Statistical Exploration of the Manifold Hypothesis

Bernardo Marenco

IMERL FACULTAD DE INGENIERÍA UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Seminario Optimización y Aprendizaje Automático
24 de abril de 2025

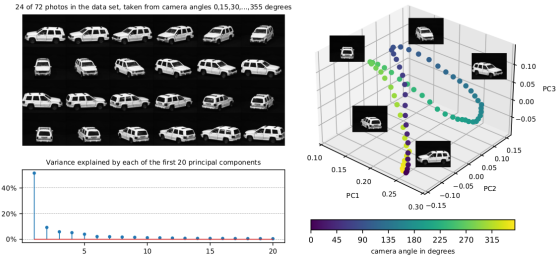
The Manifold Hypothesis

- *Statistical exploration of the manifold hypothesis* – Whiteley, Gray, Rubin-Delanchy (2025), arXiv:2208.11665v5
- High-dimensional data often concentrate near a low-dimensional *manifold* embedded in ambient space.

> (**Cayton, 2005**) "...the dimensionality of many data sets is only artificially high; though each data point consists of perhaps thousands of features, it may be described as a function of only a few underlying parameters. That is, the data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space"

- Observed empirically in images, speech, genomics, neuroscience, ...
- Underlies manifold learning, nonlinear dimension reduction, deep learning theory.

Motivating Examples



Car images: 75 grayscale images of resolution 384×288, camera angles
PCA reveals a loop structure (circle of angles).

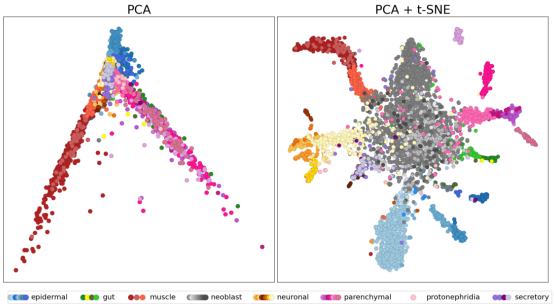


Figure 2: Planaria example. Left: first 2 dimensions of the PCA embedding. Right: representation of the data in 2 dimensions obtained by first reducing to 14 dimensions using PCA, then applying t-SNE.

Planaria cells: $n = 5000$, $p = 5821$ gene expressions, PCA+t-SNE shows branching tree-like structure.

Latent Metric Model

Assume data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ arises from:

$$Y_{ij} = X_j(Z_i) + \sigma E_{ij}.$$

where

- **Latent variables** $Z_i \sim \mu$ i.i.d, μ Borel measure in **compact** metric space $(\mathcal{Z}, d_Z)$
- **Random functions** $X_j : \mathcal{Z} \to \mathbb{R}$, $\mathbb{E}\left[X_j(z)^2\right] < \infty \, \forall z \in \mathcal{Z}$
- **Noise** E_{ij}, zero mean, unit variance. Matrix $\mathbf{E}$ is independent across columns, and element in different rows are uncorrelated

Mean Correlation Kernel

Mean correlation kernel $f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$

$$f(z, z') = \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}[X_j(z) X_j(z')].$$

Assumption 1: For each $j = 1, \ldots, p$, $\mathbb{E}\left[X_j(z) X_j(z')\right]$ is a continuous function of $(z, z') \in \mathcal{Z} \times \mathcal{Z}$.

(Mercer's theorem). *If $\mathcal{Z}$ is a compact metric space, μ a finite Borel measure supported on $\mathcal{Z}$ and $f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a symmetric, positive semi-definite, continuous function, there exists nonnegative numbers $(\lambda_k^f)_{k \geq 1}$, $\lambda_1^f \geq \lambda_2^f \geq \cdots$, and $\mathbb{R}$-valued functions $(u_k^f)_{k \geq 1}$ orthonormal in $L_2(\mu)$, such that:*

$$f(z, z') = \sum_{k=1}^{\infty} \lambda_k^f u_k^f(z) u_k^f(z'), \quad z, z' \in \mathcal{Z},$$

where the convergence is absolute and uniform.

Why is f Positive Semi-Definite?

