Advanced Data Science

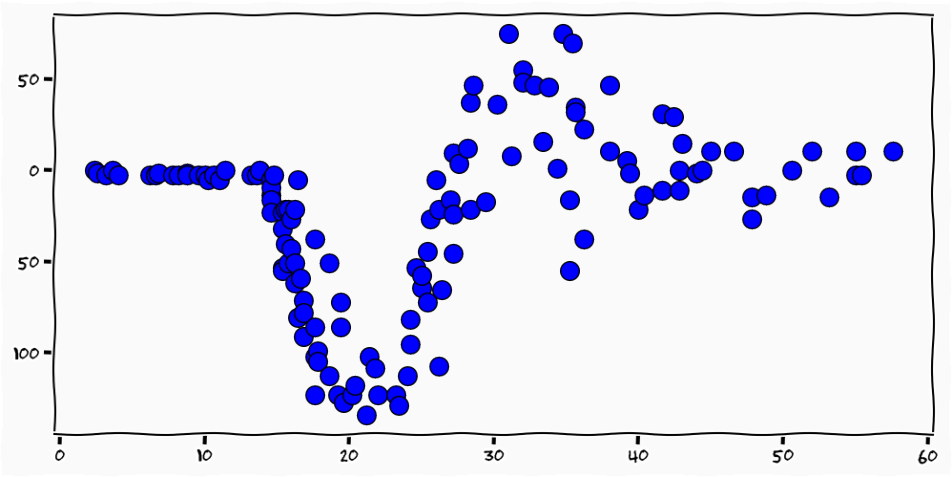
Lecture 7 : Visualisation

Carl Henrik Ek - che29@cam.ac.uk

24th of November, 2021

http://carlhenrik.com

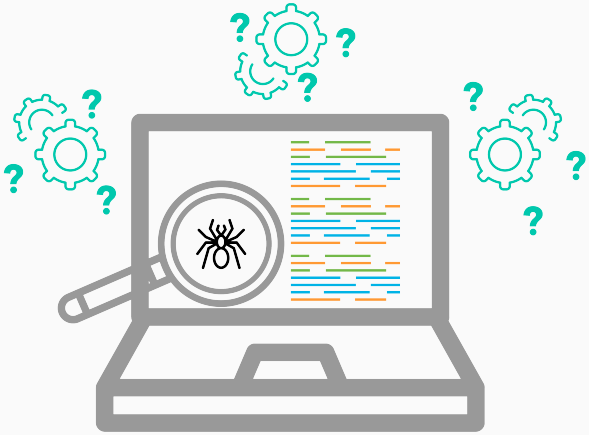
Motorcycle Data



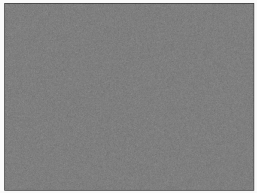
Tube Map


STOCKHOLM
T-BANA
1 Thorildsplan
2 Fridhemsplan
3 S:t Eriksplan
4 Odenplan
5 Rådmansgatan
6 Hötorget
Akalla
11
Husby
Mörby centrum
14
Danderyds sjukhus
Tensta
Hjulsta
10
Kista
Kymlinge
Rinkeby
Rissne
Hallonbergen
Bergshamra
Universitetet
Duvbo
Näckrosen
Hässelby gård
Johannelund
Sundbybergs centrum
Solna centrum
Tekniska högskolan
Vällingby
19
Vreten
13
Ropsten
Hässelby strand
Råcksta
Huvudsta
Stadion
5
Gärdet
Blackeberg
Västra skogen
4
6
Karlaplan
Islandstorget
Stadshagen
Östermalmstorg
Ängbyplan
17
18
3
Åkeshov
T-Centralen
Kungsträdgården
Brommaplan
Abrahamsberg
Stora mossen
Alvik
Kristine berg
1
2
Rådhuset
10
11
Gamla stan
Slussen
Hornstull
Liljeholmen
Medborgarplatsen
Aspudden
Zinkensdamm
Mariatorget
Skanstull
Örnsberg
Gullmarsplan
Bredäng
Midsommar-kransen
Globen
Skärmarbrink
Hammarbyhöjden
Mälar-höjden
Axels-berg
Telefonplan
Enskede gård
Björkhagen
Sätra
Hägerstensåsen
Sockenplan
Blåsut
Kärrtorp
Skärholmen
Västertorp
Svedmyra
Sandsborg
Bagarmossen
Fruängen
Stureby
Skogskyrko-gården
Skarpnäck
Vårberg
14
Bandhagen
17
Tallkrogen
Vårby gård
Hagsätra
Rågsved
Gubbängen
Norsborg
Hallunda
19
Högdalen
Hökarängen
Alby
Masmo
Farsta
13
Fittja
18
Farsta strand

