Estimating the intrinsic dimension of a data set

Kristian Nybo

Helsinki University of Technology

25.9.2007

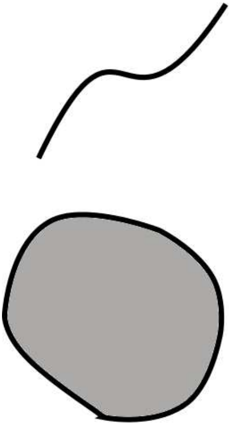
Outline

Outline

Introduction

- The *intrinsic dimension(ality)* of a data set is usually defined as the minimal number of parameters or latent variables required to describe the data
- We need to be able to estimate intrinsic dimensionality, because many DR methods need it but cannot estimate it themselves
- How do we translate the intuitive definition into something we can compute?
  - *Topological dimension* is formally exact, but hard to estimate for real data
  - Fractal dimensionality measures
  - Trial and error: ‘apply a DR method to the data, see what dimensionality works’

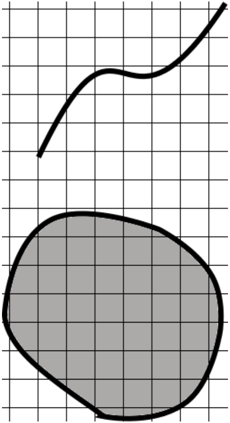
The ‘box-counting dimension’



- Determine the hypercube that circumscribes the data points
- Divide the hypercube into a grid of smaller hypercubes (‘boxes’) with edge length ϵ
- Determine $N(\epsilon)$, the number of boxes occupied by one or more data points.
- Idea: For a D-dimensional object, $N(\epsilon) \propto \epsilon^{-D} \Rightarrow D \propto -\frac{\log N(\epsilon)}{\log \epsilon}$
- Hence we define

$$d_{cap} = -\lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \epsilon} \qquad (1)$$

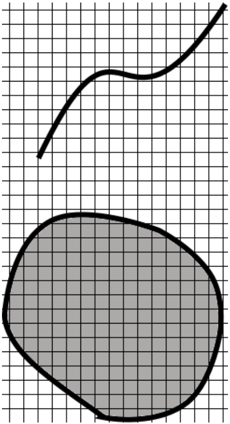
The ‘box-counting dimension’



- Determine the hypercube that circumscribes the data points
- Divide the hypercube into a grid of smaller hypercubes (‘boxes’) with edge length ϵ
- Determine $N(\epsilon)$, the number of boxes occupied by one or more data points.
- Idea: For a D-dimensional object, $N(\epsilon) \propto \epsilon^{-D} \Rightarrow D \propto -\frac{\log N(\epsilon)}{\log \epsilon}$
- Hence we define

$$d_{cap} = -\lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \epsilon} \qquad (1)$$

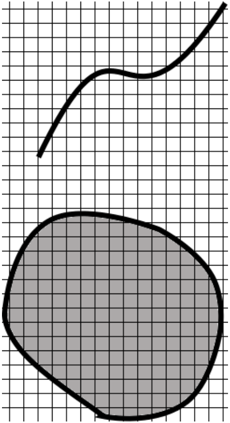
The ‘box-counting dimension’



- Determine the hypercube that circumscribes the data points
- Divide the hypercube into a grid of smaller hypercubes (‘boxes’) with edge length ϵ
- Determine $N(\epsilon)$, the number of boxes occupied by one or more data points.
- Idea: For a D-dimensional object, $N(\epsilon) \propto \epsilon^{-D} \Rightarrow D \propto -\frac{\log N(\epsilon)}{\log \epsilon}$
- Hence we define

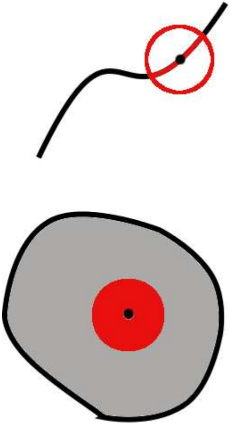
$$d_{cap} = -\lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \epsilon} \qquad (1)$$

Problems with capacity dimension



- For a real data set, the limit cannot be computed exactly
- To get even a good estimate, we need approximately 10^D data points for a D-dimensional manifold [1]

Correlation dimension



- $C_2(\epsilon)$ is the probability of two random points in the data set being within a distance ϵ of each other:

$$C_2(\epsilon) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i<j}^{N} H(\epsilon - \|\mathbf{y}_i - \mathbf{y}_j\|_2)$$
$$= P(\|\mathbf{y}_i - \mathbf{y}_j\|_2 \leq \epsilon), \quad (2)$$

where $H(x) = 1$ if $x \leq \epsilon$ and 0 otherwise.

