Statistical exploration of the Manifold Hypothesis

Nick Whiteley[1] Annie Gray[2] Patrick Rubin-Delanchy[3]

[1]Institute for Statistical Science, School of Mathematics, University of Bristol

[2]The Alan Turing Institute

[3]School of Mathematics, University of Edinburgh

Manifold structure in graph embeddings
Patrick Rubin-Delanchy
NeurIPS Spotlight, 2020

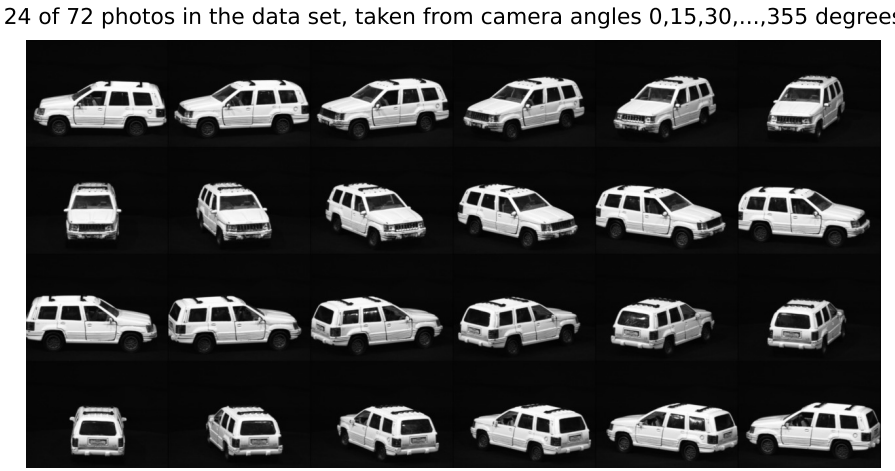
Statistical exploration of the Manifold Hypothesis
N.W., Annie Gray and Patrick Rubin-Delanchy
arXiv:2208.11665, **To appear as RSS discussion paper.**

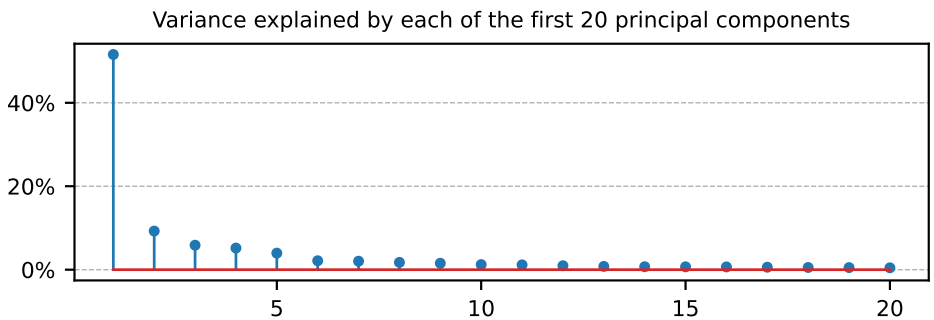
The Manifold Hypothesis

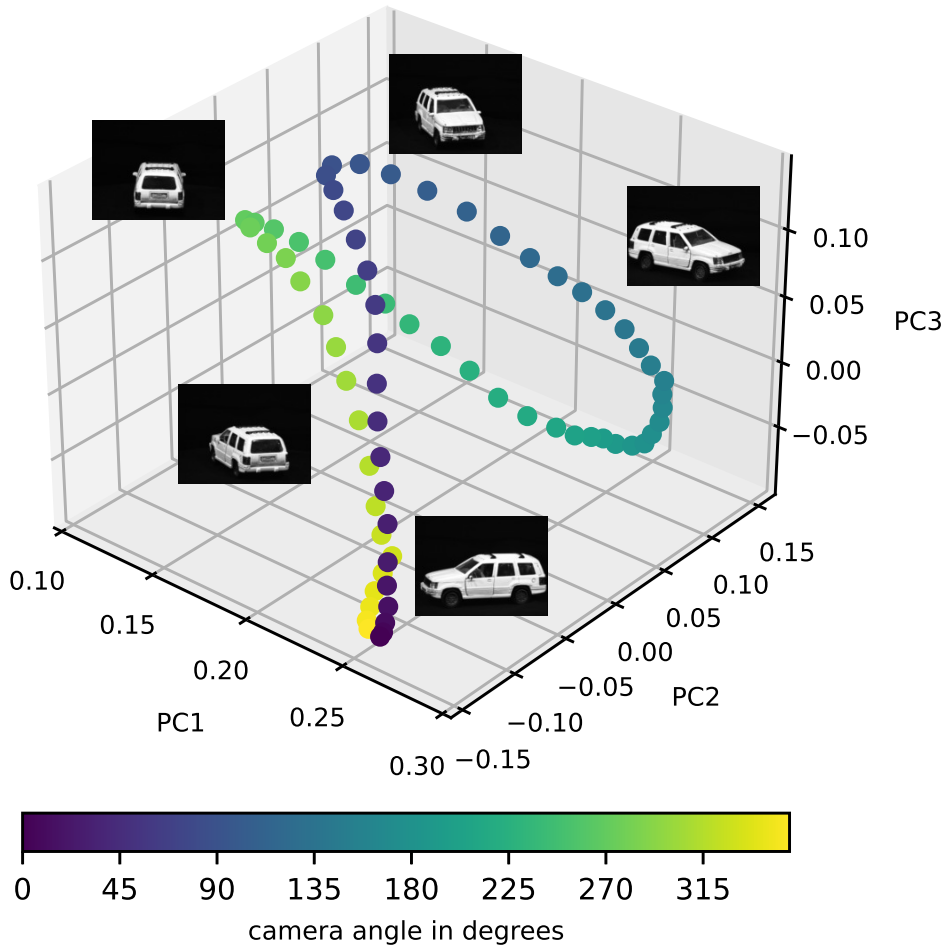
"*...the idea that the dimensionality of many data sets is only artificially high; though each data point consists of perhaps thousands of features, it may be described as a function of only a few underlying parameters. That is, the data points are actually samples from a low-dimensional manifold that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