- Each $f_j(z, z') := \mathbb{E}[X_j(z)X_j(z')]$ is a positive semi-definite kernel:
  - For any $a_1, \ldots, a_n \in \mathbb{R}$ and $z_1, \ldots, z_n \in \mathcal{Z}$,

$$\sum_{i,k=1}^{n} a_i a_k \mathbb{E}[X_j(z_i)X_j(z_k)] = \mathbb{E}\left[\left(\sum_{i=1}^{n} a_i X_j(z_i)\right)^2\right] \geq 0$$

  since the square expands to a double sum:

$$\left(\sum_{i} a_i X_j(z_i)\right)^2 = \sum_{i,k} a_i a_k X_j(z_i)X_j(z_k)$$

- $f(z, z') = \frac{1}{p}\sum_{j=1}^{p} f_j(z, z')$ is an average of PSD kernels, so it is PSD.

Feature map

By Mercer's theorem:

$$f(z,z') = \sum_{k=1}^{\infty} \lambda_k^f u_k^f(z) u_k^f(z') = \langle \phi(z), \phi(z') \rangle_{\ell_2}$$

where we define the feature map $\phi : \mathcal{Z} \to \ell_2$ as:

$$\phi(z) = \left[(\lambda_1^f)^{1/2} u_1^f(z), (\lambda_2^f)^{1/2} u_2^f(z), \dots \right]$$

Manifold: $\mathcal{M} = \{\phi(z) : z \in \mathcal{Z}\} \subset \ell^2$ since $\|\phi(z)\|_{\ell_2}^2 = f(z,z)$, $\mathcal{Z}$ is compact and f is continuous

$r = \text{rank}(f)$ i.e. largest k such that $\lambda_k^f > 0$, with $r := \infty$ if $\lambda_k^f > 0$ for all $k \geq 1$

If $r < \infty$, write:

$$\phi(z) = \left[(\lambda_1^f)^{1/2} u_1^f(z), (\lambda_2^f)^{1/2} u_2^f(z), \dots, (\lambda_r^f)^{1/2} u_r^f(z) \right]$$

Relating data inner products to feature map inner products

LMM model: data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ arises from:

$$Y_{ij} = X_j(Z_i) + \sigma E_{ij}.$$

Proposition 1: If A1 holds, then

$$\mathbf{Y}_i \overset{m.s.}{=} p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i, \quad \mathbb{E}\left[\mathbf{W}^\top\mathbf{W}\right] = \mathbf{I}_r$$

where

$$\mathbf{W}_{jk} := \frac{1}{(p\lambda_k^f)^{1/2}} \int_{\mathcal{Z}} X_j(z) u_k^f(z) \mu(dz)$$

i.e. $p^{-1/2}\mathbf{Y}_i$ is a noisy, random projection of $\phi(Z_i)$

Proof key ingredient: (u_k, λ_k) are $L_2(\mu)$-orthonormal eigenfunctions and eigenvalues of the integral operator T_f associated with the kernel f and the measure μ:

$$(T_f\psi)(z) := \int_{\mathcal{Z}} f(z, z')\,\psi(z')\,d\mu(z')$$

Relating data inner products to feature map inner products

Proposition 1: If A1 holds, then

$$\mathbf{Y}_i \overset{m.s.}{=} p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i, \quad \mathbb{E}\left[\mathbf{W}^\top\mathbf{W}\right] = \mathbf{I}_r$$

Then:

$$\begin{aligned}\frac{1}{p}\mathbb{E}\left[\langle\mathbf{Y}_i, \mathbf{Y}_j\rangle | Z_i, Z_j\right] &= \langle\phi(Z_i), \mathbb{E}\left[\mathbf{W}^\top\mathbf{W}\right]\phi(Z_j)\rangle_{\ell_2} + \sigma^2\frac{1}{p}\mathbb{E}\left[\langle\mathbf{E}_i, \mathbf{E}_j\rangle\right] \\ &= \langle\phi(Z_i), \phi(Z_j)\rangle_{\ell_2} + \sigma^2\mathbf{1}_{i=j}\end{aligned}$$

Hence, by the law of large numbers:

$$|p^{-1}\langle\mathbf{Y}_i, \mathbf{Y}_j\rangle - \langle\phi(Z_i), \phi(Z_j)\rangle_{\ell_2} - \sigma^2\mathbf{1}_{i=j}| \to 0 \text{ when } p \to \infty$$

(under weak dependency and bounded moments assumptions)

