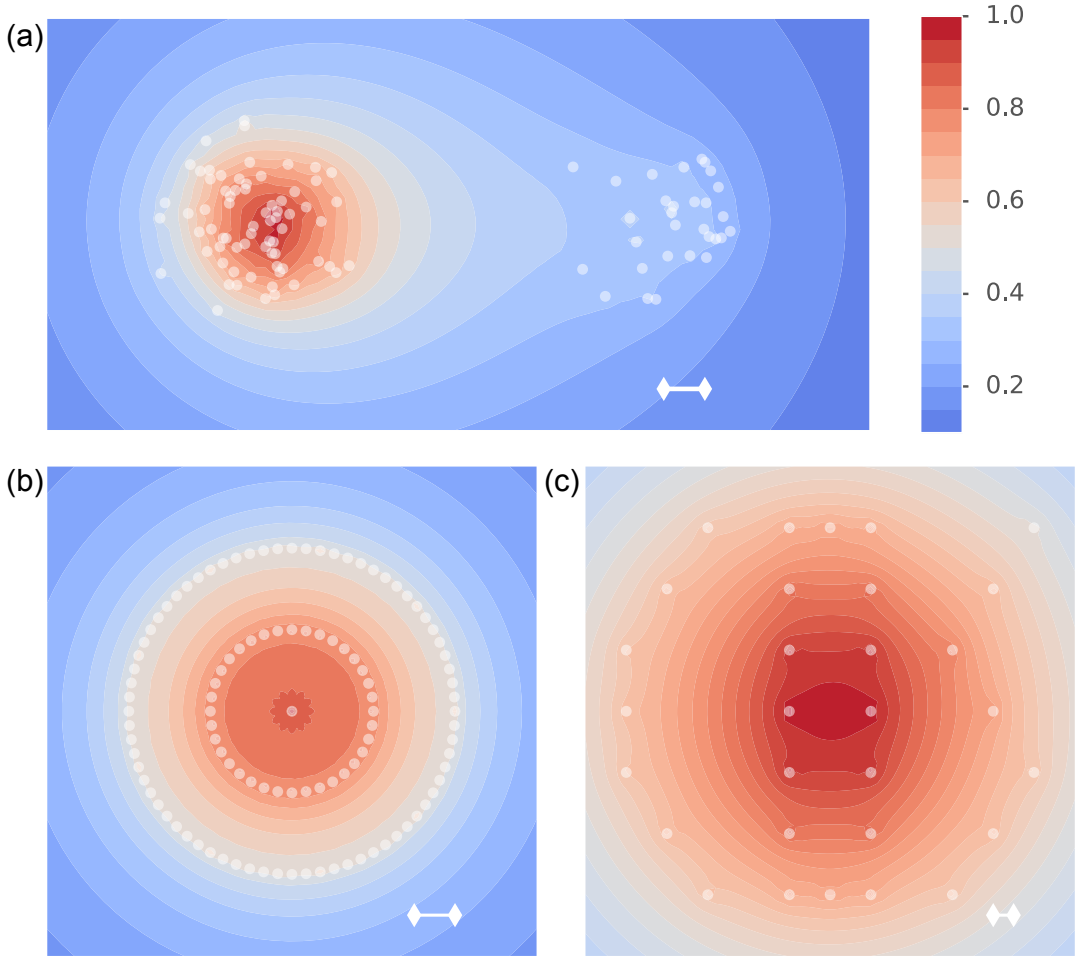
- Local volume effect

$$\nu = \bar{r}' = \frac{1}{m_s} \sum_{i=1}^{m_s} r_i'$$

- Local anisotropy

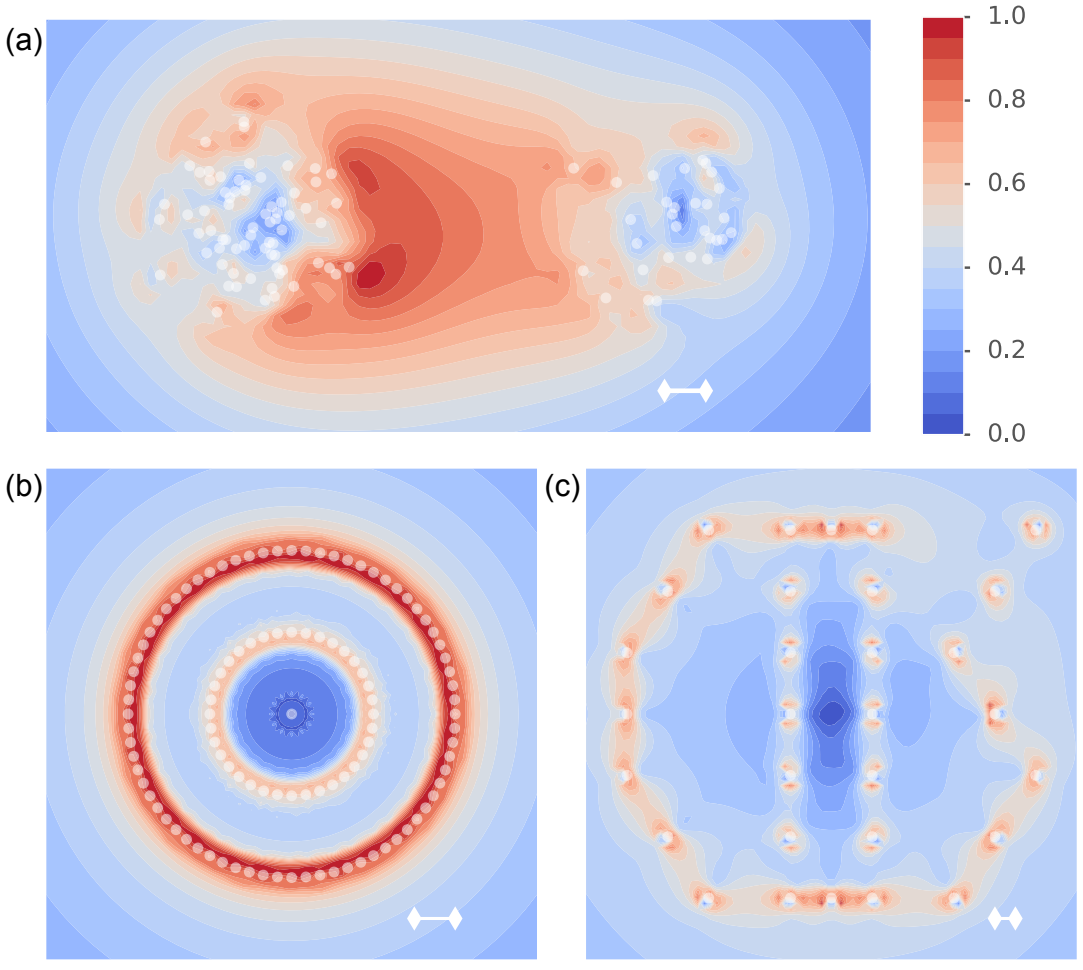
$$\kappa = \mathrm{std} \left\{ \frac{r_i'}{\bar{r}'} \right\}_{i=1}^{m_s}$$

