Manifold Learning

Estimation of the Intrinsic Dimension

Master Informatique

Parcours Data Mining

Université Lumière Lyon 2

Université de Lyon

Intrinsic vs. Extrinsic dimensions

- In practice, high-dimensional data are not truly high-dimensional
- Manifold hypothesis : high-dimensional data resides on a low-dimensional manifold
- Input data: $\mathcal{X} = \{x_\alpha\}_{\alpha \in A} \in M \subset R^D$, where M is of dimension s

Definition

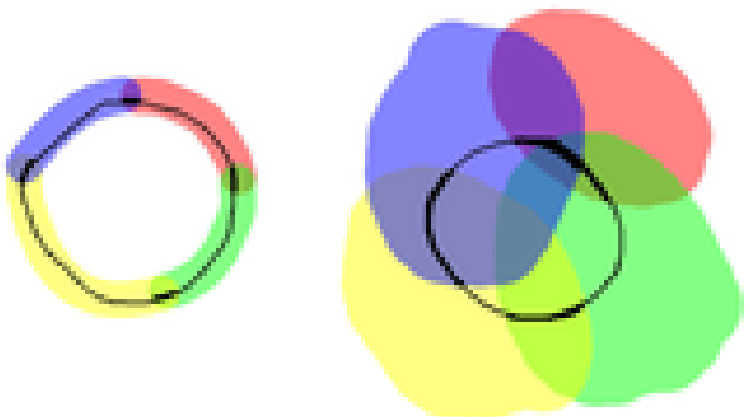
We call D the extrinsic dimension of $\mathcal{X}$ and s the intrinsic dimension of $\mathcal{X}$.

$s \ll D$: key to the feasibility of dimensionality reduction

A formal definition of the intrinsic dimension

Definition (Topological or Lebesgue Covering Dimension)

A subset $\mathcal{X}$ of a topological space S is said to have topological dimension s if every covering $\mathscr{C}$ of $\mathcal{X}$ has a refinement $\mathscr{C}'$ such that every point in $\mathcal{X}$ is covered by at most $s + 1$ open sets in $\mathscr{C}'$, and s is the smallest among these integers.



Source: Wikipedia

From theory to practice

- Formally appealing, the topological dimension is difficult to estimate
- Other definitions uses
  - Projections techniques (e.g. PCA)
  - Geometric approaches (e.g. fractal dimension or NN)

Note that the estimation of the intrinsic dimension should remain coherent with the data reduction DR method (an estimation of the dimension with a nonlinear model makes no sense if the dimensionality reduction uses a linear model)

Correlation dimension

Recall : $V_{sph}^{s}(r) = \frac{\pi^{s/2} r^{s}}{\Gamma(s/2+1)} \sim r^{s}$

- Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^D$
- If the data set lies on a $s-$dimensional manifold then the number of pair points closer than $r \sim r^s$
- Define

$$C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_{\{d(x_i,x_j)<r\}} \quad \text{and} \quad C(r) = \lim_{n \mapsto \infty} C_n(r)$$

Definition

$$s_{\text{corr}} = \lim_{r \mapsto 0} \frac{\log C(r)}{\log(r)}$$

Practical estimation

Given a set of points, estimate s_{corr} as follows

1. Compute the distances for all possible pairs of points
2. Determine the proportion of distances that are $\leq \epsilon$
3. Apply the log function, and divide by $\log \epsilon$
4. Compute the limit when ϵ tends to zero; this is s_{corr}

The difficult step is the last one. Using L'Hopital rule, one derives a better estimation

Definition

$$s_{\text{corr}}(\epsilon_1, \epsilon_2) = \frac{\log \hat{C}(\epsilon_2) - \log \hat{C}(\epsilon_1)}{\log \epsilon_2 - \log \epsilon_1}$$