This implies:

$$|p^{-1}\|Y_i - Y_j\|_2^2 - \|\phi(Z_i) - \phi(Z_j)\|_{\ell_2}^2 - 2\sigma^2| \to 0 \text{ when } p \to \infty$$

Topological Equivalence

$\phi : \mathcal{Z} \to \mathcal{M}$ is a homeomorphism (continuous, bijective, has continuous inverse) between $(\mathcal{Z}, d_{\mathcal{Z}})$ and $(\mathcal{M}, d_{\mathcal{M}})$ where $d_{\mathcal{M}}(\cdot,\cdot) = \|\cdot - \cdot\|_{\ell_2}$

Continuity of ϕ: $d_{\mathcal{Z}}(z, z') \to 0$ implies $d_{\mathcal{M}}(\phi(z), \phi(z')) = \|\phi(z) - \phi(z')\|_{\ell_2} \to 0$

$$\|\phi(z) - \phi(z')\|_{\ell_2}^2 = \|\phi(z)\|_{\ell_2}^2 + \|\phi(z')\|_{\ell_2}^2 - 2\langle\phi(z), \phi(z')\rangle_{\ell_2} = f(z,z) + f(z',z') - 2f(z,z')$$

and f is continuous by **A1**.

Since ϕ is surjective by definition, if it is one-to-one its inverse is automatically continuous since $\mathcal{Z}$ is compact

Assumption 2 (Distinguishability):

$$\sum_{j=1}^{p} \mathbb{E}[|X_j(z) - X_j(z')|^2] > 0 \quad \forall z \neq z'.$$

Obs: **A2** is equivalent to: for each $z, z' \in \mathcal{Z}$ with $z \neq z'$ there exists $\xi \in \mathcal{Z}$ s.t. $f(z,\xi) \neq f(z',\xi)$

Topological Equivalence

Proposition 2: $\phi : \mathcal{Z} \to \mathcal{M}$ is a homeomorphism if and only if **A2** holds

Proof:

$$\begin{aligned}\|\phi(z) - \phi(z')\|_{\ell_2}^2 &= \|\phi(z)\|_{\ell_2}^2 + \|\phi(z')\|_{\ell_2}^2 - 2\langle\phi(z), \phi(z')\rangle_{\ell_2} = f(z,z) + f(z',z') - 2f(z,z') \\ &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[|X_j(z)|^2\right] + \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[|X_j(z')|^2\right] - \frac{2}{p}\sum_{j=1}^{p}\mathbb{E}\left[X_j(z)X_j(z')\right] \\ &= \frac{1}{p}\sum_{j=1}^{p}\mathbb{E}\left[|X_j(z) - X_j(z')|^2\right]\end{aligned}$$

Hence ϕ is one-to-one iff $\sum_{j=1}^{p}\mathbb{E}[|X_j(z) - X_j(z')|^2] > 0 \quad \forall z \neq z'$ □

Topological Equivalence

Since ϕ is a homeomorphism between $(\mathcal{Z}, d_{\mathcal{Z}})$ and $(\mathcal{M}, d_{\mathcal{M}})$:

- Can be transformed between each other by bending, twisting, stretching and folding, but **not** cutting, puncturing or joining
- Same number of connected components, 1-dimensional loops and k-dimensional "holes" as each other$\rightarrow$ **this is why persistent homology analysis works**
- They have the same covering dimension$\rightarrow$ **if $\mathcal{Z}$ is low-dimensional, $\mathcal{M}$ is low-dimensional**

Metric Equivalence (Isometry)

Assume X_j's are weakly stationary:

- $\mathbb{E}[X_j(z)]$ is constant in z
- $\mathbb{E}[(X_j(z) - \mathbb{E}[X_j(z)])(X_j(z') - \mathbb{E}[X_j(z')])]$ only depends on $d_{\mathcal{M}}(z, z')$

Then $f(z, z')$ would also depend only on $d_{\mathcal{M}}(z, z')$

Assumption 3: $\mathcal{Z} \subset \mathbb{R}^d$, and there exists a continuous path in $\mathcal{Z}$ of finite length between any two points in $\mathcal{Z}$

Paths on a manifold

Let $x, x' \in \mathcal{M}$, a path in $\mathcal{M}$ is a continuous function $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma(1) = x'$
A partition $\mathcal{P}$ of $[0,1]$ is a non-decreasing sequence $0 = t_0 < t_1 < \cdots < t_n = 1$.
For a path γ and partition $\mathcal{P}$, define:

$$\chi(\gamma, \mathcal{P}) := \sum_{k=1}^{n} \|\gamma(t_k) - \gamma(t_{k-1})\|_{\ell^2}.$$

The length of the path is:

$$L(\gamma) := \sup_{\mathcal{P}} \chi(\gamma, \mathcal{P}),$$

where the supremum is over all partitions.