Data Science is Debugging



High-Dimensional



Data

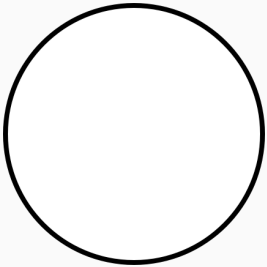
- Parametrisation of representation
  - how the data is given, often determined by measuring device

Data

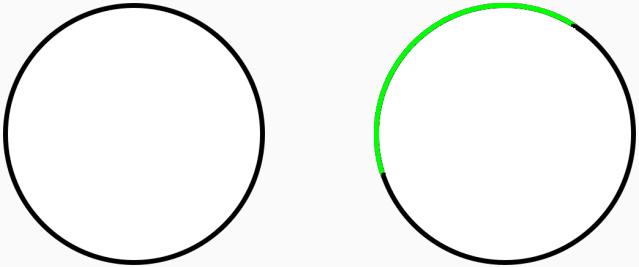
- Parametrisation of representation
  - how the data is given, often determined by measuring device
- Cannonical Representation of data
  - how the variations in the data can be parametrised

Definition (Manifold)
"a manifold is a topological space that near each point resembles Euclidean space"

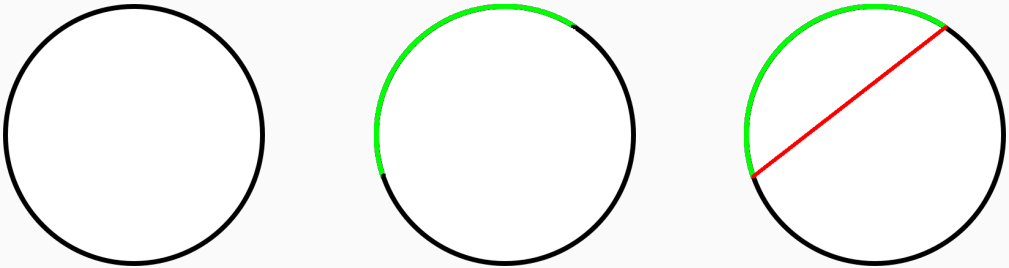
Manifold



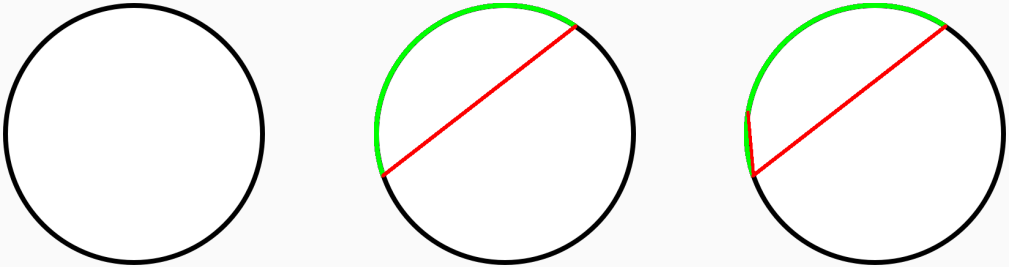
Manifold



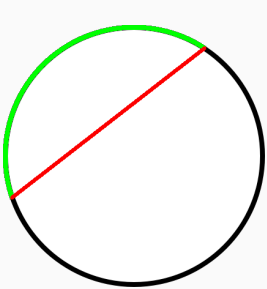
Manifold

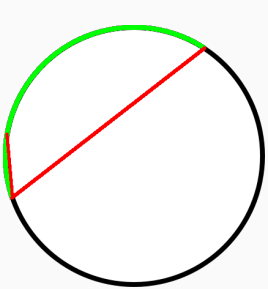


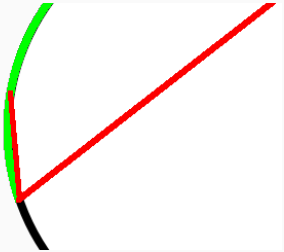
Manifold



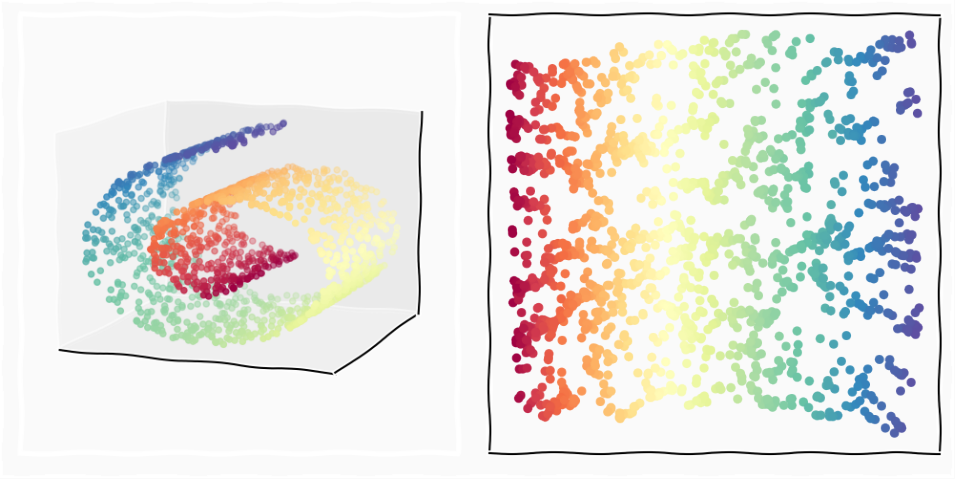
Manifold







Manifolds Learning



Today

- Can we reduce the dimensionality of data to visualise
- Can we "unravel" the manifold from data
- Problems associated with non-models

Dimensionality Reduction

Similar Matrices

$$
\begin{array}{ccccc}
 & & \mathbf{M}_A & & \\
 & \mathbf{x}_A & \rightarrow & \mathbf{M}_A\mathbf{x}_A & \\
\mathbf{P} & \downarrow & & \uparrow & \mathbf{P}^{-1} \\
 & \mathbf{x}_B & \rightarrow & \mathbf{M}_B\mathbf{x}_B & \\