Local volume effect: landscape of ν



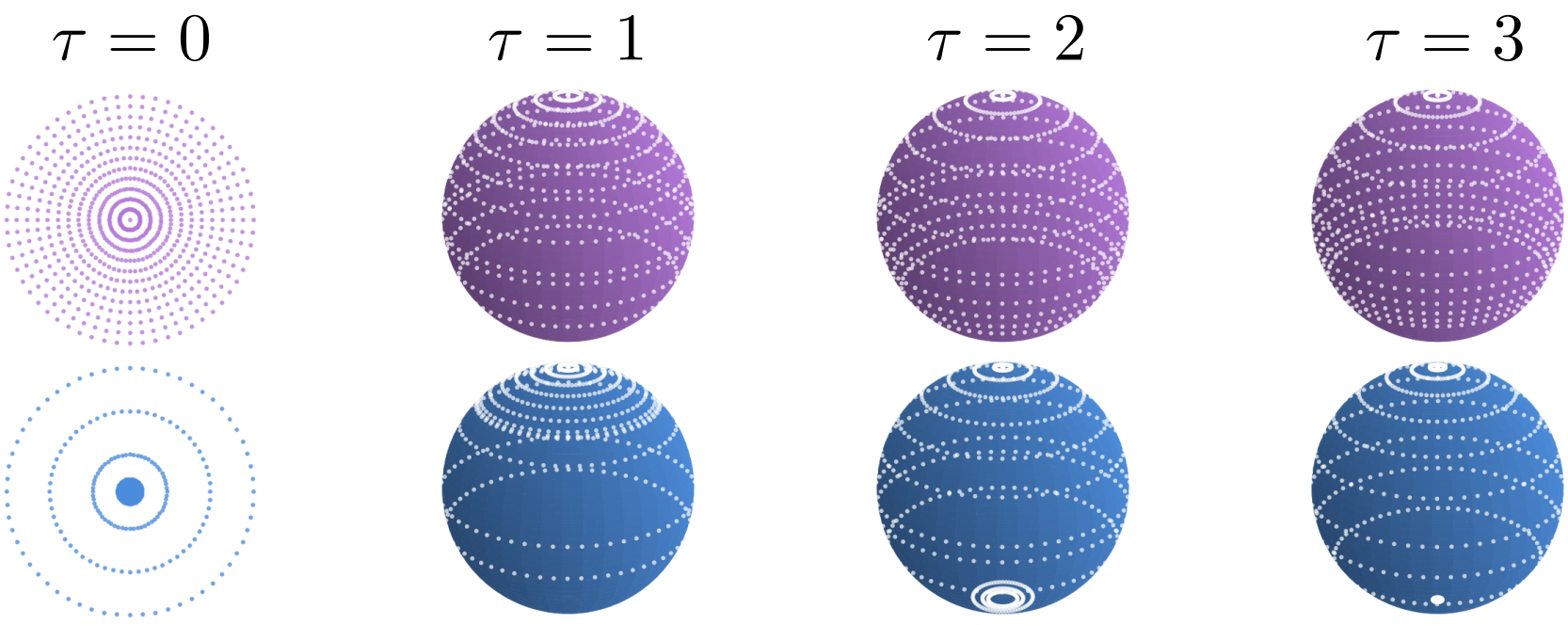
- **(a)** Two Gaussian distributions with equal variance but $m_L = 70$ and $m_R = 30$.
- **(b)** Circularly distributed points $r_{\text{outer}} = 2r_{\text{inner}}$.
- **(c)** "COS" data set.

Local anisotropy: landscape of κ



- **(a)** Two Gaussian distributions with equal variance but $m_L = 70$ and $m_R = 30$.
- **(b)** Circularly distributed points $r_{\text{outer}} = 2r_{\text{inner}}$.
- **(c)** "COS" data set.

Global deformation – $\mathbb{R}^n \longrightarrow S^n$



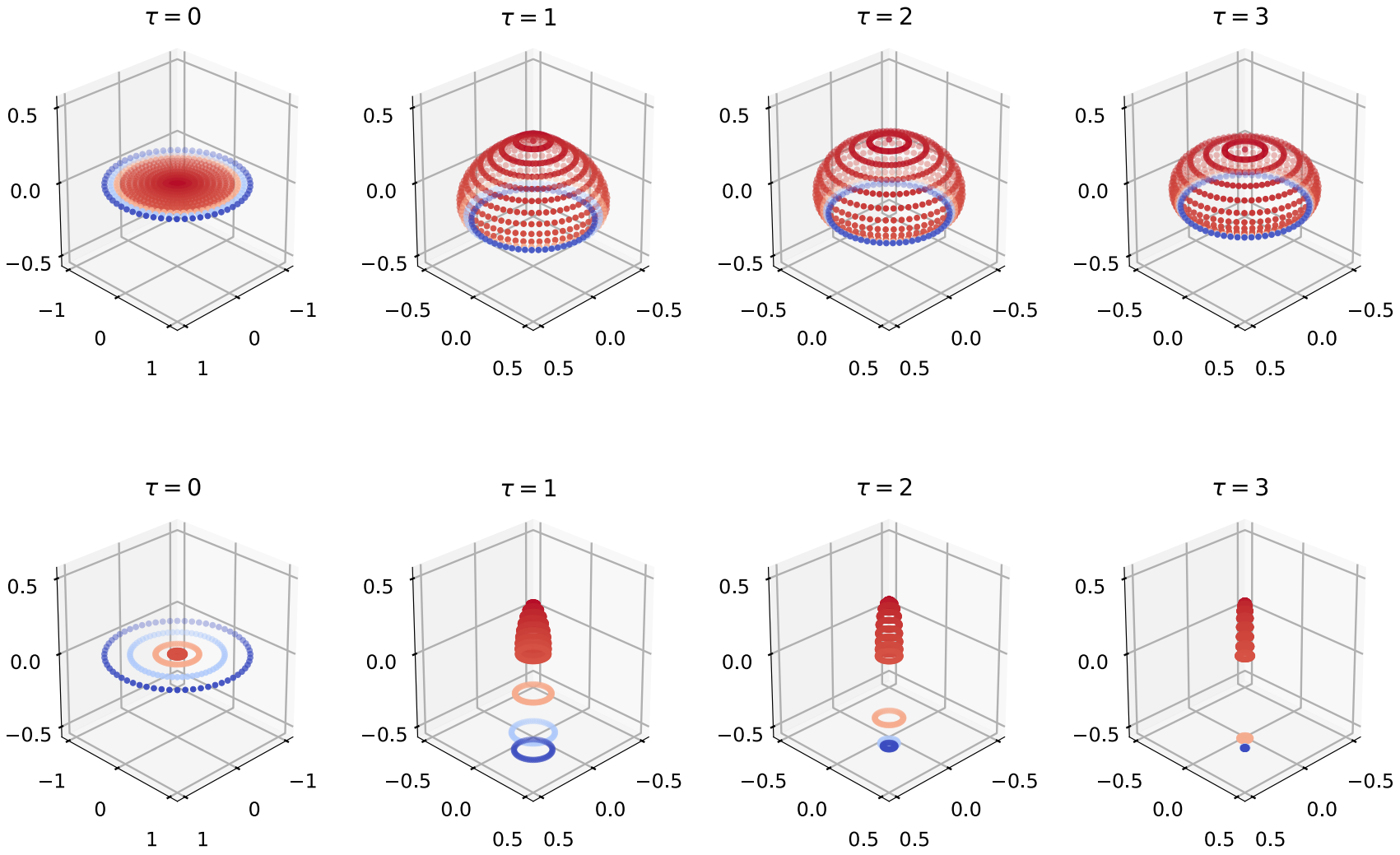
- EDT globally warped the Euclidean space $\mathbb{R}^2$ to a S^2 where points far from the origin in $\mathbb{R}^2$ are asymptotically identified as the south pole in S^2.

