Dimension reduction and manifold learning

Eddie Aamari

Département de mathématiques et applications

CNRS, ENS PSL

Master IASD / MATH — Dauphine PSL

A first high-dimensional dataset: COIL-20 (1996)

Columbia Object Image Library "COIL-20" (1996)

- Database size $n = 20$ objects $\times\ 72$ poses $= 1\,440$
- Image resolution $D = 128$ pixels $\times\ 128$ pixels $= 16\,384$

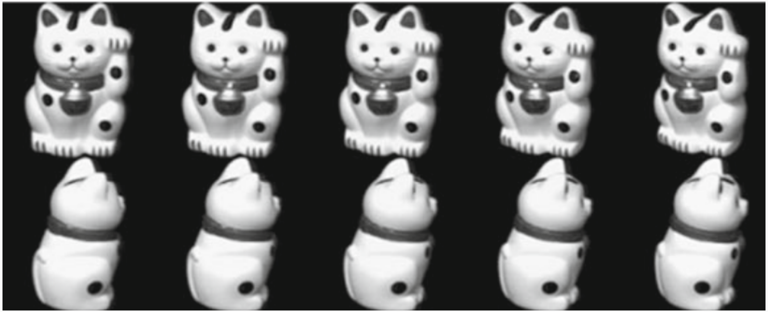


Figure 1: Pictures from the COIL20 dataset.

Synthetic dataset: COIL-20 (1996)

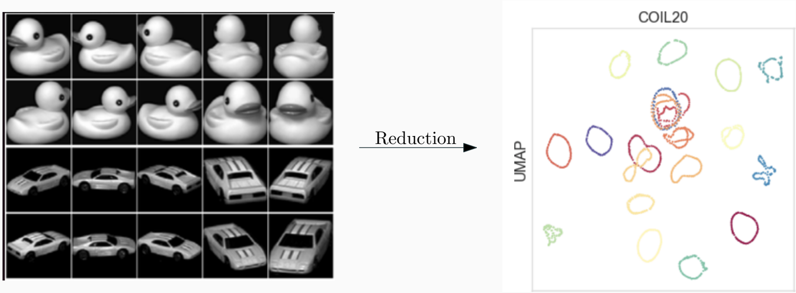


Figure 2: Low dimensional "representation" of the COIL20 dataset.

Synthetic dataset: COIL-20 (1996)

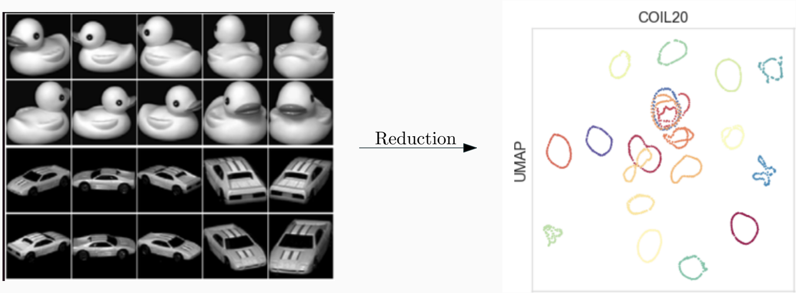


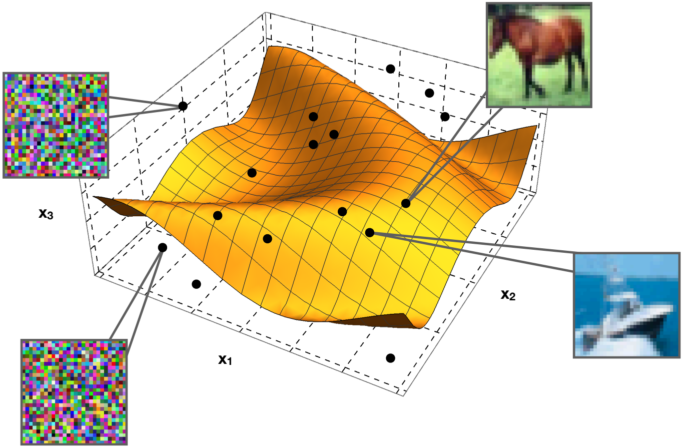
Figure 2: Low dimensional “representation” of the COIL20 dataset.

Manifold hypothesis $\equiv$ High-dim. datasets lie close to low-dim. geometric structures.