Geodesic Distances in $\mathcal{M}$ and $\mathcal{Z}$

Let $z, z' \in \mathcal{Z} \subset \mathbb{R}^d$ and $\eta : [0,1] \to \mathcal{Z}$ be a path with $\eta(0) = z$, $\eta(1) = z'$. Define length similarly:

$$\chi(\eta, \mathcal{P}) := \sum_{k=1}^{n} \|\eta(t_k) - \eta(t_{k-1})\|_{\mathbb{R}^d}, \quad L(\eta) := \sup_{\mathcal{P}} \chi(\eta, \mathcal{P}).$$

Then the geodesic distances are defined by:

$$d_{\mathcal{M}}^{\text{geo}}(x, x') := \inf_{\gamma : \gamma(0)=x, \gamma(1)=x'} L(\gamma),$$

$$d_{\mathcal{Z}}^{\text{geo}}(z, z') := \inf_{\eta : \eta(0)=z, \eta(1)=z'} L(\eta),$$

where the infima are taken over all continuous paths in $\mathcal{M}$ and $\mathcal{Z}$ respectively.

Metric Equivalence (Isometry)

Proposition 3: Assume **A1**, **A2** and **A3** hold. Define $D := \{(z, z), z \in \mathcal{Z}\} \subset \mathcal{Z} \times \mathcal{Z}$. If $f(z, z') = g(\|z - z'\|_2^2)$ for all z, z' in an open neighbourhood of D and g is twice continuously differentiable and $g'(0) < 0$, then

$$d_{\mathcal{M}}^{\text{geo}}(\phi(z), \phi(z')) = \sqrt{-2g'(0)}\, d_{\mathcal{Z}}^{\text{geo}}(z, z')$$

Proof idea: Show that ϕ is a bi-Lipschitz homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$.Then, use that for any path γ in $\mathcal{M}$ of finite length, there exists a path η in $\mathcal{Z}$ such that the following holds:
For any $\varepsilon > 0$ there exists a partition P_ε such that for any partition $P = (t_0, ..., t_n)$ satisfying $P_\varepsilon \subset P$:

$$\left| L(\gamma) - \sum_{k=1}^{n} \langle \eta(t_k) - \eta(t_{k-1}), \mathbf{H}_{\eta(t_{k-1})}(\eta(t_k) - \eta(t_{k-1})) \rangle^{1/2} \right| \leq \varepsilon$$

where $(\mathbf{H}_\xi)_{ij} := \frac{\partial^2 f}{\partial z_i \partial z_j'}\Big|_{(\xi,\xi)}$ for $\xi \in \mathcal{Z}$

Isometry on the Sphere

If $\mathcal{Z}$ is a sphere, Proposition 3 then becomes:

Proposition 4: Assume **A1** and **A2**. If $\mathcal{Z} = \{z \subset \mathbb{R}^d : \|z\|_2 = 1\}$ and $f(z, z') = g(\langle z, z' \rangle_{\mathcal{L}_2})$ for all z, z' in an open neighbourhood of D and g is twice continuously differentiable and $g'(1) > 0$, then

$$d^{\text{geo}}_{\mathcal{M}}(\phi(z), \phi(z')) = \sqrt{g'(1)} d^{\text{geo}}_{\mathcal{Z}}(z, z')$$

Smoothness to concentration in a low-dimensional subspace

When $\mathcal{Z} \subset \mathbb{R}^d$, we say that f is smooth if it is the restriction of a smooth function on $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathcal{Z} \times \mathcal{Z}$.