MDS reconstructed data distributions in $\mathbb{R}^n$

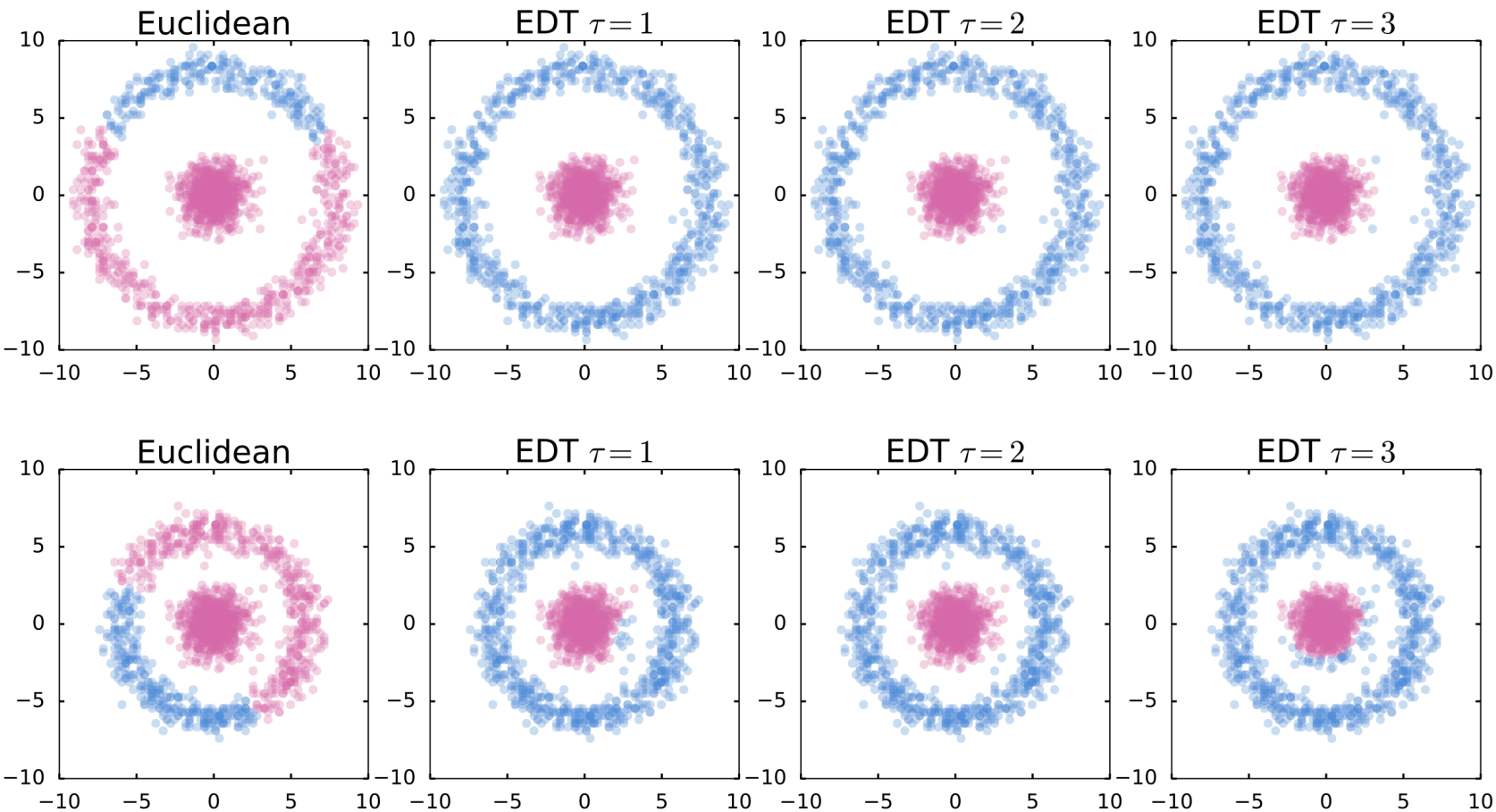
Multidimensional scaling (MDS): $d_{ij} \mapsto \{\mathbf{x}_i\}$

1. Input: a distance matrix d_{ij}
2. Centered squared distance: $B = -\frac{1}{2}JDJ$ where $D_{ij} = d_{ij}^2$ and $J = I - \frac{1}{m}11^\top$
3. Diagonalization: $B\mathbf{v}_k = \lambda_k \mathbf{v}_k$ where eigenvalues are ranked $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$; if there exists $\lambda_j < 0$, the embedding is approximate.
4. Embedding: $\mathbf{x}_i = ((\mathbf{v}_1)_i\sqrt{\lambda_1}, (\mathbf{v}_2)_i\sqrt{\lambda_2}, \cdots, (\mathbf{v}_{m^+})_i\sqrt{\lambda_{m^+}})^\top \in \mathbb{R}^{m^+}$ where $\lambda_i > 0$ for $i = 1, 2, \cdots, m^+$.