 & & \mathbf{M}_B & &
\end{array}
$$

Similar Matrices

$$\mathbf{M}_B\mathbf{x}_B = \mathbf{P}\left(\mathbf{M}_A\mathbf{x}_A\right) = \mathbf{P}\left(\mathbf{M}_A\left(\mathbf{P}^{-1}\mathbf{x}_B\right)\right) = \left(\mathbf{P}\mathbf{M}_A\mathbf{P}^{-1}\right)\mathbf{x}_B$$
$$\mathbf{M}_B = \mathbf{P}\mathbf{M}_A\mathbf{P}^{-1},$$

- the maps are *similar* or $\mathbf{M}_A \sim \mathbf{M}_B$

Eigen-decomposition

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$$

$$\mathbf{\Lambda} = \begin{cases} 0 & i \neq j \\ \lambda_i & i = j \end{cases}$$

$$\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I} \quad \Rightarrow \quad \mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}}.$$

Spectral Theorem

$$\mathbf{M} = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}.$$

- the eigen decomposition means we can write a matrix as a sum of rank one matrices
- all symmetric real matrices have a diagonal matrix that they are similar to

Rank-Nullity Theorem

$$\mathrm{Rank}(T) + \mathsf{Nullity}(T) = \mathsf{dim}(A)$$

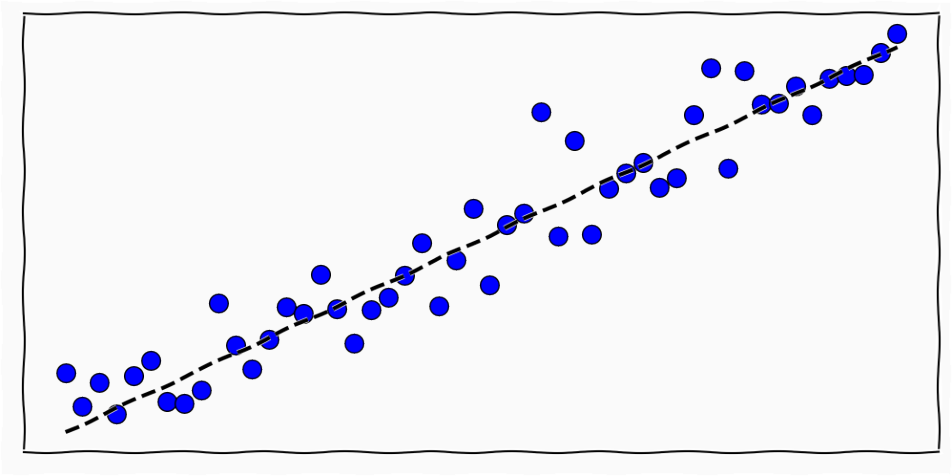
- $T : A \to B$ is a map between two vector spaces
- $\mathrm{Rank}(T)$ is the dimensionality of the *image* of T
- $\mathrm{Nullity}(T)$ is the dimensionality of the *kernel* of T