Cayton, "Algorithms for manifold learning", UCSD Tech. Rep., 2005.

low dimensional structure in high dimensional data?



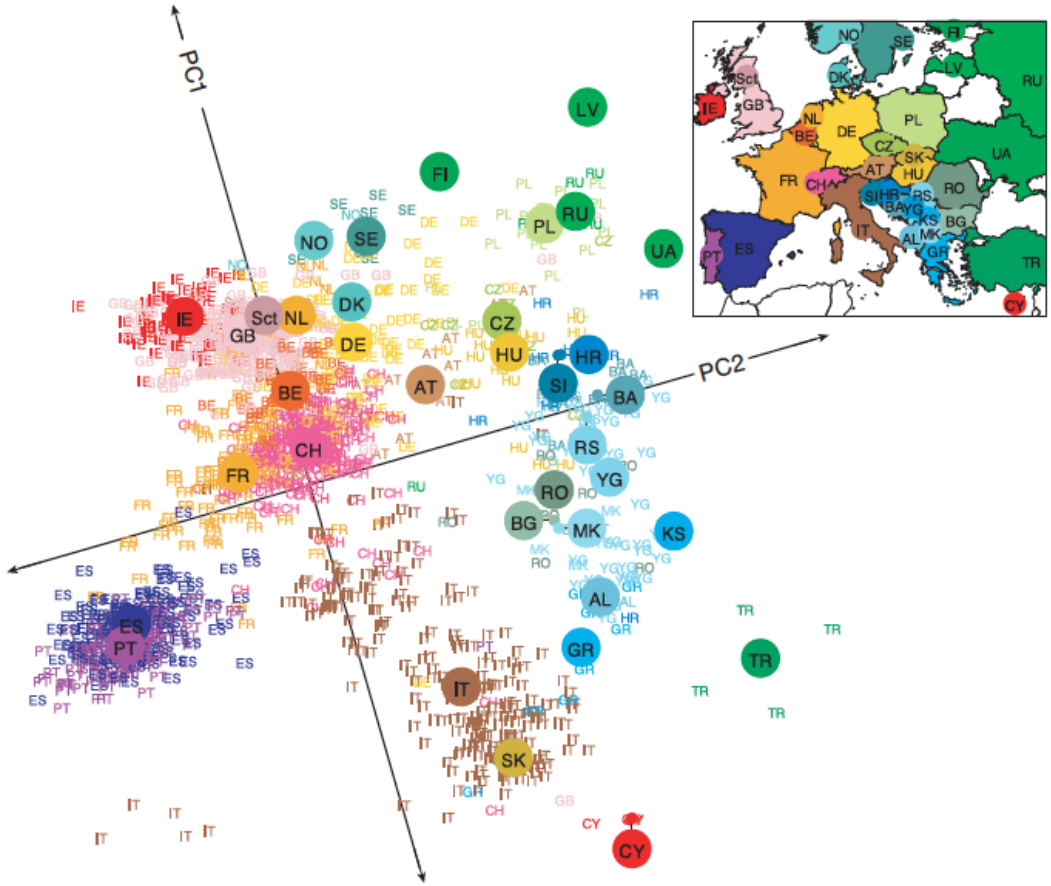




Data from the *Amsterdam library of object images*

- $n = 72$ grayscale photographs, taken from angles $0, 5, 10, \ldots, 355$ degrees.
- $p = 110592 = 384 \times 288$ pixels per photograph, reduced to 3 dimensions using PCA

low dimensional structure in high dimensional data?



Novembre et al. *"Genes mirror geography within Europe"*, Nature, 2008

- $n = 3000$ European individuals
- $p = 5 \times 10^5$ genotyped DNA sites per individual, reduced to 2 dimensions using PCA

exploring and exploiting manifold structure

- **Nonlinear dimension reduction**
  - attempt to "flatten" or "unfold" manifolds in order visualize in 2d or 3d
  - Laplacian Eigenmaps, Locally Linear Embedding, Isomap, t-SNE, U-MAP...
- **Regression and classification with covariates on manifolds**
  - nearest neighbour methods, locally linear regression, tree-based methods, ...
  - statistical performance driven by dimension of manifold, not ambient dimension
- **AI technologies and deep neural networks**
  - Nakada and Imaizumi. *Adaptive approximation and generalization of deep neural network with intrinsic dimensionality*, JMLR, 2020.
  - Huang, Wei and Chen. *Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality.* arXiv:2410.18784, 2024
  - and many others...