MDS of effective dissimilarities $d_{ij}^{(\tau)}$

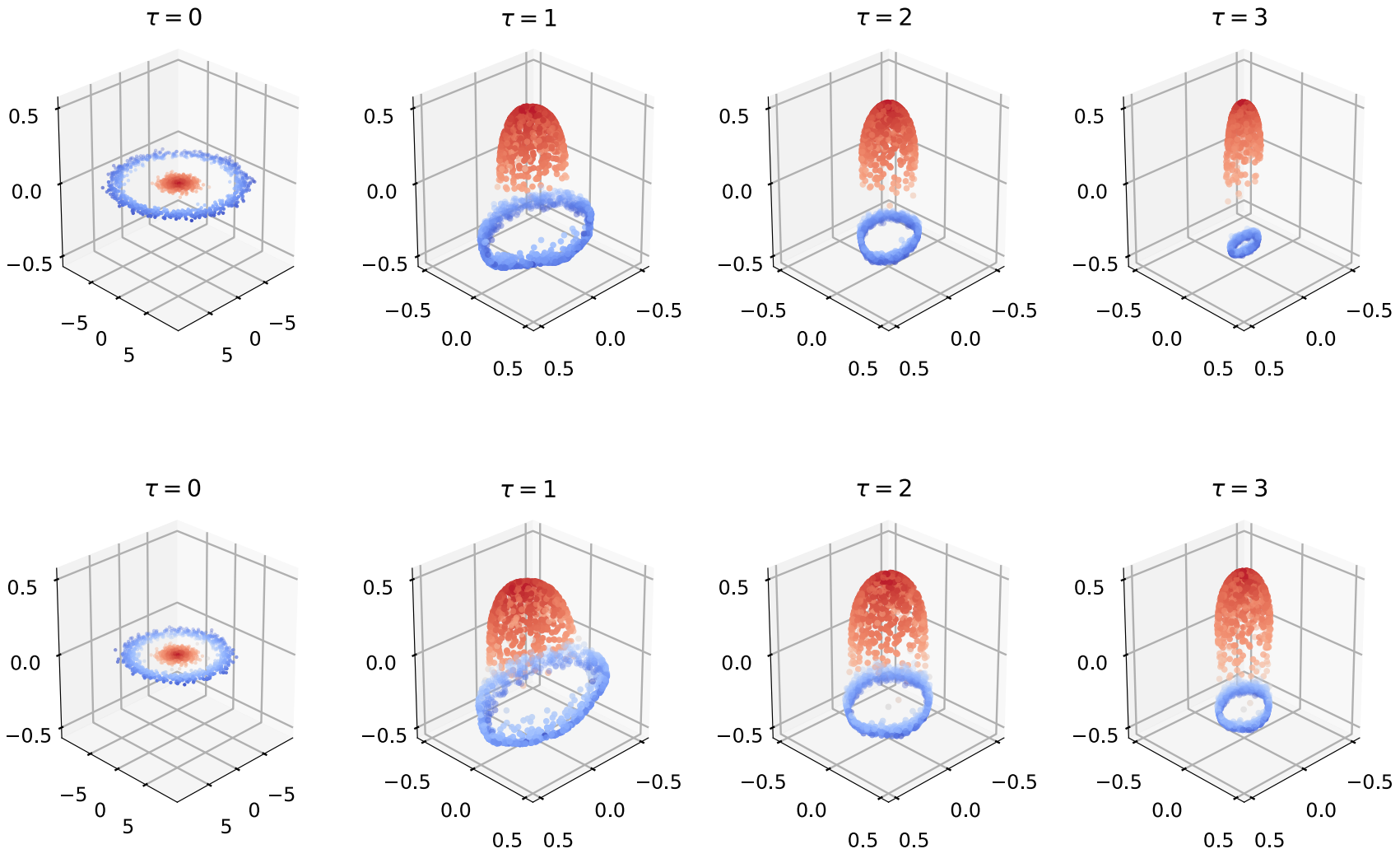


Global deformation – $\mathbb{R}^n \longrightarrow S^n$

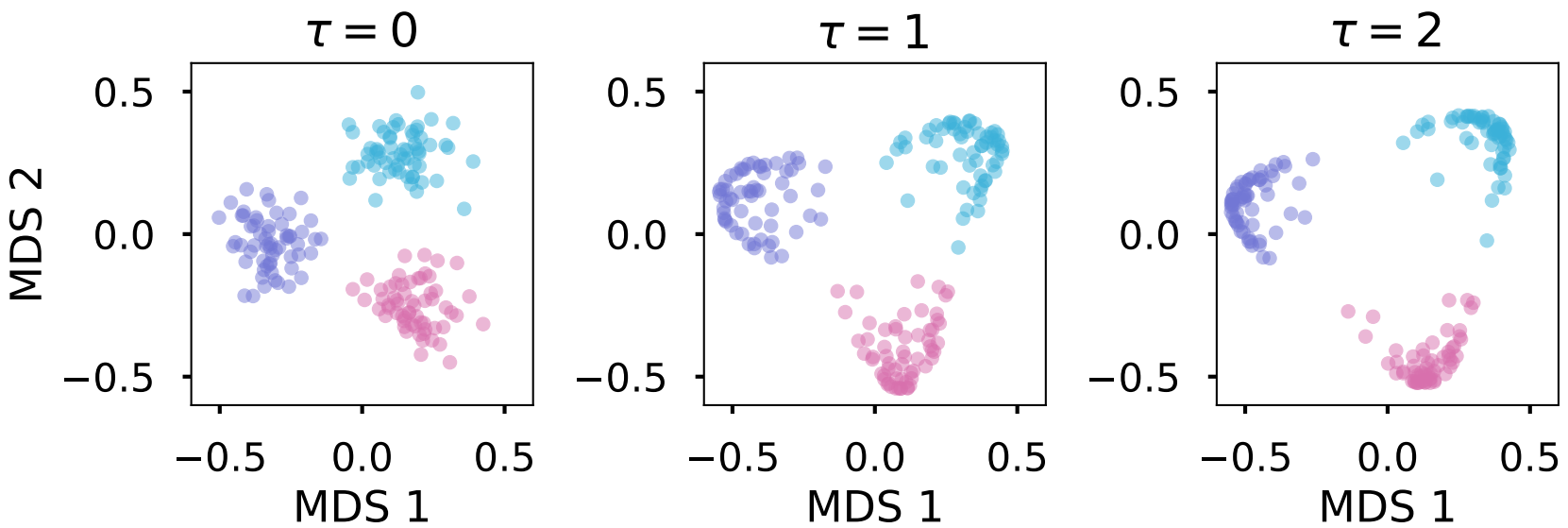


- EDT improved hierarchical clusterings of annulus data sets.

MDS of effective dissimilarities $d_{ij}^{(\tau)}$

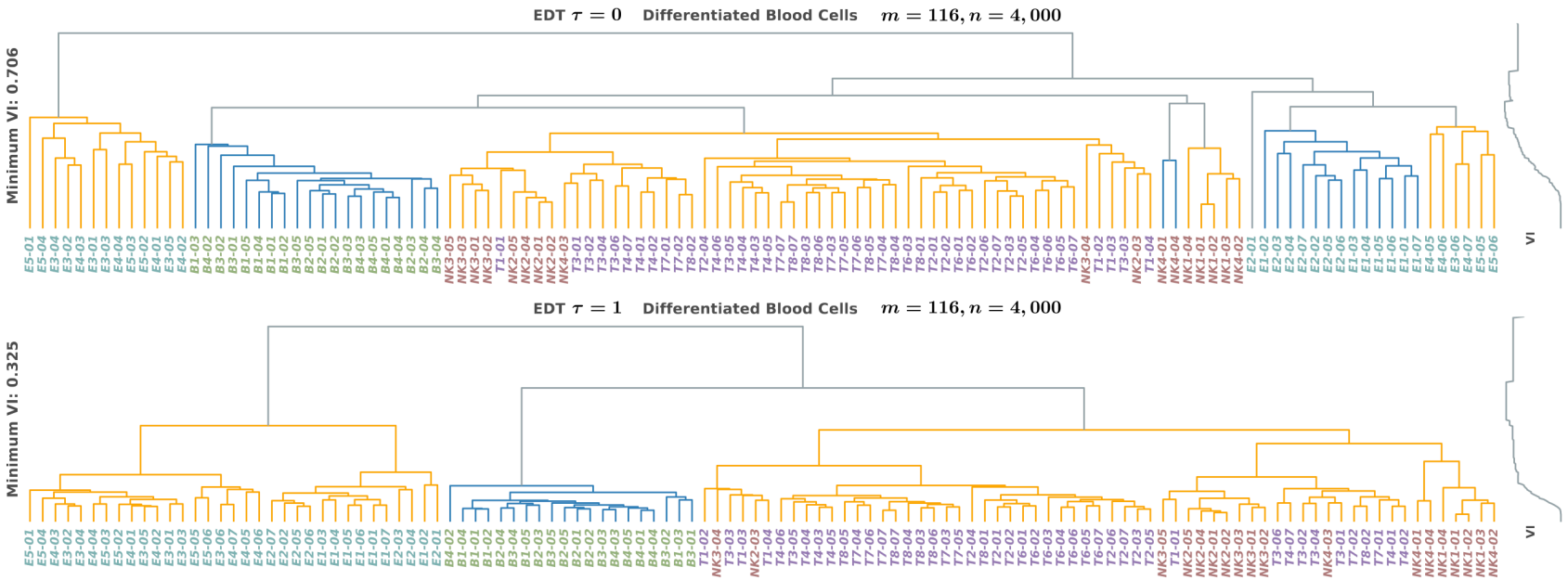


MDS visualization of cluster-condensation



- Each cluster condensed and ran away from each other.

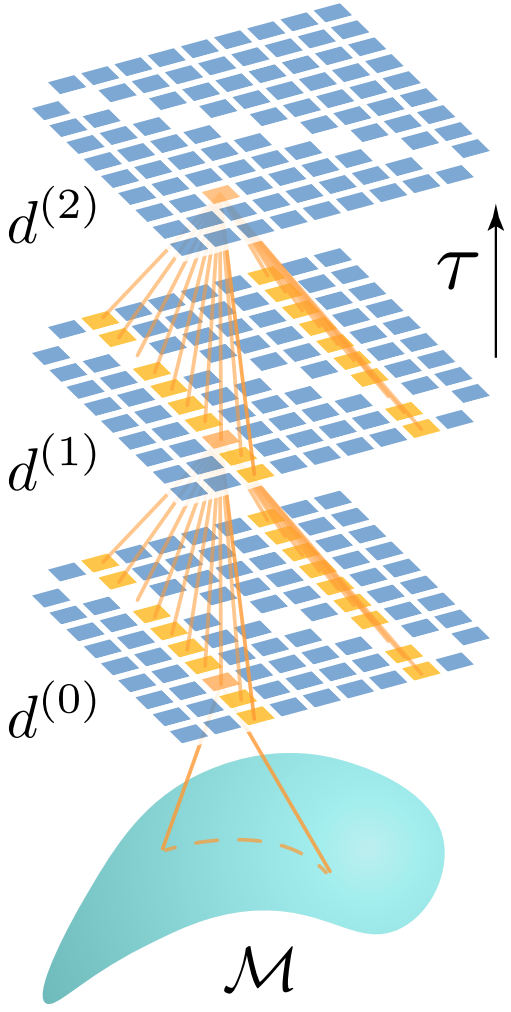
Application in blood cell differentiation data



- Red blood cells (Ej), T cells, B cells, and natural killer (NK) cells

EDT revealed geometric structures based on distances

- Clusters condensed and drifted away from each other.
- Contrasts of Euclidean distances in high dimensions were adaptively improved.
- Euclidean feature spaces were deformed into hyperspheres $\mathbb{R}^n \longrightarrow S^n$.
- Accuracy of hierarchical clustering was often improved.