Rank-Nullity Theorem

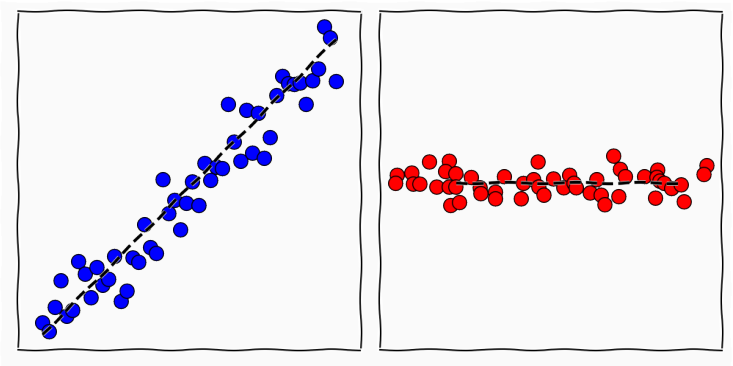
$$\mathrm{Rank}(T) + \mathsf{Nullity}(T) = \mathsf{dim}(A)$$

Task Can we find a map T such that kernel of the map is the subspace where the data have no variations?

Rank-Nullity

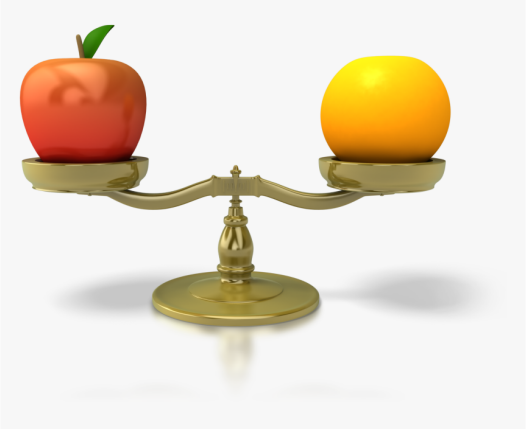


Principal Component Analysis

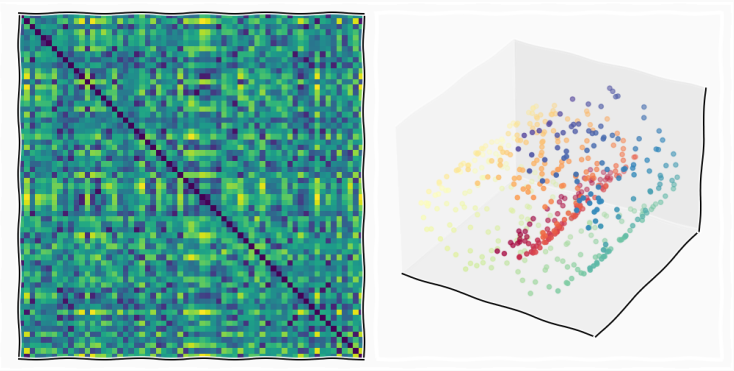


$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$$

Proximity/Dissimilarity



Multi Dimensional Scaling [Cox et al., 2008]



- Given a similarity matrix $\mathbf{\Delta}$ can we find a vectorial representation such that,

$$\mathbf{y}_i^{\mathrm{T}} \mathbf{y}_j = \mathbf{\Delta}_{ij}$$

Multi Dimensional Scaling

$$\mathbf{\Delta} = \begin{bmatrix} \delta_{00} & \delta_{01} & \cdots & \delta_{0N} \\ \delta_{10} & \delta_{11} & \cdots & \delta_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{N0} & \delta_{N1} & \cdots & \delta_{NN} \end{bmatrix}$$

$$\mathbf{D}_{ij}^2 = d_{ij}^2 = \sum_{k=1}^{d} (y_{ki} - y_{kj})^2 = \mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i + \mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j - 2\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j$$

$$\mathbf{G}_{ij} = g_{ij} = \mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j$$

$$d_{ij}^2 = g_{ii} + g_{jj} - 2g_{ij}$$

$$g_{ij} = \frac{1}{2}(g_{ii} + g_{jj} - d_{ij}^2)$$

- if we assume that the data is centred we can write the Gram matrix as a function of the distance matrix