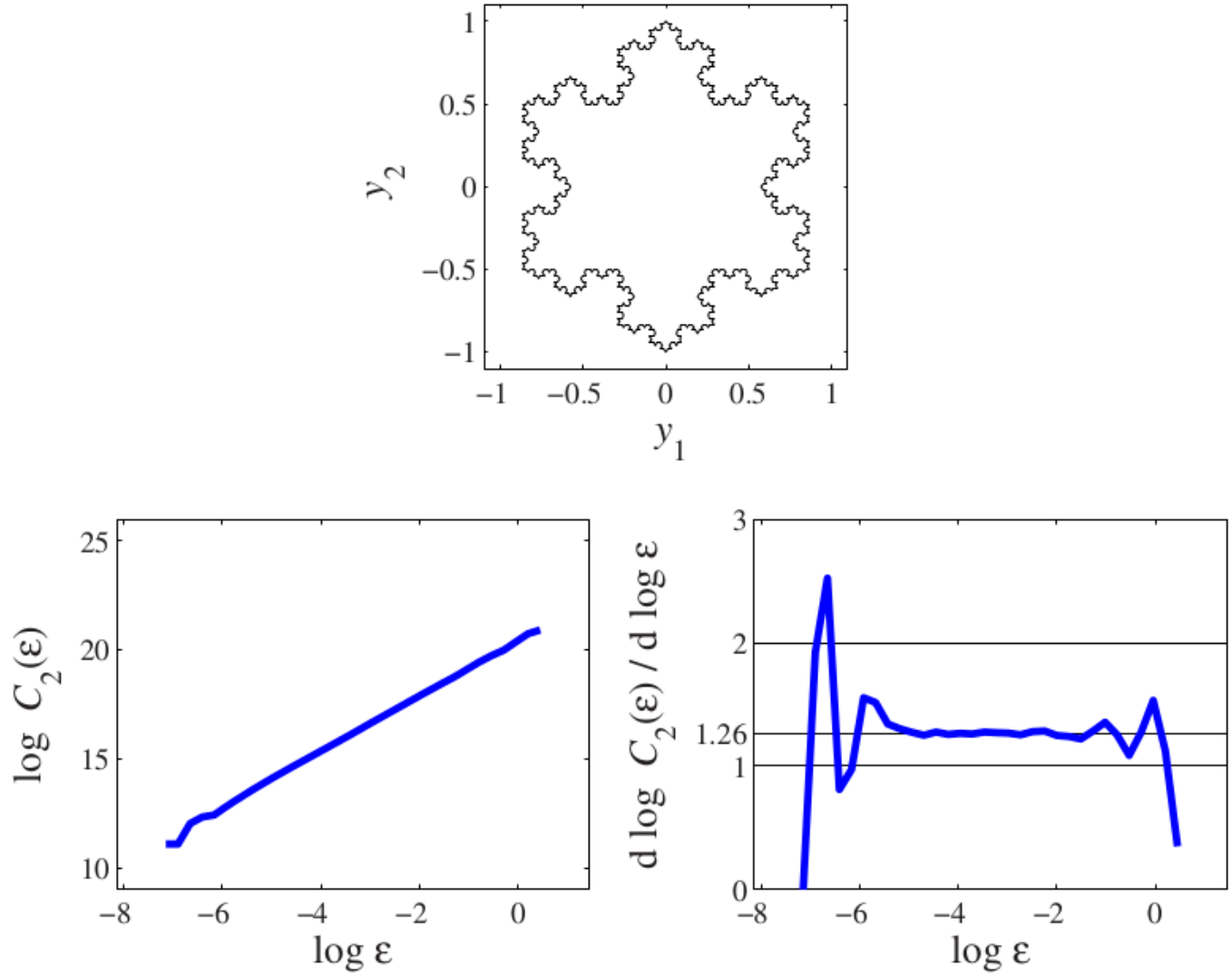


Fig. 3.3. Correlation dimension of Koch's island. The first plot shows the coastline, whose corners are the data set for the estimation of the correlation dimension. The log-log plots of the estimated correlation sum $\hat{C}_2(\epsilon)$ and its numerical derivative are displayed below.