ML, Math and Stats perspectives

- Bengio et al. *"Representaton Learning..."*. IEEE PAMI, 2013.
- Fefferman et al. *"Testing the manifold hypothesis"*. J. Amer. Math. Soc, 2016.
- Wasserman. "*Topological Data Analysis*". Annu. Rev. Stat. Appl., 2018

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$

2. **random functions** $X_1, \ldots, X_p$, each acting $\mathcal{Z} \to \mathbb{R}$, so for each $z \in \mathcal{Z}$, $X_j(z)$ is a r.v.

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$

2. **random functions** $X_1, \ldots, X_p$, each acting $\mathcal{Z} \to \mathbb{R}$, so for each $z \in \mathcal{Z}$, $X_j(z)$ is a r.v.

3. **noise** matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ of zero mean, unit variance random variables

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$
   - hyp: $\mathcal{Z}$ is compact
2. **random functions** $X_1, \ldots, X_p$, each acting $\mathcal{Z} \to \mathbb{R}$, so for each $z \in \mathcal{Z}$, $X_j(z)$ is a r.v.
3. **noise** matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ of zero mean, unit variance random variables

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} \coloneqq X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$
   - hyp: $\mathcal{Z}$ is compact
2. **random functions** $X_1, \ldots, X_p$, each acting $\mathcal{Z} \to \mathbb{R}$, so for each $z \in \mathcal{Z}$, $X_j(z)$ is a r.v.
   - hyp: for each j, $z \to z' \Rightarrow \mathbb{E}[|X_j(z) - X_j(z')|^2] \to 0$.
3. **noise** matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ of zero mean, unit variance random variables

The Latent Metric Model

Definition: Latent Metric Model

data matrix: $\mathbf{Y} = [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{Y}_{ij} := X_j(Z_i) + \sigma \mathbf{E}_{ij}$$

comprising three independent sources of randomness:

1. **latent variables** $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mu$, where μ supported on metric space $(\mathcal{Z}, d)$
   - hyp: $\mathcal{Z}$ is compact
2. **random functions** $X_1, \ldots, X_p$, each acting $\mathcal{Z} \to \mathbb{R}$, so for each $z \in \mathcal{Z}$, $X_j(z)$ is a r.v.
   - hyp: for each j, $z \to z' \Rightarrow \mathbb{E}[|X_j(z) - X_j(z')|^2] \to 0$.
3. **noise** matrix $\mathbf{E} \in \mathbb{R}^{n \times p}$ of zero mean, unit variance random variables
   - hyp: $\mathbf{E}_{ij}$ are uncorrelated across i and independent across j

The Latent Metric Model

Definition: implict kernel and feature map

- **implicit kernel**

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

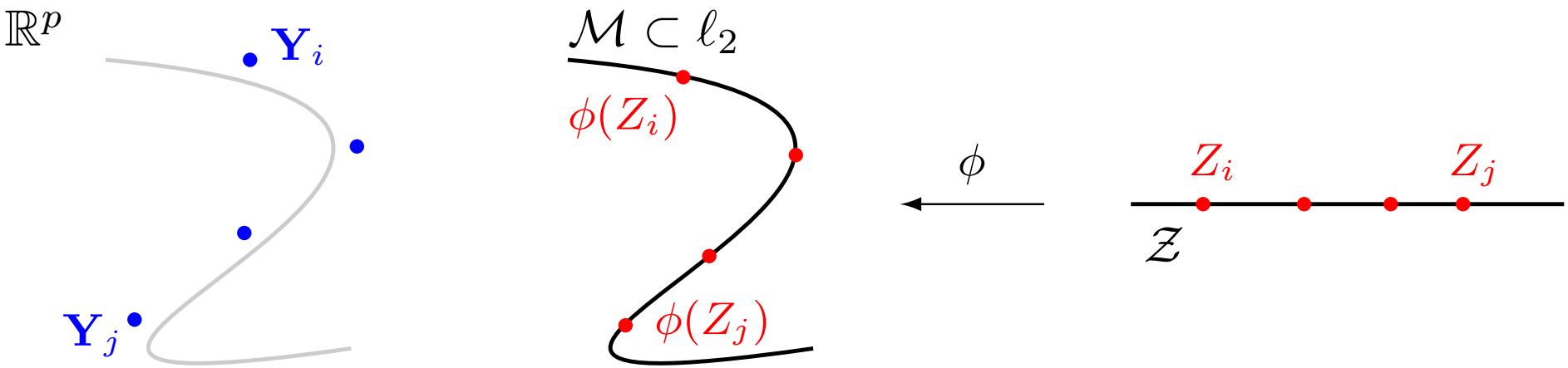
- by Mercer's theorem, $\exists$ **feature map** $\phi : \mathcal{Z} \to \ell_2$ s.t.

$$f(z, z') = \langle \phi(z), \phi(z') \rangle_{\ell_2}.$$

- **image** of ϕ,

$$\mathcal{M} := \{\phi(z); z \in \mathcal{Z}\} \subset \ell_2$$

sketch


$\mathbb{R}^p$
$\mathbf{Y}_i$
$\mathbf{Y}_j$
$\mathcal{M} \subset \ell_2$
$\phi(Z_i)$
$\phi(Z_j)$
ϕ
Z_i
Z_j
$\mathcal{Z}$

Intrinsic random projections

Proposition

Let $r \in \{1, 2, \dots, \} \cup \{\infty\}$ be the rank of the implicit kernel f and define the matrix $\mathbf{W} \in \mathbb{R}^{p \times r}$ with elements

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

where λ_k^f, u_k^f is the k-th eigenvalue/function pair assoc. with f, μ. Then:

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

where $\mathbf{I}_r$ is the identity matrix with r rows and columns.