- MDS Objective,

$$\hat{\mathbf{Y}} = \text{argmin}_{\mathbf{Y}} \|\mathbf{D} - \mathbf{\Delta}\|_F.$$

- MDS Objective,

$$\hat{\mathbf{Y}} = \text{argmin}_{\mathbf{Y}} \|\mathbf{D} - \mathbf{\Delta}\|_F.$$

- Element-Wise Matrix norm,

$$\|\mathbf{M}\|_{p,q} = \left(\sum_{j=1}^{n} \left(\sum_{i=1}^{m} |m_{ij}|^p \right)^{\frac{p}{q}} \right)^{\frac{1}{q}}$$

$$\|\mathbf{M}\|_F = \sqrt{\text{trace}(\mathbf{M}^{\text{T}}\mathbf{M})} = \sqrt{\text{trace}(\mathbf{M}^2)}.$$

- Any proximity matrix will be square and symmetric
- The spectral theorem says that it is therefore *similar* to a diagonal matrix
- The Frobenious norm of a diagonal matrix is just square-root of the *trace* of the square of the *eigenvalues*

$$\text{argmin}_{\mathbf{D}} \|\mathbf{D} - \mathbf{\Delta}\|_F^2 = \text{argmin}_{\mathbf{D}} \text{trace}\,(\mathbf{D} - \mathbf{\Delta})^2$$

Multi Dimensional Scaling

$$\begin{aligned}\operatorname{argmin}_{\mathbf{D}}\|\mathbf{D}-\boldsymbol{\Delta}\|_F^2 &= \operatorname{argmin}_{\mathbf{D}}\operatorname{trace}\left(\mathbf{D}-\boldsymbol{\Delta}\right)^2 \\ &= \operatorname{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}}\operatorname{trace}\left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}}-\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)^2\end{aligned}$$

$$
\begin{aligned}
\operatorname{argmin}_{\mathbf{D}} \|\mathbf{D} - \boldsymbol{\Delta}\|_F^2 &= \operatorname{argmin}_{\mathbf{D}} \operatorname{trace} \left(\mathbf{D} - \boldsymbol{\Delta}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q}, \hat{\boldsymbol{\Lambda}}} \operatorname{trace} \left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q}, \hat{\boldsymbol{\Lambda}}} \operatorname{trace} \left(\mathbf{V}^{\mathrm{T}} \left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right) \mathbf{V}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\text{argmin}_{\mathbf{D}} \|\mathbf{D} - \boldsymbol{\Delta}\|_F^2 &= \text{argmin}_{\mathbf{D}} \text{trace}\left(\mathbf{D} - \boldsymbol{\Delta}\right)^2 \\
&= \text{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \text{trace}\left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)^2 \\
&= \text{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \text{trace}\left(\mathbf{V}^{\mathrm{T}}\left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)\mathbf{V}\right)^2 \\
&= \text{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \text{trace}\left(\mathbf{V}^{\mathrm{T}}\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}}\mathbf{V} - \mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{argmin}_{\mathbf{D}} \|\mathbf{D}-\boldsymbol{\Delta}\|_F^2 &= \operatorname{argmin}_{\mathbf{D}} \operatorname{trace}\left(\mathbf{D}-\boldsymbol{\Delta}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \operatorname{trace}\left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \operatorname{trace}\left(\mathbf{V}^{\mathrm{T}}\left(\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}} - \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\right)\mathbf{V}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \operatorname{trace}\left(\mathbf{V}^{\mathrm{T}}\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}}\mathbf{V} - \mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\mathbf{V}\right)^2 \\
&= \operatorname{argmin}_{\mathbf{Q},\hat{\boldsymbol{\Lambda}}} \operatorname{trace}\left(\mathbf{V}^{\mathrm{T}}\mathbf{Q}\hat{\boldsymbol{\Lambda}}\mathbf{Q}^{\mathrm{T}}\mathbf{V} - \boldsymbol{\Lambda}\right)^2.
\end{aligned}
$$

$$\mathbf{D} = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}},$$

$$\|\mathbf{D} - \mathbf{\Delta}\|_F = \sqrt{\sum_{i=d+1}^{N} \lambda_i^2}$$

- To get the best d dimensional solution we pick the top d eigenvalues

$$\mathbf{D} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$$

$$\begin{aligned}\mathbf{D} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} &= \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ &= \left(\mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^{\mathrm{T}}\right)\end{aligned}$$