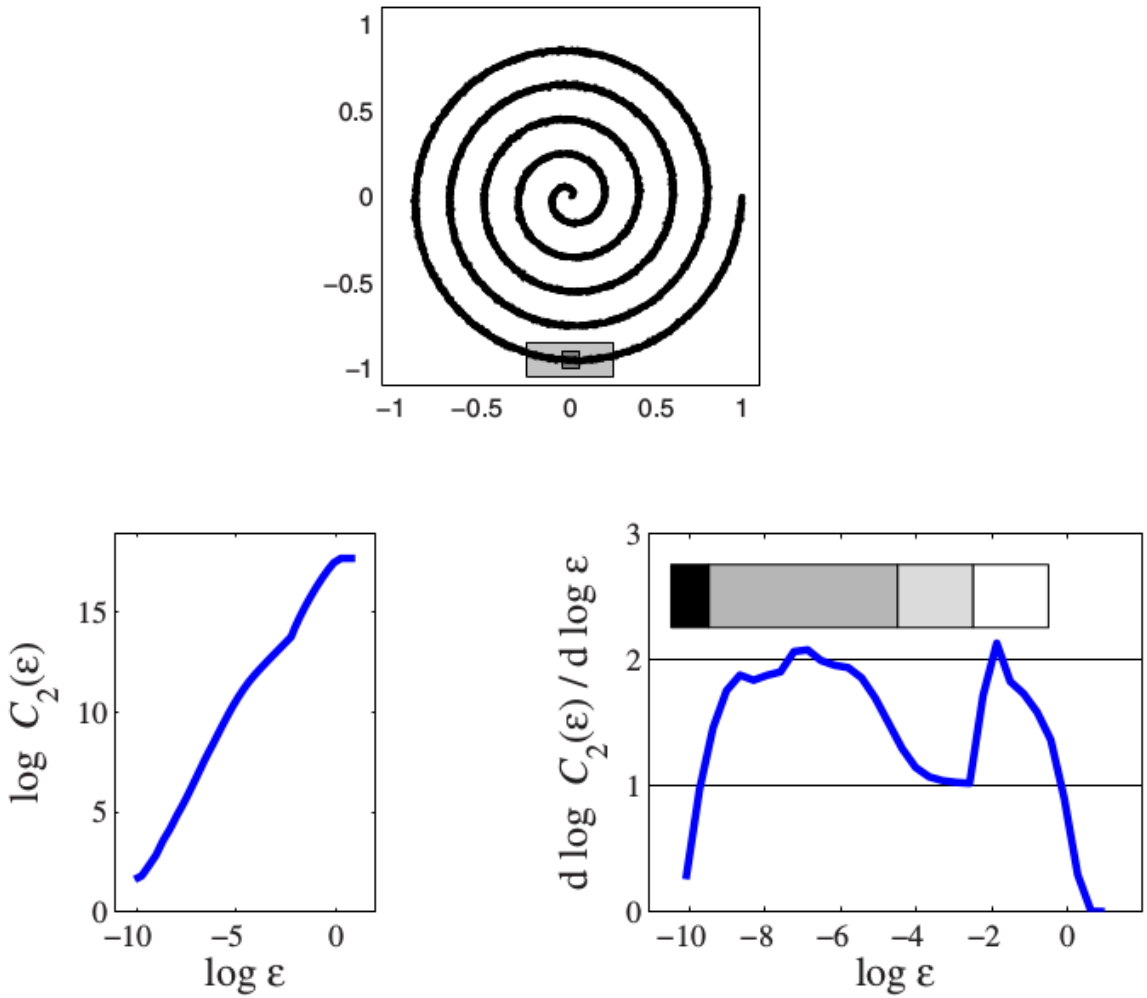
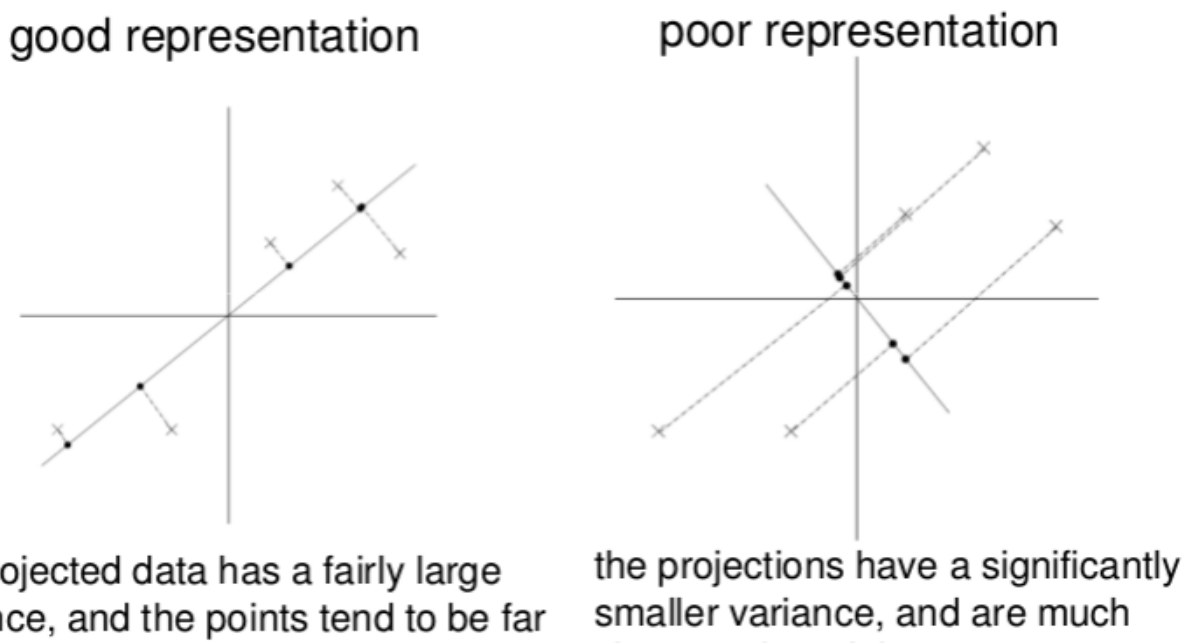