Intrinsic random projections

Proposition

Let $r \in \{1, 2, \dots, \} \cup \{\infty\}$ be the rank of the implicit kernel f and define the matrix $\mathbf{W} \in \mathbb{R}^{p \times r}$ with elements

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

where λ_k^f, u_k^f is the k-th eigenvalue/function pair assoc. with f, μ. Then:

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

where $\mathbf{I}_r$ is the identity matrix with r rows and columns.

$$\begin{aligned}\frac{1}{p}\mathbb{E}[\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle | Z_i, Z_j] &= \langle \phi(Z_i), \mathbb{E}[\mathbf{W}^\top \mathbf{W}] \phi(Z_j) \rangle_{\ell_2} + 0 + 0 + \sigma^2 \frac{1}{p}\mathbb{E}[\langle \mathbf{E}_i, \mathbf{E}_j \rangle] \\ &= \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} + \sigma^2 \mathbf{I}[i = j].\end{aligned}$$

Intrinsic random projections

Proposition

Let $r \in \{1, 2, \dots, \} \cup \{\infty\}$ be the rank of the implicit kernel f and define the matrix $\mathbf{W} \in \mathbb{R}^{p \times r}$ with elements

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

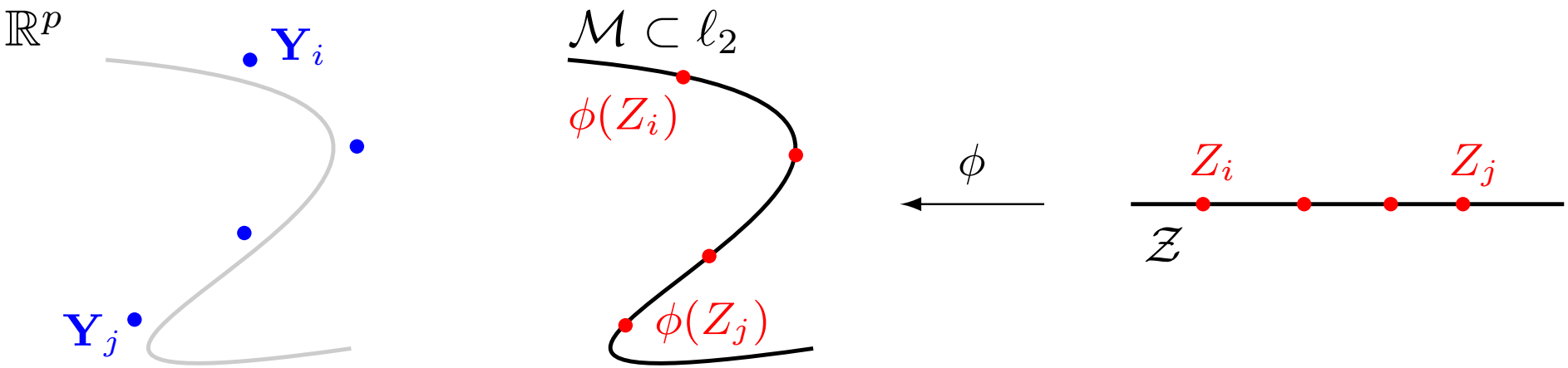
where λ_k^f, u_k^f is the k-th eigenvalue/function pair assoc. with f, μ. Then:

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

where $\mathbf{I}_r$ is the identity matrix with r rows and columns.

$$\begin{aligned}\frac{1}{p}\mathbb{E}[\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle | Z_i, Z_j] &= \langle \phi(Z_i), \mathbb{E}[\mathbf{W}^\top \mathbf{W}] \phi(Z_j) \rangle_{\ell_2} + 0 + 0 + \sigma^2 \frac{1}{p}\mathbb{E}[\langle \mathbf{E}_i, \mathbf{E}_j \rangle] \\ &= \langle \phi(Z_i), \phi(Z_j) \rangle_{\ell_2} + \sigma^2 \mathbf{I}[i = j].\end{aligned}$$

sketch


$\mathbb{R}^p$
$\mathbf{Y}_i$
$\mathbf{Y}_j$
$\mathcal{M} \subset \ell_2$
$\phi(Z_i)$
$\phi(Z_j)$
ϕ
Z_i
Z_j
$\mathcal{Z}$