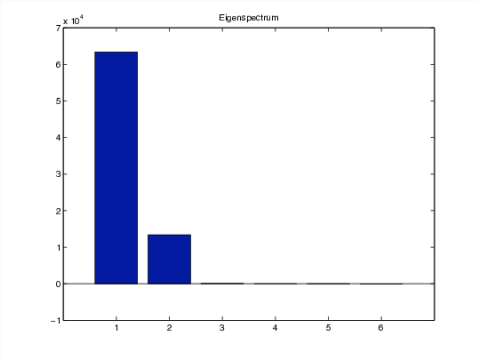
$$\begin{aligned}\mathbf{D} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} &= \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \\ &= \left(\mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)\left(\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^{\mathrm{T}}\right) \\ &= \left(\mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)\left(\mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{\mathrm{T}}\end{aligned}$$

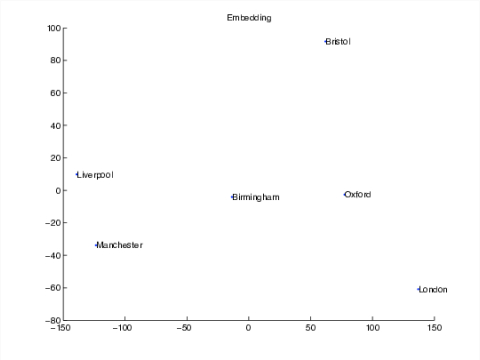
$$
\begin{aligned}
\mathbf{D} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}} &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \\
&= \left(\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\right)\left(\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^{\mathrm{T}}\right) \\
&= \left(\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\right)\left(\mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\right)^{\mathrm{T}} \\
&\Rightarrow \mathbf{Y} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}
\end{aligned}
$$

Example

	Man	Ox	Lon	Bri	Liv	Birm
Man	0	203	262	224	46	114
Ox	203	0	83	95	217	91
Lon	262	83	0	170	285	161
Bri	224	95	170	0	217	122
Liv	46	217	285	217	0	126
Birm	114	91	161	122	126	0

Example


Eigenspectrum
x 10^4
7
6
5
4
3
2
1
0
−1
1
2
3
4
5
6


Embedding
100
80
60
40
20
0
−20
−40
−60
−80
−150
−100
−50
0
50
100
150
Bristol
Liverpool
Birmingham
Oxford
Manchester
London

PCA Equvivalence [1]

- In MDS we diagonalise a $N \times N$ matrix

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$$

[1] see attached notes

PCA Equvivalence [1]

- In MDS we diagonalise a $N \times N$ matrix

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$$

- In PCA we diagonalise a $D \times D$ matrix

$$\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$$

[1] see attached notes

PCA Equvivalence [1]

- In MDS we diagonalise a $N \times N$ matrix

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$$

- In PCA we diagonalise a $D \times D$ matrix

$$\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$$

- Rank

$$\text{Rank}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right) = \text{Rank}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right).$$

[1] see attached notes

Proximity Graph

- We have a method to find a geometrical embedding from a similarity relationship

Proximity Graph

- We have a method to find a geometrical embedding from a similarity relationship
- *a manifold is a topological space that near each point resembles Euclidean space*

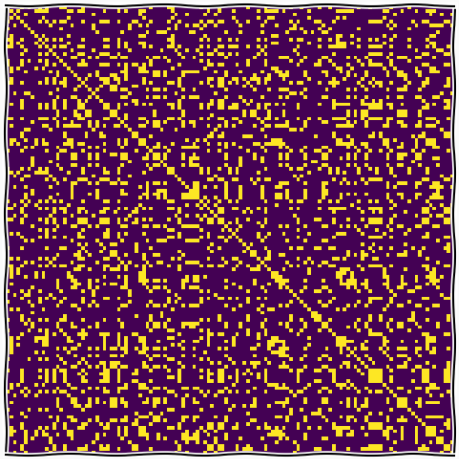
Proximity Graph

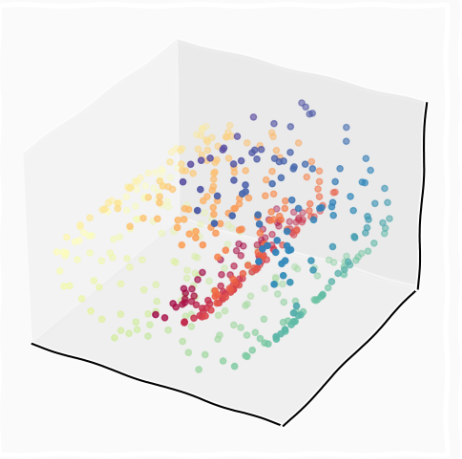
- We have a method to find a geometrical embedding from a similarity relationship
- *a manifold is a topological space that near each point resembles Euclidean space*
- $\Rightarrow$ we can *measure* local distances faithfully