- $C_2(\epsilon) \propto \epsilon^D$, so we define

$$d_{cor} = \lim_{\epsilon\to 0} \frac{\log C_2(\epsilon)}{\log \epsilon} \quad (3)$$

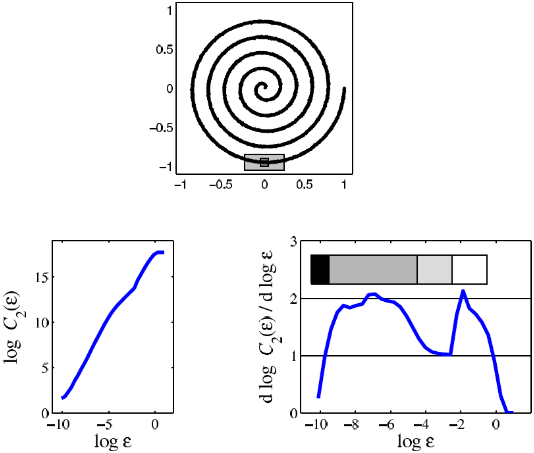
Estimating correlation dimension in practice

- L'Hopital:

$$\lim_{\epsilon \to 0} \frac{\log C_2(\epsilon)}{\log \epsilon} = \lim_{\epsilon \to 0} \frac{\partial \log C_2(\epsilon)}{\partial \log \epsilon} = \lim_{\epsilon_1 \to 0, \epsilon_2 \to 0} \frac{\log C_2(\epsilon_2) - \log C_2(\epsilon_1)}{\log \epsilon_2 - \log \epsilon_1} \quad (4)$$

- To estimate (4), ϵ_1 and ϵ_2 are usually chosen from a region where the log-log plot of $C_2(\epsilon)$ versus ϵ is almost constant
- Alternatively, we can calculate a second-order estimate for the derivative: $f'(x) = \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x} + \mathcal{O}(\Delta x^3)$
- Tsonis criterion: $10^{2+0.4P}$ points required for a good estimate [1]

An example of estimation



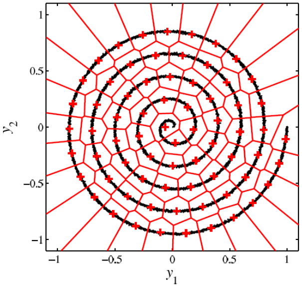
- The dependency of the estimate on ϵ is a feature, not a bug: ϵ represents the scale at which we observe the data, and the perceived dimensionality depends on that scale.

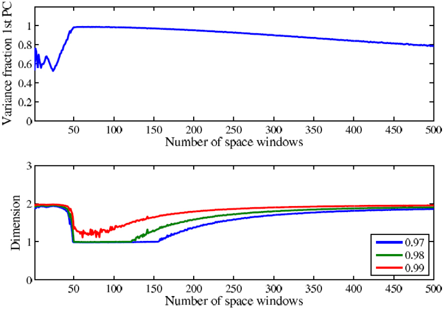
Outline

Outline of local PCA

- Divide the data set into small patches ("space windows") by clustering
- Apply PCA separately to each patch (assumption: the manifold is locally approximately linear)
- Estimate the dimensionality of the data as a weighted average of the dimensionalities of the patches
- Local PCA has the advantage that, in addition to the global dimensionality of a data set, it can also estimate local variations in dimensionality