Homeomorphism

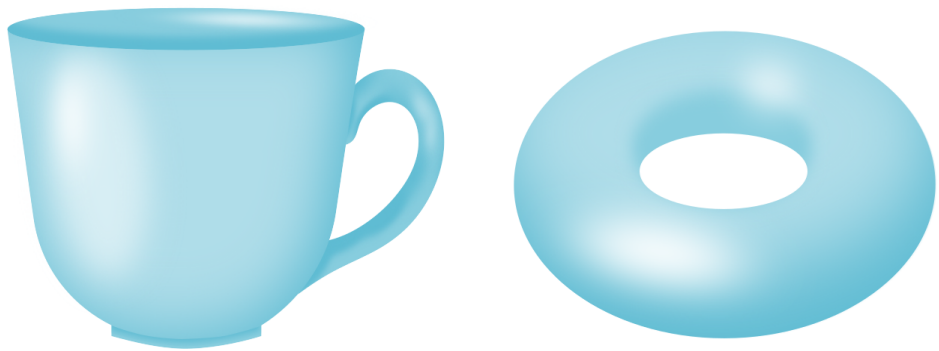
Definition: homeomorphism

A mapping between two metric spaces is a **homeomorphism** if it is continuous, one-to-one, and has a continuous inverse.

Homeomorphism

Definition: homeomorphism

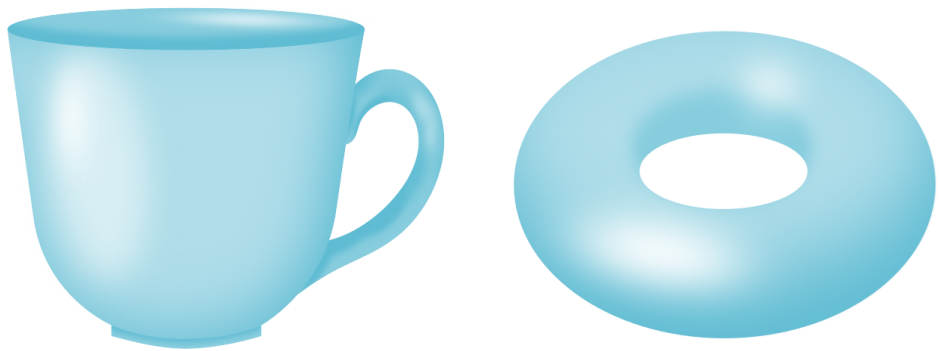
A mapping between two metric spaces is a **homeomorphism** if it is continuous, one-to-one, and has a continuous inverse.



Homeomorphism

Definition: homeomorphism

A mapping between two metric spaces is a **homeomorphism** if it is continuous, one-to-one, and has a continuous inverse.



Proposition

ϕ is a homeomorphism between $\mathcal{Z}$ and $\mathcal{M}$ if and only if

$$\forall z, z' \in \mathcal{Z}, \quad z \neq z' \quad \Longrightarrow \quad \sum_{j=1}^{p} \mathbb{E}\left[\left|X_j(z) - X_j(z')\right|^2\right] \neq 0.$$

Isometry

Definition: weak stationarity

Weak stationarity of any one of the random functions X_j means that:

- $\mathbb{E}[X_j(z)]$ is constant in z, and
- $\mathrm{Cov}[X_j(z), X_j(z')]$ depends only on distance $d(z, z')$

Isometry

Definition: weak stationarity

Weak stationarity of any one of the random functions X_j means that:

- $\mathbb{E}[X_j(z)]$ is constant in z, and
- $\mathrm{Cov}[X_j(z), X_j(z')]$ depends only on distance $d(z, z')$

...it follows that if $X_1, \ldots, X_p$ are **all weakly stationary**, then $f(z, z')$ **depends only on** $d(z, z')$.

Isometry

Definition: path and path length

- a **path** in $\mathcal{Z}$ is a continuous mapping $\eta : [a, b] \to \mathcal{Z}$ for some a, b.
- write $\gamma : [a, b] \to \mathcal{M}$ given by $\gamma_t := \phi(\eta_t)$ the **corresponding path** in $\mathcal{M}$,
- **lengths** of η and γ:

$$L(\eta) := \sup_{\mathcal{T}} \sum_{k=1}^{n} d(\eta_{t_k}, \eta_{t_{k-1}}), \quad L(\gamma) := \sup_{\mathcal{T}} \sum_{k=1}^{n} \|\gamma_{t_k} - \gamma_{t_{k-1}}\|_2,$$

the $\sup$ is over $n \geq 1$ and $\mathcal{T} = (t_0, \ldots, t_n)$ s.t. $t_0 = a \leq t_1 \leq \cdots \leq t_n = b$.

Isometry

Definition: path and path length

- a **path** in $\mathcal{Z}$ is a continuous mapping $\eta : [a, b] \to \mathcal{Z}$ for some a, b.
- write $\gamma : [a, b] \to \mathcal{M}$ given by $\gamma_t := \phi(\eta_t)$ the **corresponding path** in $\mathcal{M}$,
- **lengths** of η and γ:

$$L(\eta) := \sup_{\mathcal{T}} \sum_{k=1}^{n} d(\eta_{t_k}, \eta_{t_{k-1}}), \quad L(\gamma) := \sup_{\mathcal{T}} \sum_{k=1}^{n} \|\gamma_{t_k} - \gamma_{t_{k-1}}\|_2,$$

the sup is over $n \geq 1$ and $\mathcal{T} = (t_0, \ldots, t_n)$ s.t. $t_0 = a \leq t_1 \leq \cdots \leq t_n = b$.