Low-rank approximation: for $s < r = \text{rank} f$, define the truncated map $\phi_s : \mathcal{Z} \to \ell_2$ as

$$\phi_s(z) = \left[(\lambda_1^f)^{1/2} u_1^f(z), (\lambda_2^f)^{1/2} u_2^f(z), \ldots, (\lambda_s^f)^{1/2} u_s^f(z), 0, \ldots \right]$$

Eigenvalues λ_i^f measure of how well $\mathcal{M}_s := \phi_s(\mathcal{Z})$ approximates $\mathcal{M}$ through mean square error:

$$\mathbb{E}\left[\|\phi(Z_i) - \phi_s(Z_i)\|_{\ell_2}^2 \right] = \sum_{k>s} \lambda_k^f \mathbb{E}\left[|u_k^f(Z_i)| \right] = \sum_{k>s} \lambda_k^f$$

Since rate of decay of the λ_i^f is related to the smoothness of the kernel f taking s suitably large the first s coordinates of ϕ provide a good approximation to $\mathcal{M}$, even if $r = \infty$

Smoothness to concentration in a low-dimensional subspace

When $s \leq p$ smoothness also implies each Y_i concentrates around the (at most) s-dimensional subspace of R^p spanned by the first s columns of $\mathbf{W}$

Remember **Proposition 1**: $\mathbf{Y}_i \overset{m.s.}{=} p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i, \quad \mathbb{E}\left[\mathbf{W}^\top\mathbf{W}\right] = \mathbf{I}_r$, so:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{Y}_i - p^{1/2}\mathbf{W}\phi_s(z_i)\right\|_2^2\right] &= \mathbb{E}\left[\left\|p^{1/2}\mathbf{W}\phi(Z_i) + \sigma\mathbf{E}_i - p^{1/2}\mathbf{W}\phi_s(z_i)\right\|_2^2\right] \\ &= p\mathbb{E}\left[\|\phi(Z_i) - \phi_s(Z_i)\|_{\ell_2}^2\right] + \sigma^2\mathbb{E}\left[\|\mathbf{E}_i\|_2^2\right] = p\sum_{k>s}\lambda_k^f + p\sigma^2\end{aligned}$$

Visual example

$\mathcal{Z} \subset \mathbb{R}^3$ is a torus, μ is uniform distribution. Sample $Z_1, \ldots, Z_{4000}$:

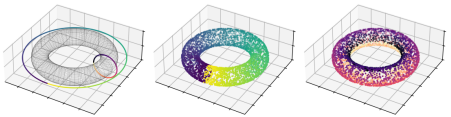
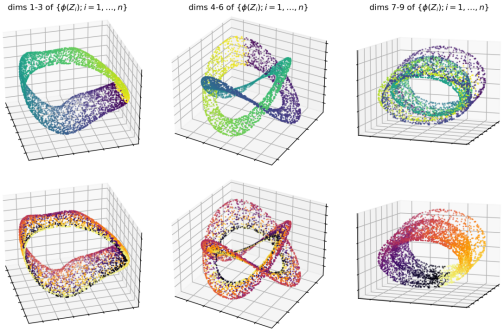


Figure 3: Torus example. Left: grey wireframe of $\mathcal{Z}$, a torus, with colour bars indicating coordinates with respect to two circles. Both the middle and right plots show the same $n = 4000$ points, $Z_1, \ldots, Z_{4000}$, which are sampled uniformly on the torus, coloured by their coordinates with respect to each of the two circles.

$X_1, \ldots, X_p$ i.i.d gaussian with covariance function $f(z, z') = \exp(-\|z - z'\|_2^2)$

Numerical approximations to the first 1-3, 4-6 and 7-9 dimensions of $\phi(Zi)$ using PCA:

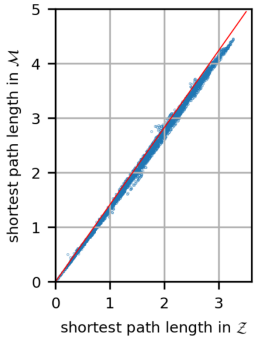


Global shape of $\mathcal{M}$, when viewed three dimensions at a time, is qualitatively different to the global shape of $\mathcal{Z}$

Visual example

However,assumptions **A1, A2** and **A3**, hold in this example

Theoretic scaling factor for geodesic distances is $\sqrt{-g'(0)} = \sqrt{2}$



Geodesic distances computed using nearest-neighbour graph (more on this coming up)