Fig. 3.4. Correlation dimension of a noisy spiral. The first plot shows the data set (10,000 points). The log-log plots of the estimated correlation sum $\hat{C}_2(\epsilon)$ and its numerical derivative are displayed below. The black, dark gray, light gray, and white boxes in the third plot illustrate that the correlation dimension depends on the observation scale. They correspond, respectively, to the scale of the isolated points, the noise, pieces of the spiral curve, and the whole spiral.

Recall: PCA in some slides

Pearson, 1901; Hotelling, 1933; Karhunen, 1946; Loève, 1948.

- Idea
  - Decorrelate zero-mean data
  - Keep large variance axes
  - Fit a plane though the data cloud and project
- Representation quality



Assume inputs are centered (i.e. $\sum_i x_i = 0$)

- Given a unit vector u and a point x, the length of the projection of x onto u is given by $x^T u$
- Maximize projected variance

$$\text{var}(y) = \frac{1}{n}\sum_i \left(x_i^T u\right)^2 = \frac{1}{n}\sum_i u^T x_i x_i^T u$$

$$= u^T \left(\frac{1}{n}\sum_i x_i x_i^T\right) u$$

- The inner matrix is called Gramm matrix $G = \frac{1}{n}\sum_i x_i x_i^T$.
- Maximizing $u^T G u$ s.t. $\|u\| = 1$ gives the principal eigenvector of G.

To project the data into a $p-$dimensional subspace ($s \ll D$) we take

- $u_1, \ldots, u_s$ the top s eigenvectors of G (which forms a orthogonal basis)
- The low dimensional outputs are $y_i = (u_1^T x_i, u_2^T x_i, \ldots, u_s^T x_i)^T$
- How to interpret the PCA:
  - Eigenvectors: principal axes of maximum variance subspace.
  - Eigenvalues: variance of projected inputs along principle axes.
  - Estimated dimensionality: number of significant (nonnegative) eigenvalues.