Proposition

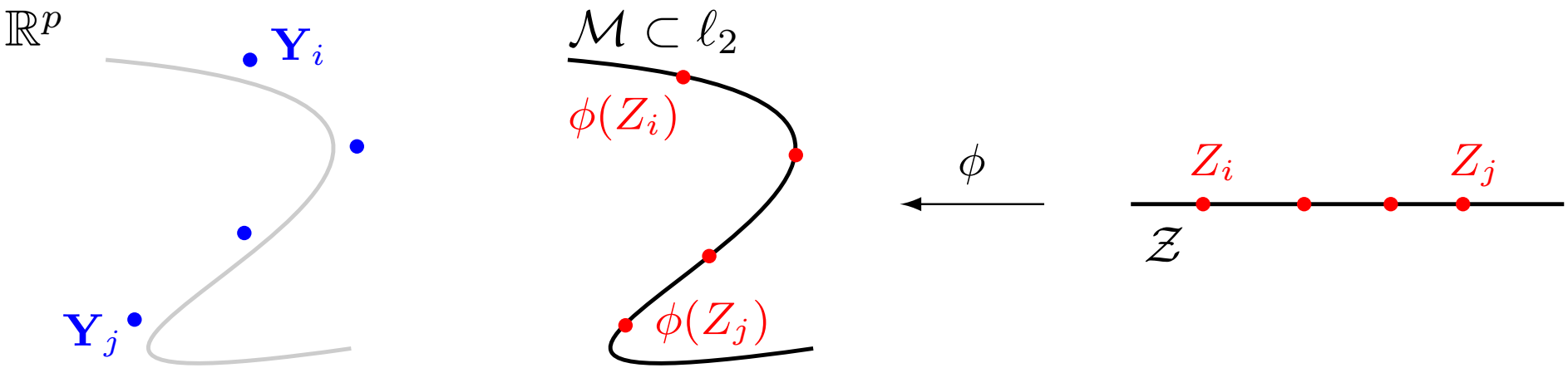
Assume there exists $\epsilon > 0$ and a C^2 function g such that $g'(0) < 0$ and

$$f(z, z') = g(d(z, z')^2) \quad \text{whenever} \quad d(z, z') \leq \epsilon.$$

Then for any path η in $\mathcal{Z}$ such that $L(\eta) < \infty$,

$$L(\gamma) = \sqrt{-2g'(0)} L(\eta).$$

sketch


$\mathbb{R}^p$
$\mathbf{Y}_i$
$\mathbf{Y}_j$
$\mathcal{M} \subset \ell_2$
$\phi(Z_i)$
$\phi(Z_j)$
ϕ
Z_i
Z_j
$\mathcal{Z}$

Uniform consistency of PCA embedding

Theorem

Let $\mathbf{Y} \sim$ Latent Metric Model and assume additionally that:

- $X_1, X_2, \ldots$ are mutually independent,
- $\sup_{j\geq 1} \sup_{z\in\mathcal{Z}} \mathbb{E}[|X_j(z)|^4] < \infty$, $\sup_{j\geq 1} \sup_{i\geq 1} \mathbb{E}[|\mathbf{E}_{ij}|^4] < \infty$,
- the kernel f has rank $r < \infty$, with positive eigenvalues bounded above and below uniformly in p.

Let $\zeta_1, \ldots, \zeta_n \in \mathbb{R}^r$ be the r-dimensional PCA embedding of $\mathbf{Y}$. Then there exists a random orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r\times r}$ depending on p, n such that:

$$\max_{i=1,\ldots,n} \left\| p^{-1/2}\mathbf{Q}\zeta_i - \phi(Z_i) \right\|_2 \in O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{n}{p}}\right), \quad \text{as} \quad p, n \to \infty.$$

some take home messages

... data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space."

Cayton, "Algorithms for manifold learning", UCSD Tech. Rep., 2005.

1. **manifold structure emerges from elementary statistical concepts**: latent variables, correlations, stationarity $\leftrightarrow$ homeomorphism, isometry, etc.

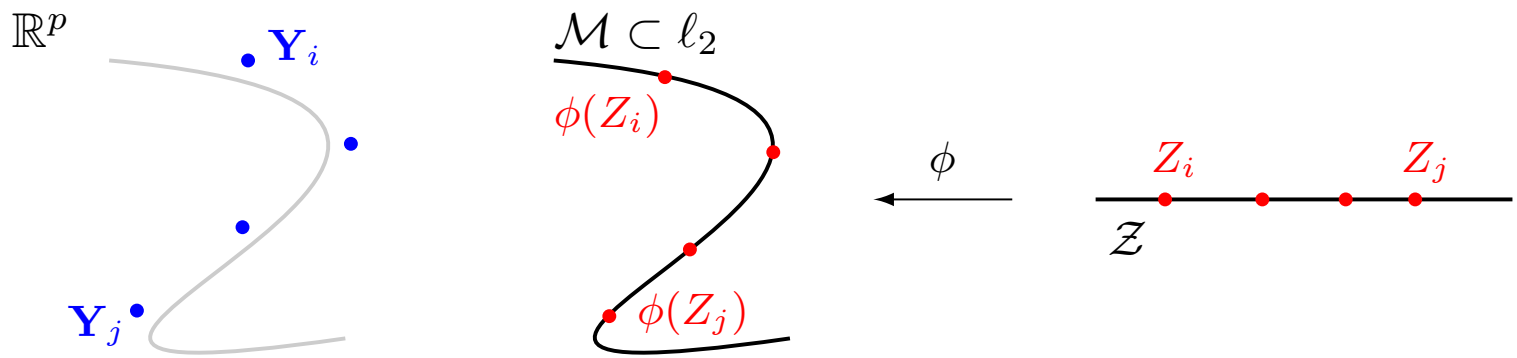
some take home messages

... data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space."