Proximity Graph

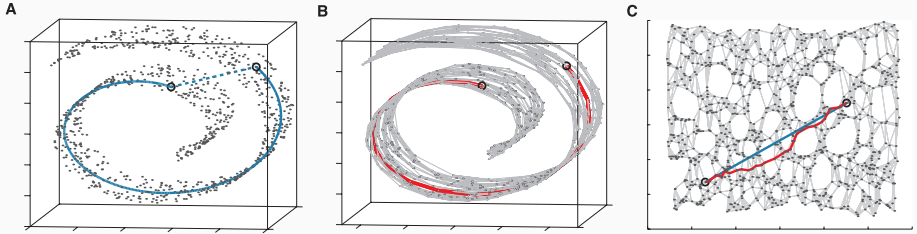
- We have a method to find a geometrical embedding from a similarity relationship
- *a manifold is a topological space that near each point resembles Euclidean space*
- $\Rightarrow$ we can *measure* local distances faithfully
- Learning manifold implies completing similarity relationship

Learning Manifold



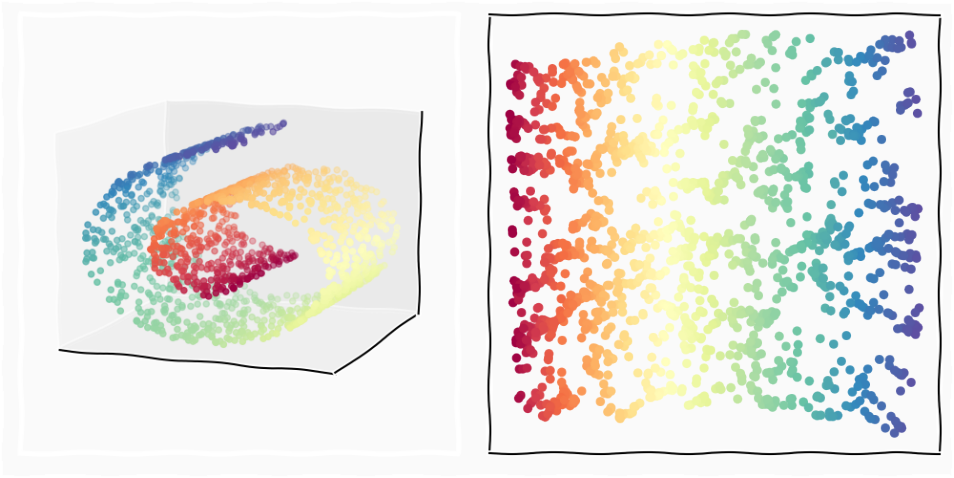


Isomap [Tenenbaum et al., 2000]



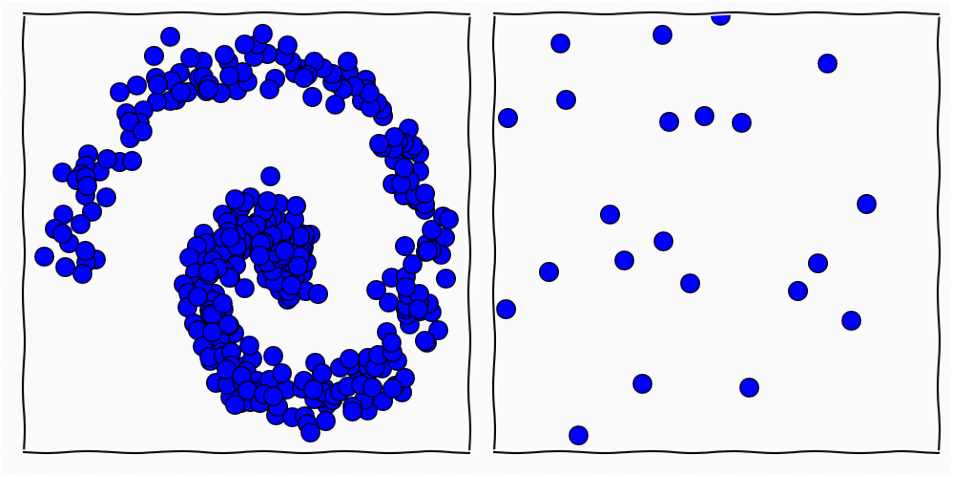
1. Compute local similarity
2. Compute shortest path in graph
3. Apply MDS