Local PCA for a noisy spiral





Estimating dimensionality by trial and error

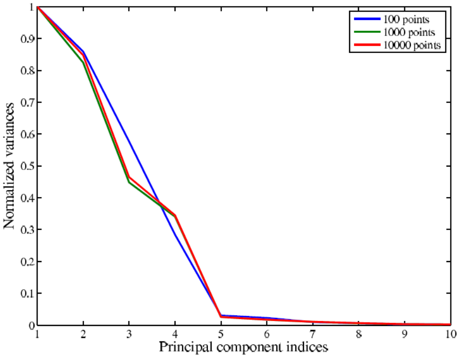
- Many NLDR methods minimize some kind of reconstruction error
- The reconstruction error should be minimal when the dimensionality of the projection equals the dimensionality of the manifold
- Thus we can try to estimate the dimensionality by observing how the reconstruction error varies with the dimensionality of the projection
- Disadvantage: very high computational cost

Outline

The data set

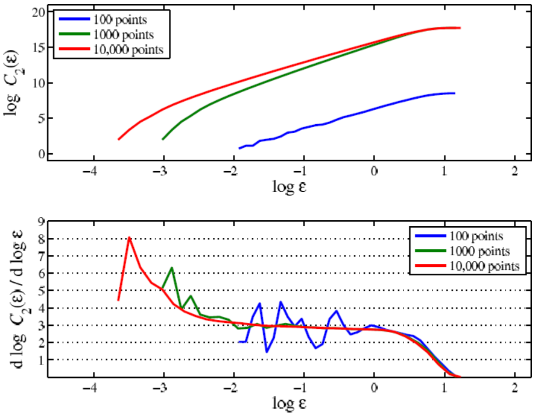
- In a three-dimensional cube, 10 distance sensors are placed at random locations
- The data points are uniformly distributed inside the cube
- Each data point is represented by a 10-dimensional vector, where each component is the point's distance to one of the sensors
- White Gaussian noise is added to each vector
- Thus the data set is a (slightly noisy) 3-dimensional nonlinear manifold embedded in 10-dimensional space

PCA



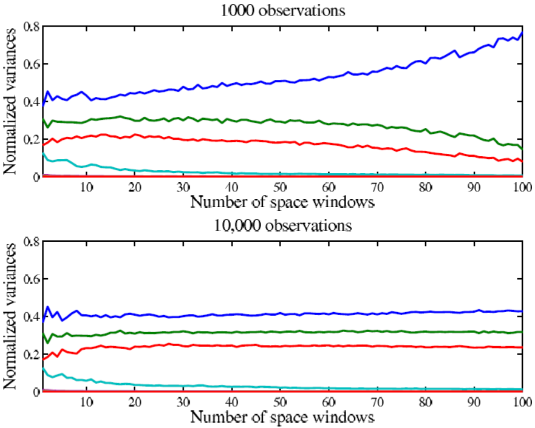
- Tends to overestimate dimensionality (this is to be expected, as the dependencies are not linear)
- Works for a small number of observations
- Fast

Correlation dimension



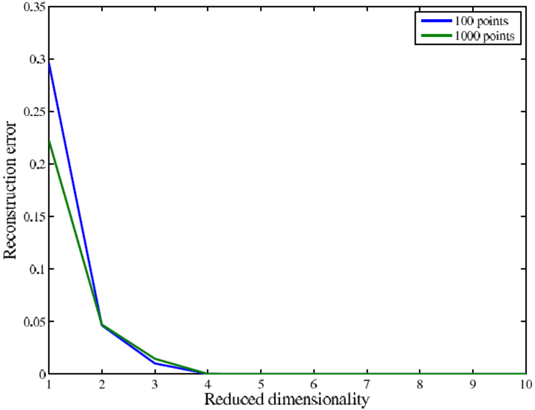
- Correct dimensionality
- More sensitive to the number of observations
- Much slower than PCA

Local PCA



- Correct dimensionality, even for a low number of windows
- Sensitive to the number of observations
- Much slower than PCA, but faster than correlation dimension

Trial and error with Sammon's mapping



- Overestimates dimensionality, but not by as much as PCA
- Works with a small number of observations
- Very slow

Summary

- PCA is not very accurate, but it is predictable and very fast
- Local PCA is accurate and fast
- Correlation dimension is slower, but it gives the dimension on all scales

References

Julien Clinton Sprott
Calculation of fractal dimension
http://sprott.physics.wisc.edu/phys505/lect12.htm