Cayton, "Algorithms for manifold learning", UCSD Tech. Rep., 2005.

1. **manifold structure emerges from elementary statistical concepts**: latent variables, correlations, stationarity $\leftrightarrow$ homeomorphism, isometry, etc.

2. perhaps **more nuanced** than traditional formulation of manifold hypothesis



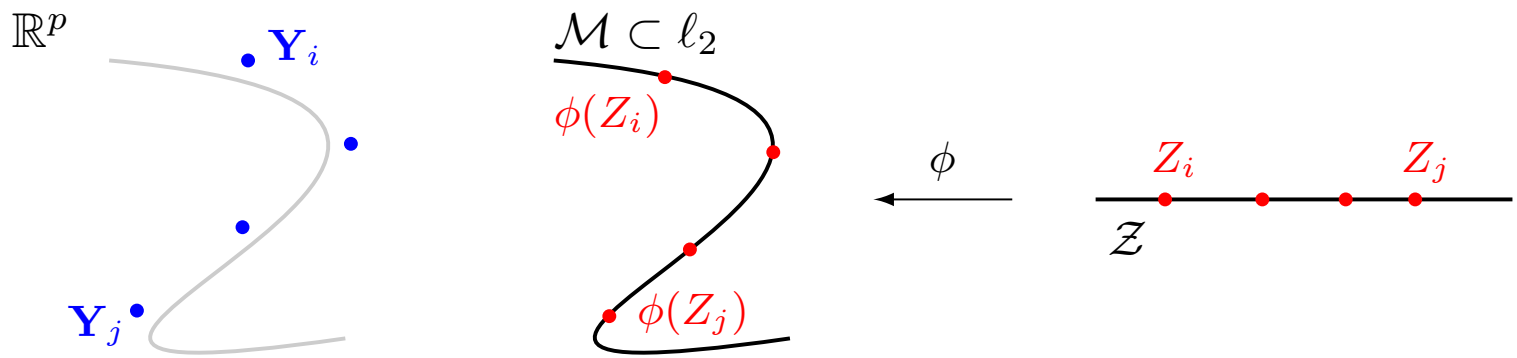
some take home messages

... data points are actually samples from a low-dimensional manifold that is embedded in a high-dimensional space."

Cayton, "Algorithms for manifold learning", UCSD Tech. Rep., 2005.

1. **manifold structure emerges from elementary statistical concepts**: latent variables, correlations, stationarity $\leftrightarrow$ homeomorphism, isometry, etc.

2. perhaps **more nuanced** than traditional formulation of manifold hypothesis



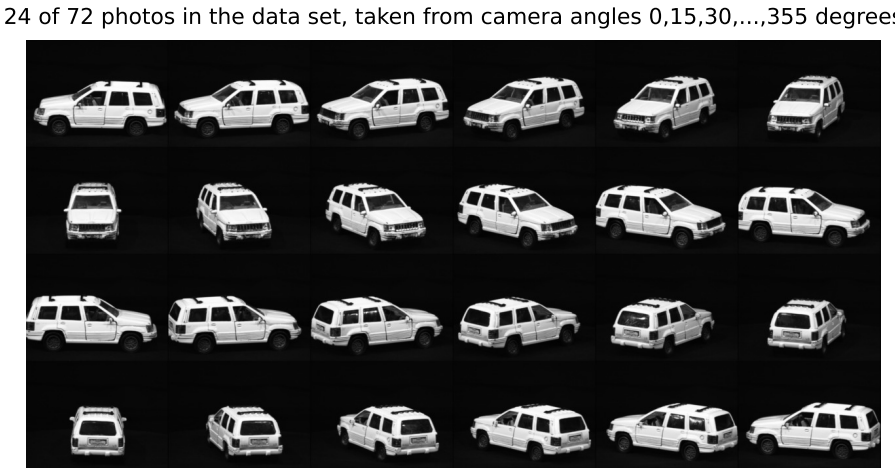
3. **rethink high-dimensional behaviour** of: PCA, spiked covariance models, Gaussian Process Latent Variables Model, Hierarchical Clustering, nonlinear dimension reduction, ...