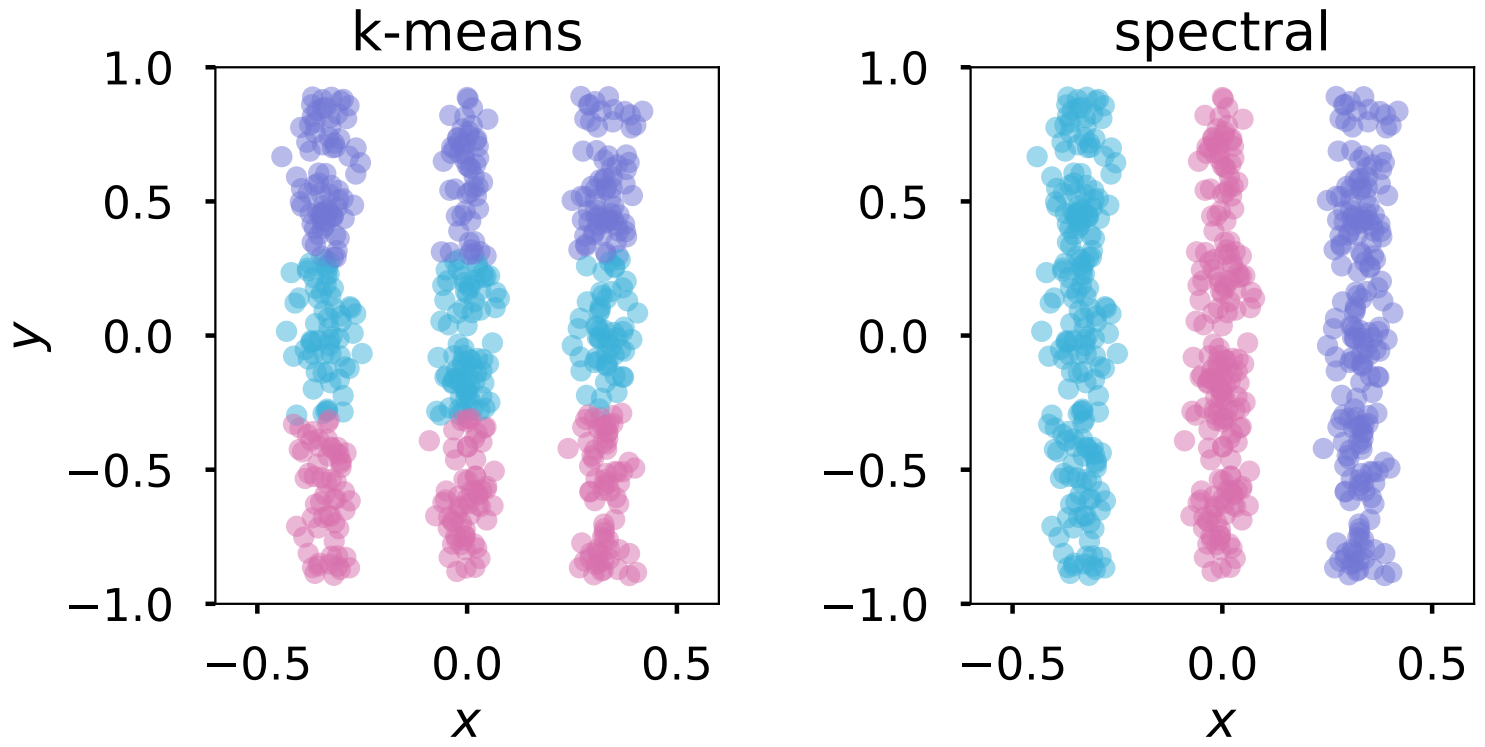


Chapter 3 Quantum Transport Clustering

Question

What if the clusters are not in spherical shapes but show complex geometric patterns?

Spectral clustering is the state of the art



- **Spectral clustering** (SC) was motivated by spectral graph theory (Ng, Jordan, and Weiss 2001).
- For points in $\mathbb{R}^n$, often take $A_{ij}^{\text{r.b.f.}} = \exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{r_\epsilon^2}\right]$.

Quantum version of spectral clustering

- SC is based on eigenvectors of a **symmetric** graph Laplacian

$$H = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

  where A_{ij} is **adjacency** between nodes (i, j), and **degree** matrix $D = \mathrm{diag}(\sum_j A_{ij}) = \mathrm{diag}(\sum_j A_{ij})$.
- SC has been interpreted using classical random walk (von Luxburg 2007).
- Hermitian matrix H can be naturally interpreted as a **Hamiltonian**, with eigenvalues E_n and eigenvectors $|\psi_n\rangle$ for $n = 0, 1, 2, \cdots$

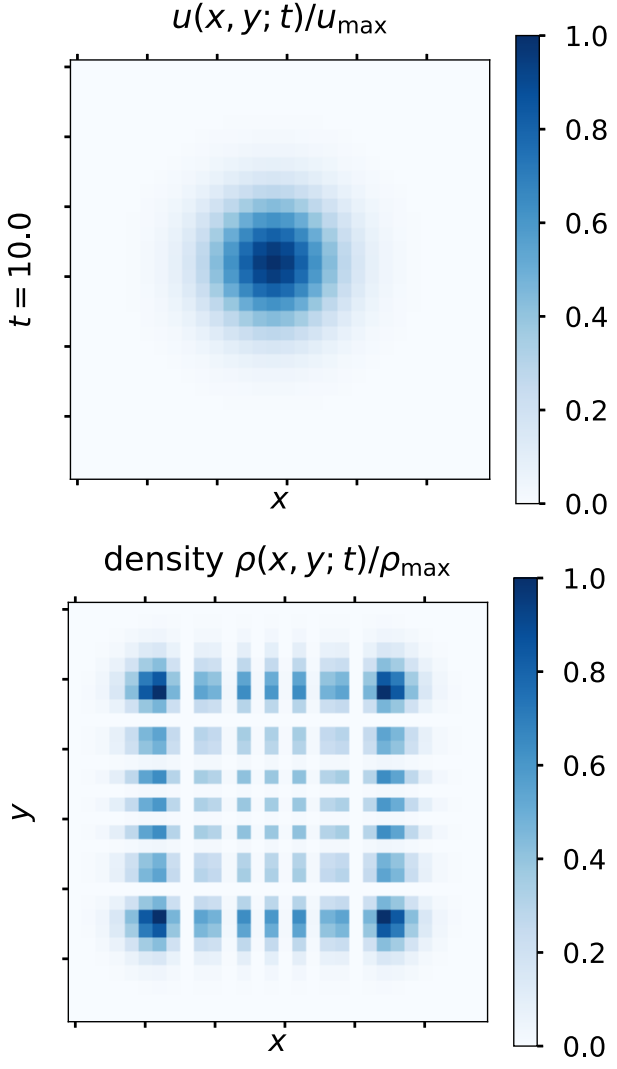
Can we measure similarity with wave functions?

- Propagator $\psi_{i|j}(t) = \langle i|\exp\{-\mathrm{i}tH\}|j\rangle$
  - Oscillates in space and time
  - Anderson localization (1958)
- Laplace transform (Anderson 1958)

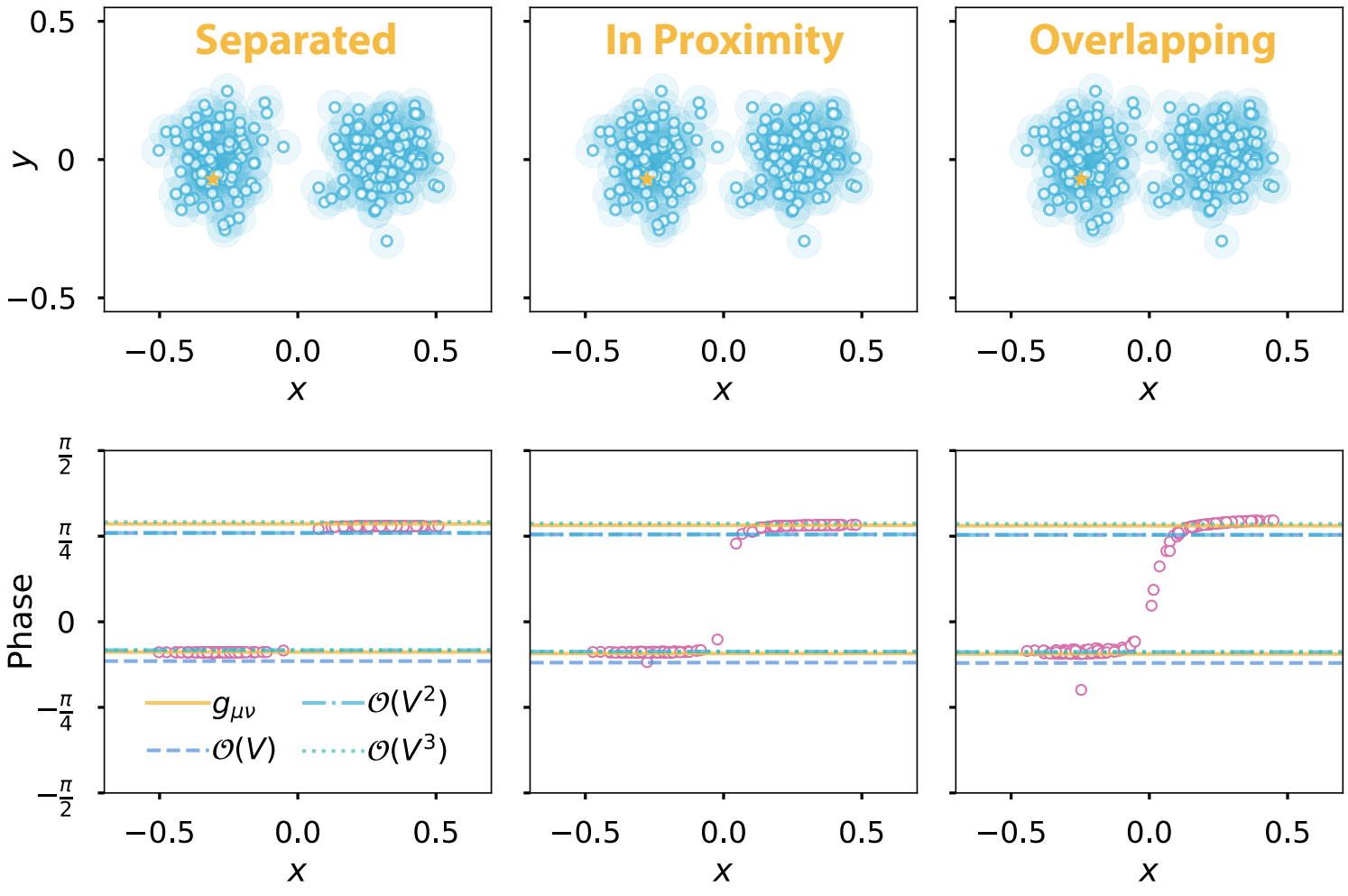
$$\mathcal{L}[\psi_{i|j}](s) = \langle i|(s+\mathrm{i}H)^{-1}|j\rangle = \sum_n \frac{\langle i|\psi_n\rangle\langle\psi_n|j\rangle}{s+\mathrm{i}E_n}.$$