$\hookrightarrow$ models local non-linear local correlations within the data;

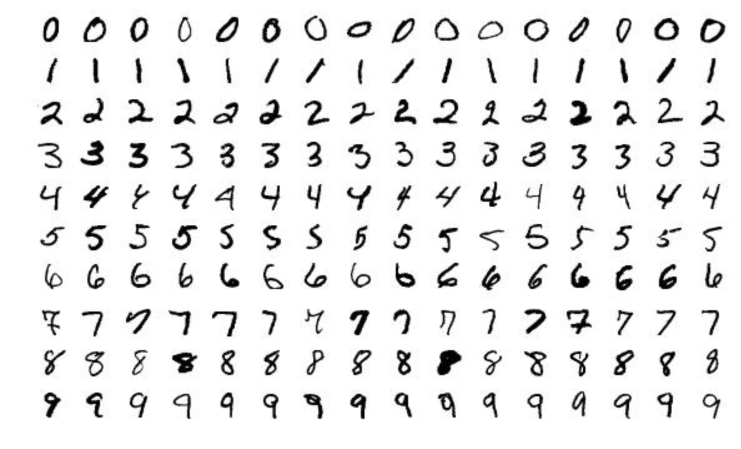
$\hookrightarrow$ is a sparsity assumption independent of coordinate systems.

Artist vision of the manifold hypothesis



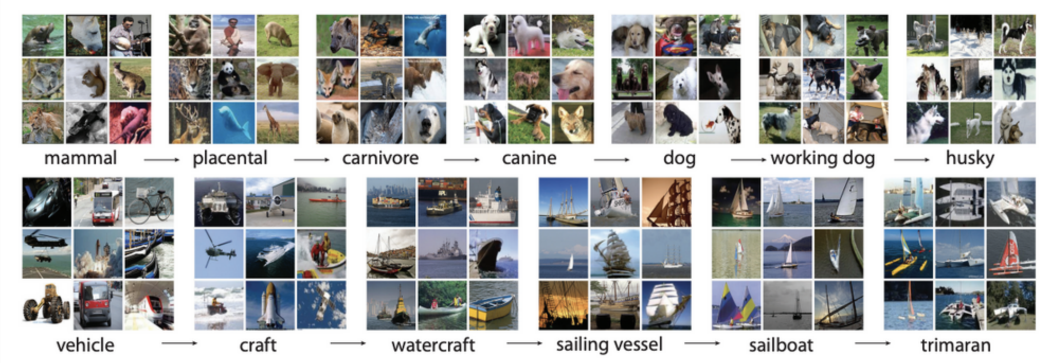
Less synthetic database : MNIST (1994)

- Database size $n = 60\,000$
- Image resolution $D = 784$



Real database : ImageNet (2010)

- Database size $n \simeq 14\,000\,000$
- Average image resolution $D \simeq 180\,000$



mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

High-dimensional data actually is intrinsically low-dimensional

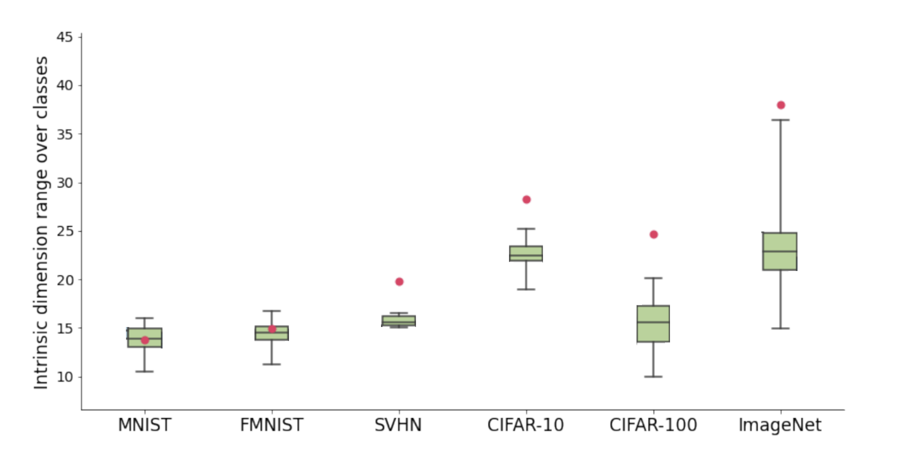
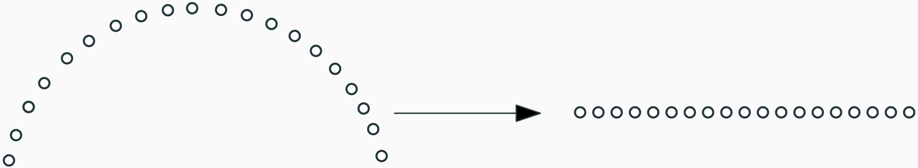