Manifolds Learning

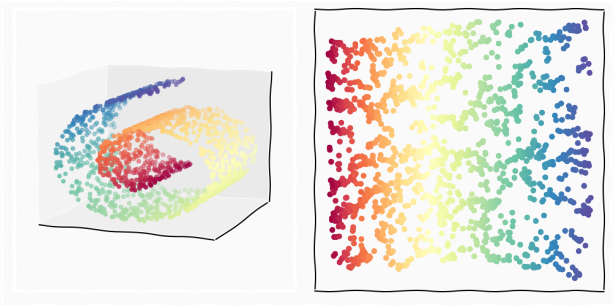


Models for Dimensionality Reduction

Locality

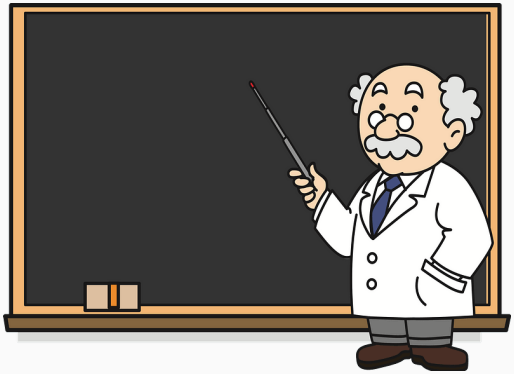


Generative Model



$$\mathbf{y}_i = f(\mathbf{x}_i)$$

Task



General Model

$$p(f, \mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid f, \mathbf{X})p(f)p(\mathbf{X})}{p(\mathbf{Y})}$$

$$p(\mathbf{Y}) = \int p(\mathbf{Y} \mid f, \mathbf{X})p(f)p(\mathbf{X})\mathrm{d}f\mathrm{d}\mathbf{X}$$

- To get the posterior we need to marginalise out our belief/preference in
  - mapping
  - latent representation
- Generally this integration will be intractable

$$p(\mathbf{Y}, \mathbf{W}, \mathbf{X}) = p(\mathbf{Y}|\mathbf{W}, \mathbf{X})p(\mathbf{X})p(\mathbf{W})$$
$$p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) = \mathcal{N}(\mathbf{X}\mathbf{W} + \mu, \beta^{-1}\mathbf{I}),$$

- we assume the data is corrupted by Gaussian noise we get a likelihood
- we assume the mapping to be linear such that $\mathbf{Y} = \mathbf{X}\mathbf{W}$

Inference Task

$$p(\mathbf{W}, \mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{Y}, \mathbf{W}, \mathbf{X})}{p(\mathbf{Y})}$$

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) p(\mathbf{X}) p(\mathbf{W})$$

- Assume both priors to be Gaussian
- Even in the linear case the integral is intractable

PCA [Tipping et al., 1999]

$$p(\mathbf{y}, \mathbf{x}|\mathbf{W}) = \mathcal{N}\left(\begin{bmatrix} \mu \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{W}\mathbf{W}^{\mathrm{T}} + \beta^{-1}\mathbf{I} & \mathbf{W} \\ \mathbf{W}^{\mathrm{T}} & \mathbf{I} \end{bmatrix}\right)$$

$$p(\mathbf{y}|\mathbf{W}) = \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \beta^{-1}\mathbf{I})$$

- We can integrate out $\mathbf{X}$ and seek the maximum likelihood solution for $\mathbf{W}$
- The solution is the diagonalisation of the sample covariance as in the PCA algorithm

Why?

- Learning a representation of data is inherently ill-constrained
- A statistical model allows us to clearly define assumptions that leads to a solution
- Naturally leads to non-linearisation
  - require approximations of intractable integral

Summary

Summary

- Getting an intuition for data is key to fitting models
- High dimensionality is often an effect of the representation not the data
- MDS, PCA are very useful methods

Tricks of the trade

- Do PCA, always
- Learn to see gram/distance matrices
- Look at projections of the data
- evaluate neighbourhoods

Next Time

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

- statistical models are based on assumptions
- we include assumptions using integration
- integration is hard :-)

eof

References

- Cox, M and T Cox (Jan. 2008). "Multidimensional scaling". In: *Handbook of data visualization*.
- Tenenbaum, Joshua B, Vin de Silva, and John C Langford (Dec. 2000). "A Global Geometric Framework for Nonlinear Dimensionality Reduction". In: *Science* 290.5500, pp. 2319–2323.
- Tipping, Michael E and Christopher M Bishop (1999). "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.