Dimension Reduction by PCA

Embed Y_i's into $\mathbb{R}^s$ via top-s eigenvectors V_s of $Y^\top Y$:

$$\zeta_i = V_s^\top Y_i$$

Assuming X_i's are independent, have finite fourth moment and $\text{rank}(f) = r < \infty$

Theorem 1: If $s = r$, there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r \times r}$ such that as $n, p/n \to \infty$,

$$\max_i \left\| p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i) \right\|_2 = O_p\left(\tfrac{1}{\sqrt{n}} + \sqrt{\tfrac{n}{p}} \right).$$

$\phi(Z_1), \ldots, \phi(Z_n)$ can be recovered from $p^{-1/2}\zeta_1, ..., p^{-1/2}\zeta_n$, up to an orthogonal transformation

Therefore, PCA can be viewed as de-noising + signal extraction

Viewed as sets, point clouds $\{p^{-1/2}\zeta_i\}_{i=1,\ldots,n}$ and $\{\phi(Z_i)\}_i$ converge to each other in Hausdorff distance (up to $\mathbf{Q}$)

Workflow Overview

1. **Dimension Selection** ($\widehat{r}$) via Wasserstein distance
2. **PCA Embedding** to $\widehat{\mathbb{R}^r}$
3. **Spherical Projection** $\zeta_i/\|\zeta_i\|$
4. **Nearest Neighbour Graph** on projected points
5. **Analysis**: shortest paths, MST, topology (persistent homology)

Step 1: Dimension Selection

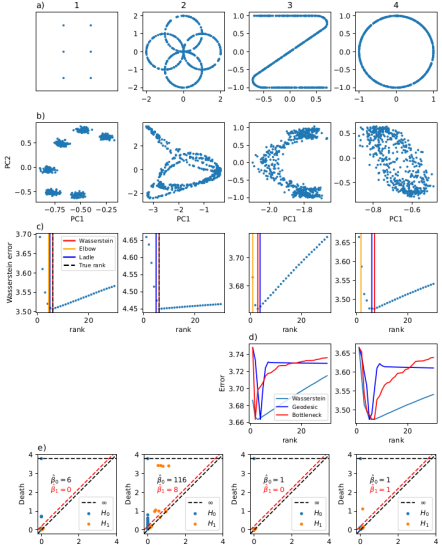
Split data into two halves. For each ρ, project first half onto top-ρ PC subspace, compute Wasserstein $\mathcal{W}_2$ to second half in $\mathbb{R}^p$. Choose

$$\widehat{r} = \arg \min_{1 \le \rho \le \rho_{\max}} \mathcal{W}_2\big(Y^{(1)}\Pi_\rho,\ Y^{(2)}\big).$$

where for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times d}$:

$$\mathcal{W}_2^2(\mathbf{A}, \mathbf{B}) := \min_{\pi} \frac{1}{m} \sum \|\mathbf{A}_i - \mathbf{B}_{\pi i}\|_2^2$$

Balances bias–variance under LMM.



Step 2: PCA Embedding

Compute $\zeta_i = \widehat{V}_r^\top Y_i$, rescale $p^{-1/2}$. Theorem 1 ensures ζ_i *approx.* $\phi(Z_i)$.

Step 3: Spherical Projection

When $f(z, z)$ varies, use extended model $Y_{ij} = \alpha_i X_j(Z_i) + \sigma E_{ij}$. After PC embedding, project

$$\zeta_i^{\mathrm{sp}} = \frac{\zeta_i}{\|\zeta_i\|},$$

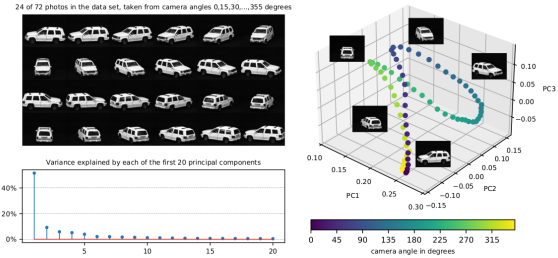
recovers $\phi(Z_i)$ up to scale.

Step 4: Nearest Neighbour Graph

Build ϵ-NN or k-NN graph on $\{\hat{\zeta}_i^{\mathrm{SP}}\}$ with weights $d_S(u, v) = \arccos(u, v)$. Compute:

- Geodesic estimates D_{ij} via shortest paths.
- Persistent homology of the graph.
- Minimum spanning tree for visualization.