- Let $G(z) \equiv (z-H)^{-1}$ with $z \in \mathbb{C}$, then

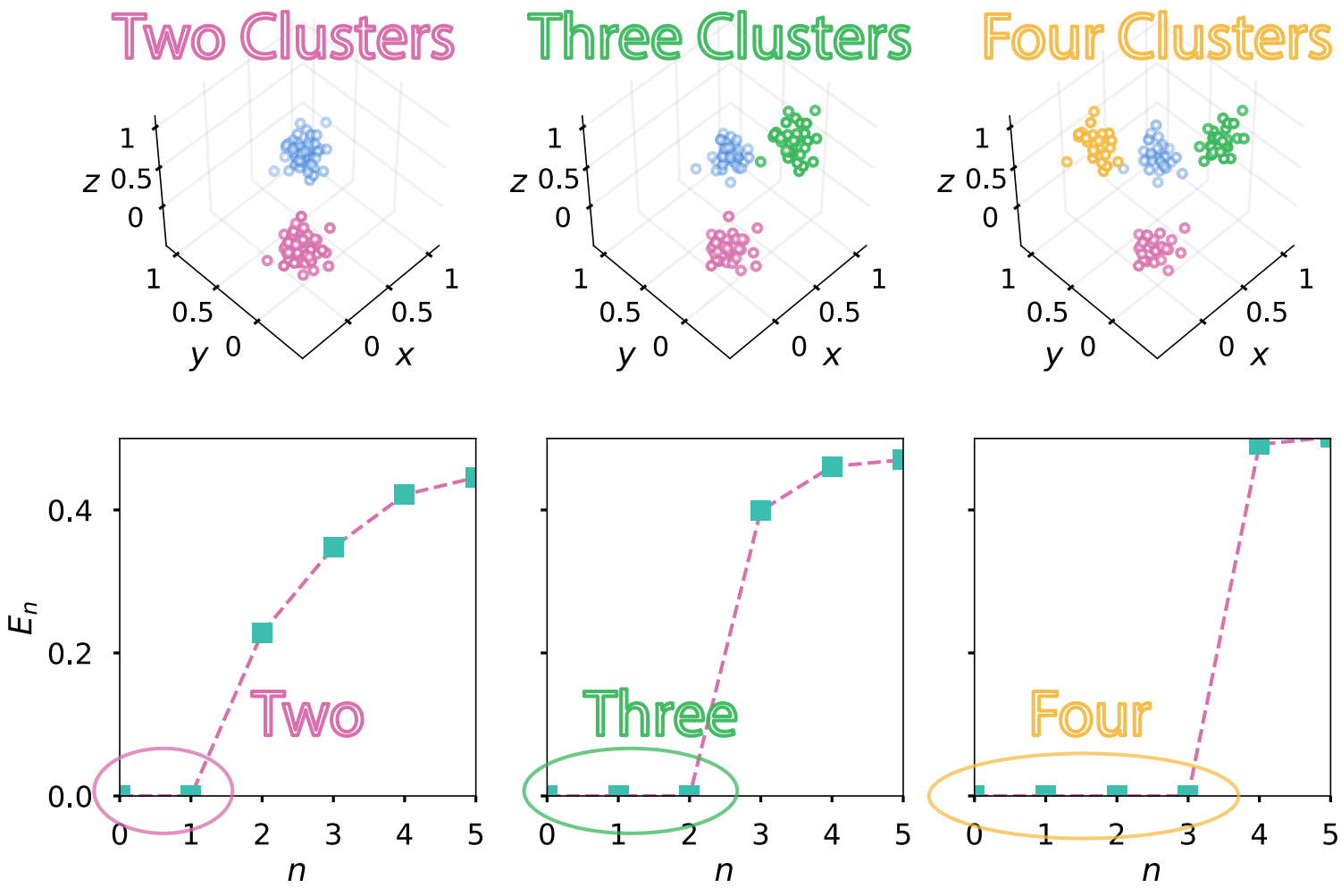
$$\mathcal{L}[\psi_{i|j}](s) = \mathrm{i}\langle i|G(\mathrm{i}s)|j\rangle.$$



Phase of $\mathcal{L}[\psi_{i|j}](s)$ distinguished clusters



Low energy modes capture macroscopic structures

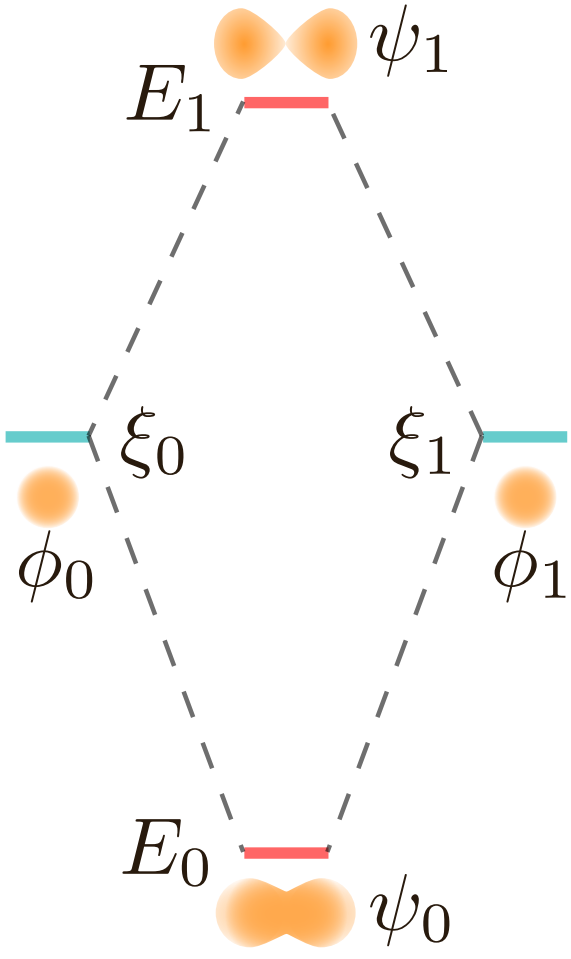


Quantum tunneling couples "atoms" into a "molecule"