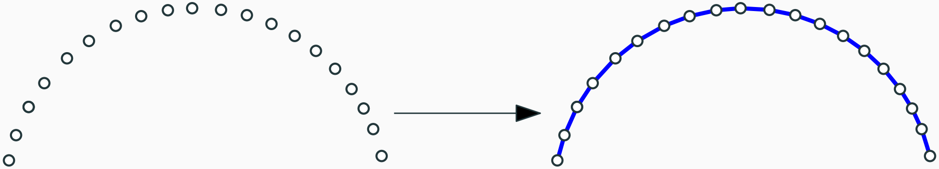
Figure 4: Boxplot of dimension estimates accross classes & dataset [Brown et al., 2023]

Dimension reduction



Dimensionality reduction (DR) refers to the problem of embedding a point set into a lower-dimensional space.

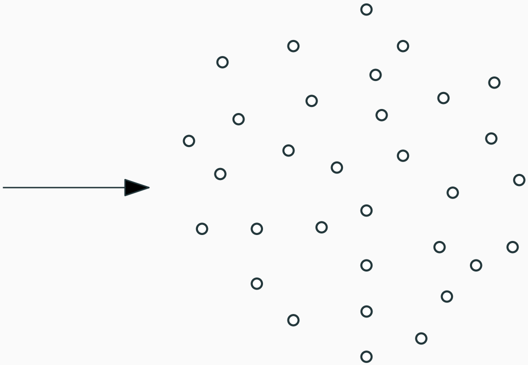
Manifold estimation



Manifold estimation refers to the problem of estimating the underlying (curved) low-dimensional space.

Multidimensional scaling

$$\begin{pmatrix} 0 & \delta_{1,2} & \cdots & \delta_{1,n} \\ \delta_{2,1} & 0 & \cdots & \delta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n,1} & \delta_{n,2} & \cdots & 0 \end{pmatrix}$$



Multidimensional scaling (MDS) is the term used in psychometry/psychology and statistics to refer to the problem of embedding a weighted graph into a Euclidean space.

Ill-posedness of dimension reduction

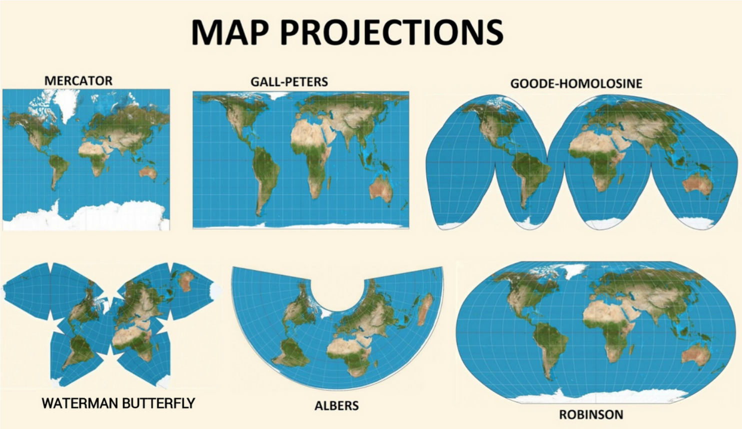


Figure 5: There exists no transformation of the sphere onto $\mathbb{R}^2$ that fully preserves distances.