Example 1: Car Images



- $n = 72$, $p = 110592$, angles $\theta_i = 0, 5, \ldots, 355°$.
- Selected $\widehat{r} = 11$. Persistent H0/H1: loop recovered.
- Shortest-path vs angular distance: linear (isometry up to scale).

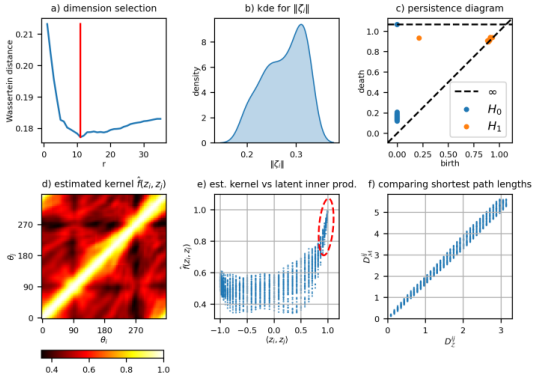


Figure 7: Images example. a) Wasserstein dimension selection; red vertical line indicates minimum at $\hat{r} = 11$. b) Kernel density estimate for the magnitudes of the PCA embedding vectors. c) Persistence diagram shows evidence of a single "loop" in the embedding. d) Estimated kernel as a function of latent positions in angular form $\theta_i = \arctan(z_i^{(2)}/z_i^{(1)})$. e) Estimated kernel as a function of latent inner product $\langle z_i, z_j \rangle$, the red dashed ellipse highlights $\hat{f}(z_i, z_j)$ in the region $\langle z_i, z_j \rangle \approx 1$. f) Evidence of a linear relationship between shortest path lengths computed from the nearest neighbour graph $\mathcal{G}$ (y-axis), and from the latent positions (x-axis).

Example 2: Planaria Transcriptomics

- $n = 5000$, $p = 5821$. Cell types hypothesized to form tree.
- $\widehat{r} \approx 14$. Nearest neighbour graph shows branching.
- Homology: $\beta_0 = 1$, $\beta_1 = 0$, consistent with tree.

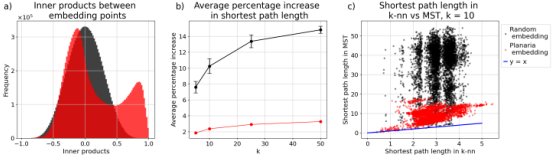
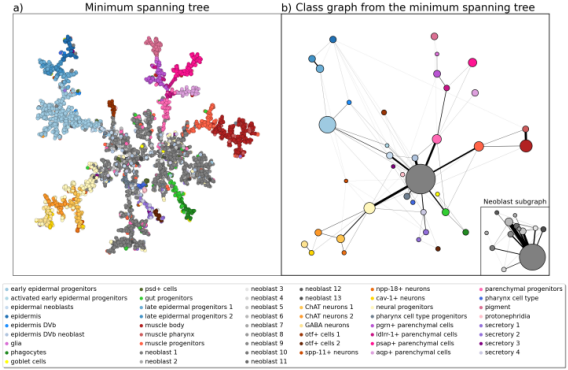


Figure 8: Single-cell transcriptomics example. a) histogram of inner products between distinct points in the PCA and random embeddings. b) average percentage increase in shortest path length in the minimum spanning tree compared to the k-nn graph, over different values of k. Results for the random embedding are shown in black, over 10 simulations with error bars indicated 2×standard error, c) comparing the shortest path lengths for samples in 10-nn graph and the MST.



Example 3: Temperature Time Series

- Daily temperature curves, n years, p time points.
- Hypothesis: latent domain is sphere (geographical location).
- LMM reveals seasonal loop and anomalies.

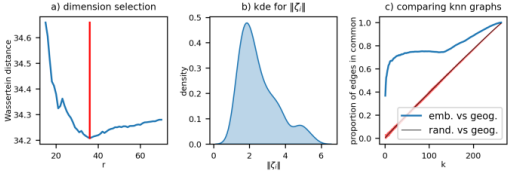


Figure 10: Temperatures example. a) Wasserstein dimension selection; red line indicates minimum at $\hat{r} = 36$. b) Kernel density estimate of the probability density of PC score magnitudes. c) The blue curve shows proportion of edges in common between embedding k-nn graph and geographic k-nn graph. The black line shows the mean proportion in common between the k-nn graph of a 100 uniformly random embeddings and the geographic k-nn graph. The red band indicates the range between maximum and minimum proportions across these 100 random embeddings.