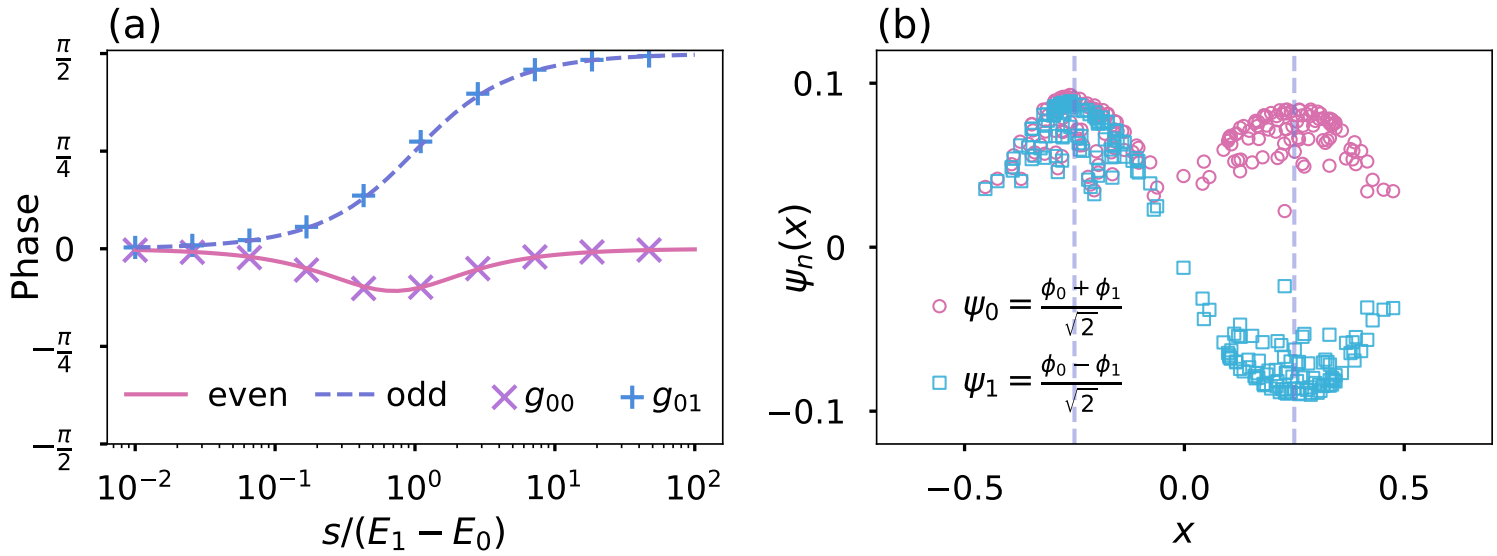
- Effective tight-binding model

$$\hat{H} \equiv \sum_{\mu,\nu} h_{\mu\nu} |\phi_\mu\rangle\langle\phi_\nu| \text{ with } h_{\mu\nu} = \xi_\mu \delta_{\mu\nu} + v_{\mu\nu}$$

- $\xi_\mu = \langle\phi_\mu|H|\phi_\mu\rangle$ is atomic ground state energy.
- $v_{\mu\nu} = \langle\phi_\mu|H|\phi_\nu\rangle$ is tunneling matrix element.
- Eigenvectors of hermitian matrix $h_{\mu\nu}$ combine **atomic** orbitals $\{\phi_\mu\}_{\mu=0}^{q-1}$ into **molecular** orbitals $\{\psi_n\}_{n=0}^{q-1}$.

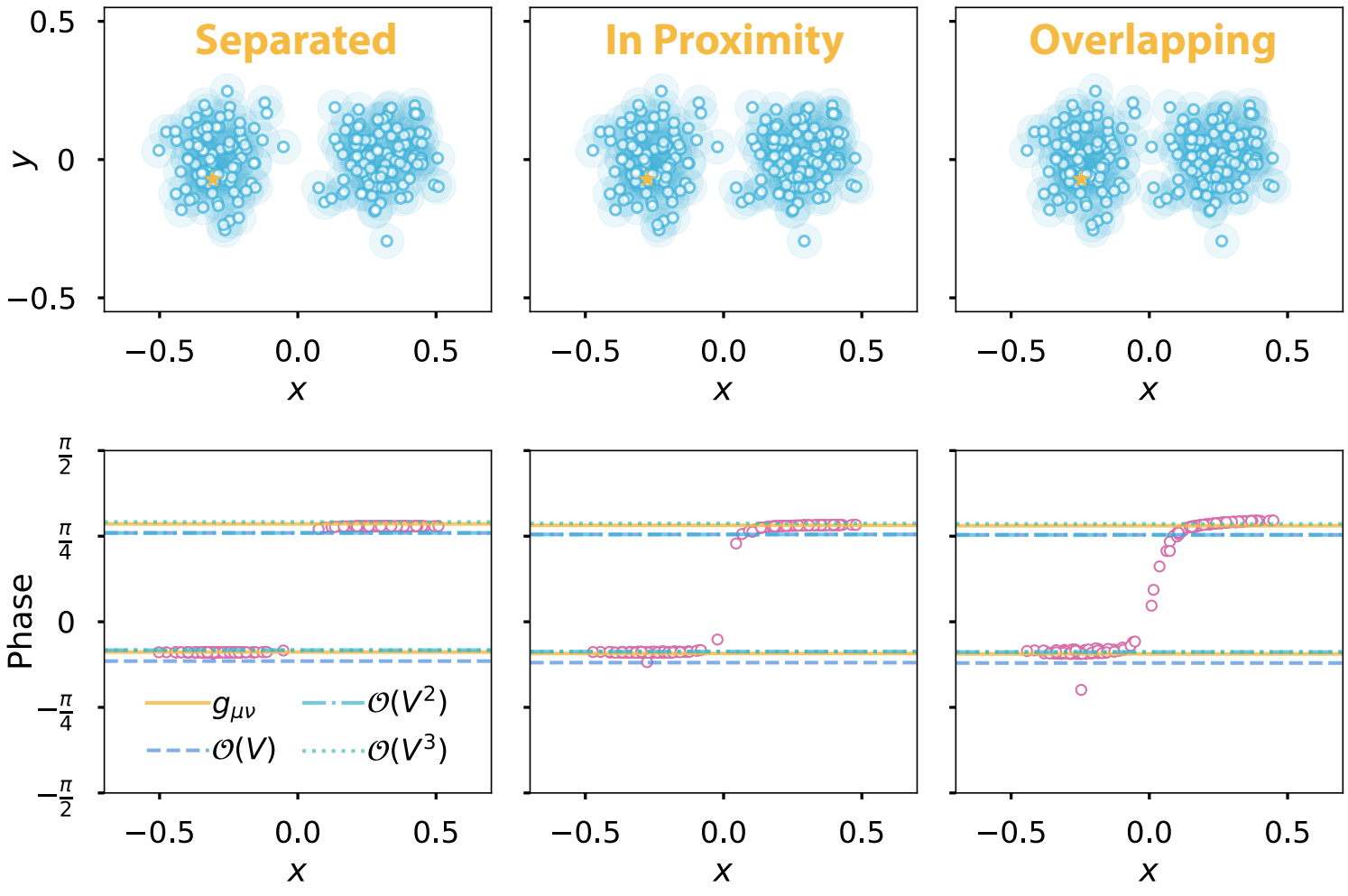


Two-cluster mapped to double-well instanton solution



- **Effective resolvent** is $\hat{G}(z) = (z - \hat{H})^{-1}$ with matrix element $g_{\mu\nu}(z) = (z - h_{\mu\nu})^{-1}$ and $z \in \mathbb{C}$.
- Set $z = \mathrm{i}s$ and take **phase** of $\mathrm{i}g_{\mu\nu}(\mathrm{i}s)$.