Incredible variety of dimension reduction

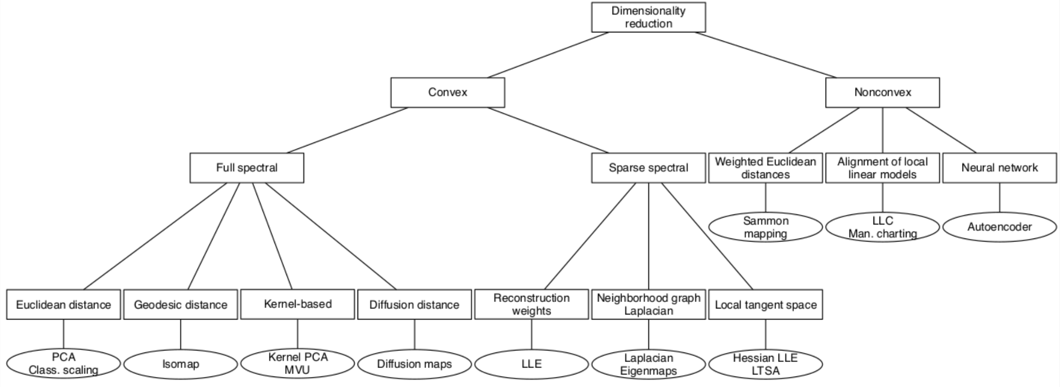


Figure 6: from Van Der Maaten, Postma, Herik, et al. 2009

Overview

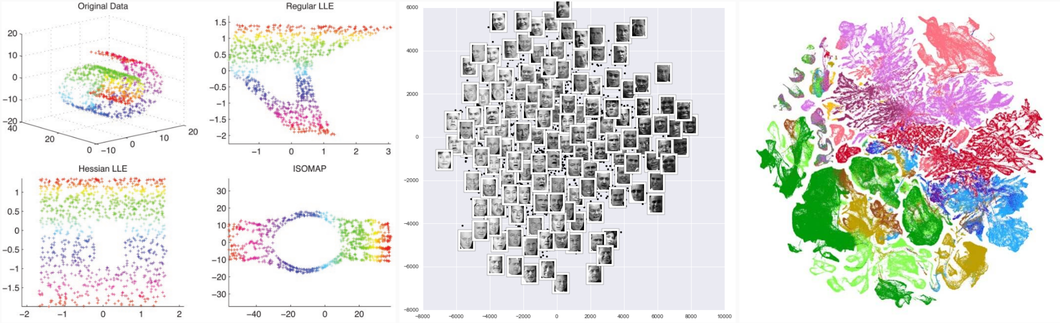


Figure 7: Visualizing complex simple / high-dimensional data in the plane.

(left) Toy 3D data (middle) Image data (right) Single-cell transcriptomics

Program

- **Goals**
  - Understand geometric phenomena in high-dimensional data
  - Get insights underlying the most common dimension reduction methods
  - Practice dimension reduction on toy and real data
  - Develop a critical approach to existing methods and design new ones

Program

- **Goals**
  - Understand geometric phenomena in high-dimensional data
  - Get insights underlying the most common dimension reduction methods
  - Practice dimension reduction on toy and real data
  - Develop a critical approach to existing methods and design new ones
- **Subjects covered (non-exhaustive)**
  - Manifold hypothesis
  - Multidimensional scaling, PCA, random projections
  - ISOMAP, Laplacian eigenmaps, kernel PCA, t-SNE, UMAP, ...

Program

- **Goals**
  - Understand geometric phenomena in high-dimensional data
  - Get insights underlying the most common dimension reduction methods
  - Practice dimension reduction on toy and real data
  - Develop a critical approach to existing methods and design new ones
- **Subjects covered (non-exhaustive)**
  - Manifold hypothesis
  - Multidimensional scaling, PCA, random projections
  - ISOMAP, Laplacian eigenmaps, kernel PCA, t-SNE, UMAP, ...
- **References**
  - Elements of dimensionality reduction and manifold learning. Ghojogh et al., 2023
  - Introduction to high-dimensional statistics. Giraud, 2021
  - Nonlinear dimensionality reduction. Lee, & Verleysen, 2007

Logistics

- **Format**
    - 7 × 3h class (**no class on January 14th!**)
    - Courses split between theory and practice
        - Lectures (blackboard / slides)
        - Hands-on sessions in Python (bring laptop!)

Logistics

- **Format**
  - 7 × 3h class (**no class on January 14th!**)
  - Courses split between theory and practice
    - Lectures (blackboard / slides)
    - Hands-on sessions in Python (bring laptop!)
- **Exam**
  - Oral presentation of a research article
  - 15mn autonomous + 5mn questions
  - Groups of 2 or 3 depending on the number of students

Logistics

- **Format**
  - 7 × 3h class (**no class on January 14th!**)
  - Courses split between theory and practice
    - Lectures (blackboard / slides)
    - Hands-on sessions in Python (bring laptop!)
- **Exam**
  - Oral presentation of a research article
  - 15mn autonomous + 5mn questions
  - Groups of 2 or 3 depending on the number of students

Quentions?