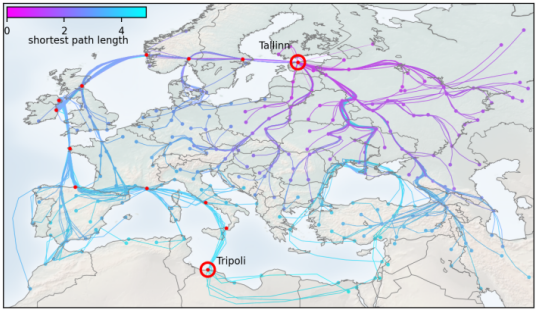


Figure 12: Temperatures example. Shortest paths in the embedding k-nn graph $\mathcal{G}$ from Tallinn, Estonia, to all other towns and cities. Each shortest path is visualized as a spline, with knot points given by the geographic locations of its constituent towns and cities. The red dots highlight the shortest path from Tallinn to Tripoli, Libya.

Special Case of LMM: IID Features

Suppose $X_j(z) = h_j(z)$ where h_j are i.i.d. samples from a Gaussian process with mean zero and covariance kernel f. Then:

- Mean correlation kernel is exactly $f(z, z')$.
- The feature map $\phi(z)$ recovers the GP kernel structure.
- PCA performs kernel PCA on f.

Special Case of LMM: Linear Features

Let $\mathcal{Z} \subset \mathbb{R}^d$ and $X_j(z) = w_j^\top z$ where $w_j \in \mathbb{R}^d$ are i.i.d. with mean zero and identity covariance.

- Then $f(z, z') = z^\top z'$, i.e., a linear kernel.
- Feature map: $\phi(z) = z$, the identity embedding.
- LMM recovers classical linear factor models (e.g., probabilistic PCA).

Special Case of LMM: Spiked Covariance Model

Let $X \in \mathbb{R}^{n \times p}$ be a data matrix with $\mathbb{E}[X^\top X]$ of rank r. Perform its eigendecomposition:

$$\frac{1}{n}\mathbb{E}[X^\top X] = V \Lambda V^\top,$$

with $V \in \mathbb{R}^{p \times r}$ orthonormal and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$. Define

$$Z = XV\Lambda^{-1/2}.$$

Then one checks:

$$\mathbb{E}[Z^\top Z] = nI_r, V^\top V = I_r,$$

$$X = Z\Lambda^{1/2}V^\top, \quad \text{a.s.},$$

where the last equality follows because

$$\mathbb{E}\|X - Z\Lambda^{1/2}V^\top\|_F^2 = \operatorname{tr}\mathbb{E}[(X - Z\Lambda^{1/2}V^\top)^\top(X - Z\Lambda^{1/2}V^\top)] = 0.$$

Adding isotropic noise E yields **spiked covariance model**

$$Y = Z\Lambda^{1/2}V^\top + \sigma E.$$

Special Case of LMM: Finite Mixture Model

Suppose $\mathcal{Z} = \{1, \ldots, K\}$ and Z_i indicate mixture components. Let means $m_1, \ldots, m_K \in \mathbb{R}^p$ and

$$X_j(z) = m_z^{(j)}$$

be the jth coordinate of m_z. Then:

- $f(z, z') = \frac{1}{p}\langle m_z, m_{z'} \rangle$, a rank-$K$ kernel.
- Data follow a mixture of K point clusters on the manifold $\{m_1, \ldots, m_K\}$.
- Nearest-neighbor graph recovers cluster structure.

Conclusions and Future Work

- LMM provides statistical basis for the Manifold Hypothesis.
- Homeomorphism and isometry connect latent and observed manifolds.
- PCA + Wasserstein + graph methods enable exploration.
- Future: non-Euclidean latent domains, faster algorithms, robustness analysis.

References

[I] Whiteley, Gray, Rubin-Delanchy (2025), arXiv:2208.11665v5

[II] Lawrence Cayton (2005), *Algorithms for manifold learning*