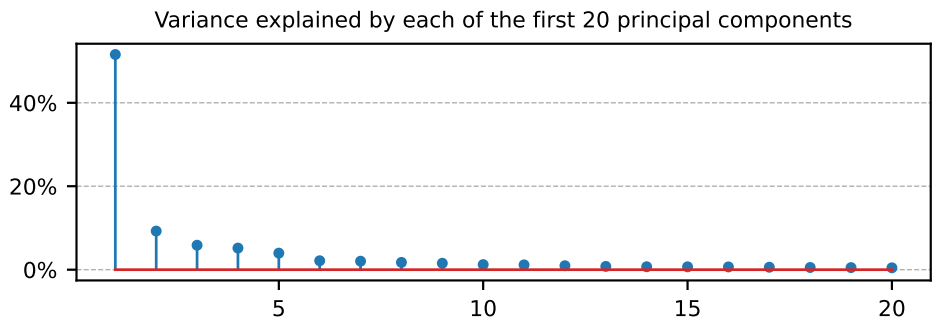
Statistical exploration of the Manifold Hypothesis
N.W., Annie Gray and Patrick Rubin-Delanchy
arXiv:2208.11665, **To appear as RSS discussion paper.**

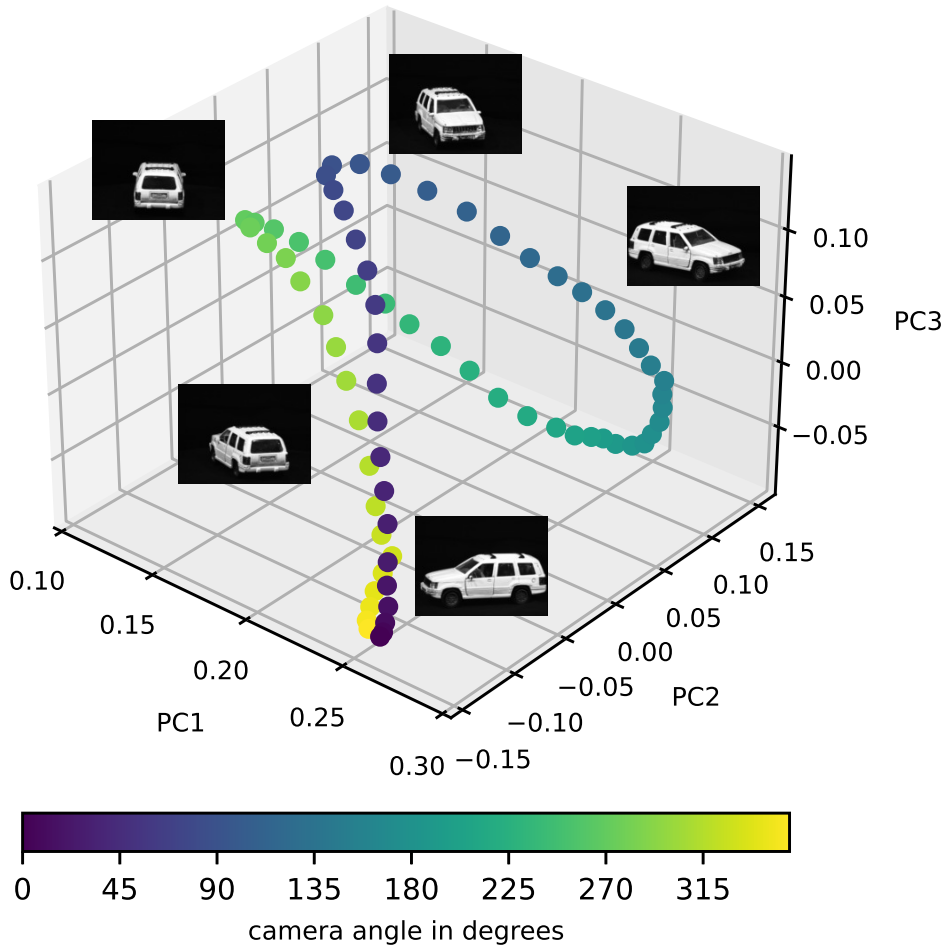
How high is 'high'? Rethinking the roles of dimensionality in topological data analysis and manifold learning.
Hannah Sansford, N.W., and Patrick Rubin-Delanchy
arxiv:2505.16879, 2025

The Origins of Representation Manifolds in Large Language Models.
Alex Modell, Patrick Rubin-Delanchy and N.W.,
arxiv.org:2505.18235, 2025

return to images example







Data from the *Amsterdam library of object images*

- $n = 72$ grayscale photographs, taken from angles $0, 5, 10, \ldots, 355$ degrees.
- $p = 110592 = 384 \times 288$ pixels per photograph, reduced to 3 dimensions using PCA

Linear dimension reduction by PCA

Definition: PCA embedding

- input: data matrix $\mathbf{Y} = [\mathbf{Y}_1 | \dots | \mathbf{Y}_n]^\top \in \mathbb{R}^{n \times p}$ and dimension $r \leq \min\{p, n\}$
- compute $v_1, \dots, v_r$ orthonorm. eig-vecs associated with r largest eig-vals of $\mathbf{Y}^\top \mathbf{Y} \in \mathbb{R}^{p \times p}$,
- the **PCA embedding** $\zeta_1, \dots, \zeta_n \in \mathbb{R}^r$ is:

$$\zeta_i := \mathbf{V}^\top \mathbf{Y}_i, \qquad \text{where} \qquad \mathbf{V} := [v_1 | \cdots | v_r]$$

return to images example

- **Hyp.1** = $\mathcal{Z}$ is a circle + homeomorphism
- **Hyp.2** = **Hyp.1** + $z_1, \ldots, z_n$ are equi-spaced around $\mathcal{Z}$ + isometry