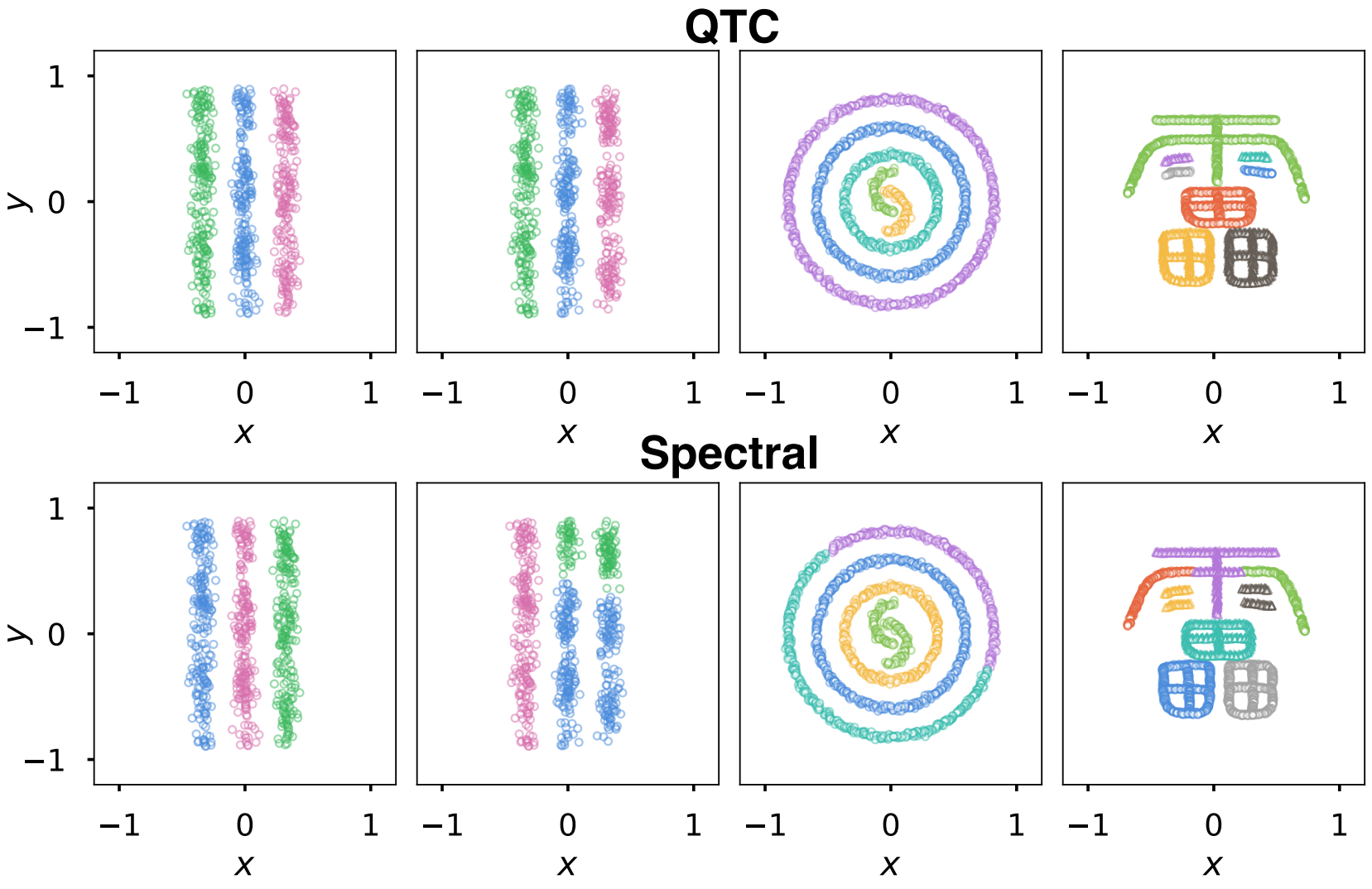
The phases resulted from tunnelings between clusters



Quantum Transport Clustering (QTC)

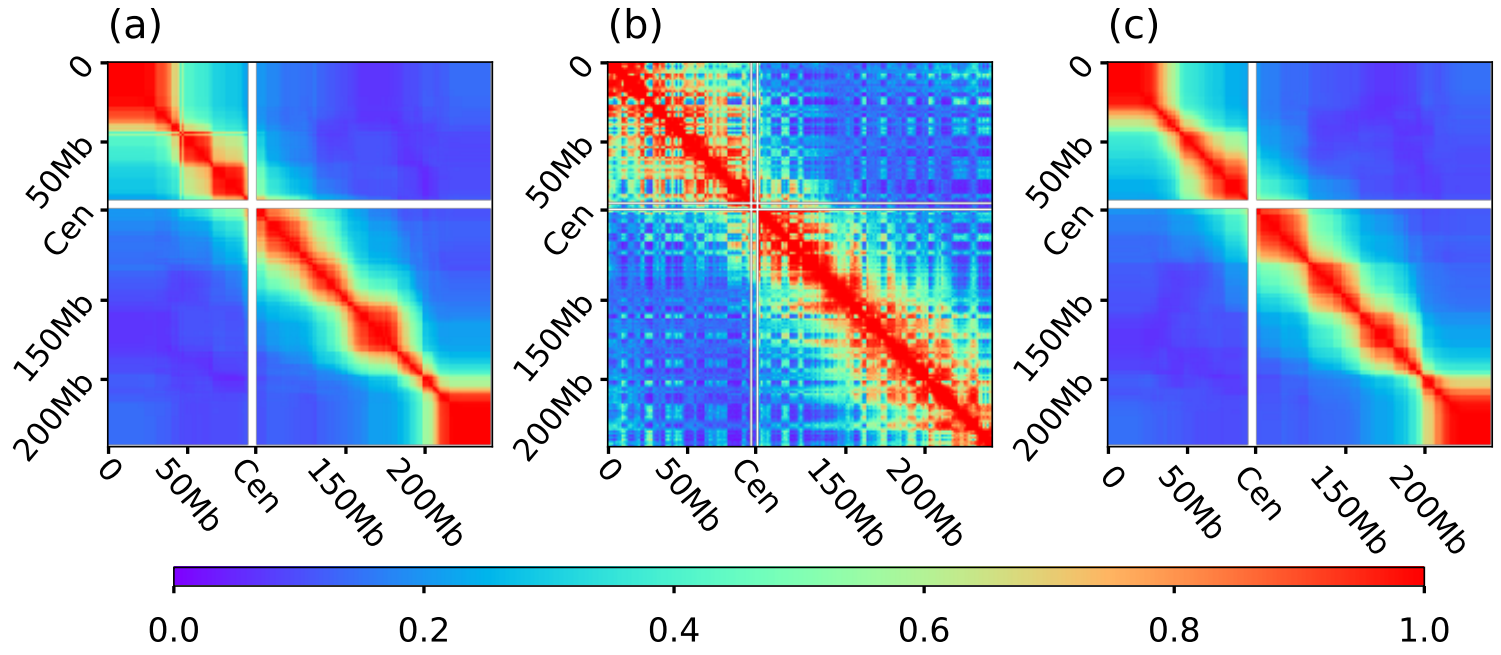
- The phase $\Theta_{i|j} = \arg \mathrm{i} G_{i|j}(\mathrm{i}s)$ embeds node i at $(\cos\Theta_{i|j}, \sin\Theta_{i|j}) \in S^1$.
- $\Theta_{\cdot|j}$ can be turned into integer labels $\Omega_{\cdot j}$
- Each initial j gives a different point of view.
- With m initializations, we get an **ensemble** of labels organized into Ω:
  - Direct extraction (`espresso`) – distinct clustering vectors $\Omega_{\cdot\alpha}$ and their frequency w_α with $\sum_\alpha w_\alpha = 1$.
  - Consensus matrix (`coldbrew`) – $C_{ij} = \frac{1}{m}\sum_{k=1}^{m} \delta(\Omega_{ik} - \Omega_{jk})$.

QTC was more robust than spectral embedding



QTC captured 3D structures in human genome

- QTC supported the hypothesis by Fudenberg et. al. (2011).



(**a**, **c**) QTC **consensus matrix** obtained using low-grade glioma and glioblastoma **somatic copy number alteration** data (TCGA) of human chromosome 2. (**b**) HiC 3D contact map (ENCODE).

QTC is more informative and natural in formalism

- Source code: `github.com/jssong-lab/QTC`
- QTC does not require hard truncation of spectrum.
- Embedding dimension $\dim S^1 = 1$.
- Probabilistic interpretations:
  1. Empirical distribution of distinct clusterings
     $w_\alpha = \mathbb{P}(\text{clustering} \sim \Omega_\alpha)$
  2. Empirical consensus of any two nodes
     $C_{ij} = \mathbb{P}(i, j \text{ in same cluster})$
- The low-energy modes of graph Laplacian are responsible for macroscopic patterns.

Diffusion processes improved machine learning methods

1. When the **geometry** of feature space is not $\mathbb{R}^n$, we can use heat kernel to measure similarities between samples.
2. EDT is able to reveal hidden structures based on **pairwise distances** through a discrete-time data point drifting process.
3. QTC is able to separate convoluted communities in network and distributions of **complex geometric shapes** in $\mathbb{R}^n$.

Conclusion

Diffusion processes often spontaneously captured the structures in data and improved the performance of machine learning algorithms.

Acknowledgements

- I would like to thank my advisor Professor Jun S. Song who has been continuously encouraging me to push myself to unlock my potentials.
- This project would not have been possible without the help and comments from my friends and colleagues.
- The works were supported by the **Sontag Foundation** and the **Grainger Engineering Breakthroughs Initiative**.