Estimating Intrinsic Dimension via Clustering

Brian Eriksson and Mark Crovella
brian.c.eriksson@gmail.com, crovella@cs.bu.edu

BOSTON UNIVERSITY

Problem Statement

Modern data analysis relies on the study of objects in a high *D*-dimensional space. Frequently, this data lies on a lower-dimensional manifold of dimension, *d*. This intrinsic dimension is often much smaller than the observed dimension (*d* << *D*).

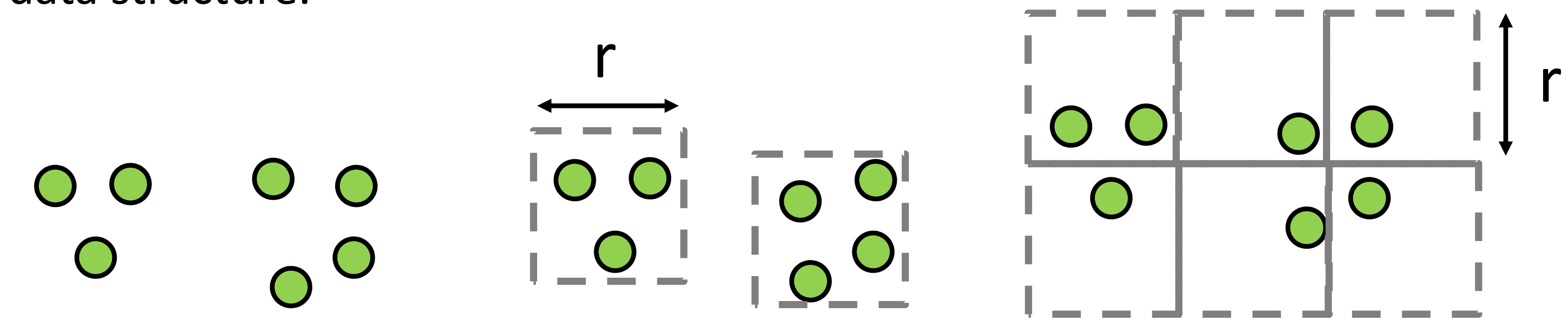
The intrinsic dimension is commonly estimated using a covering of the data:

Covering diagonal, *r*

Minimum covering number, *m(r)* ➡ $m(r) = r^{-d}$

Applications include: intrinsic dimension analysis include Internet topology analysis, computer vision, and computational finance.

Problem : Standard approaches, such as box counting, are agnostic to the data structure.



Set of objects

Minimum Covering
m(r) = 2

Box Counting Covering
$m_{box}(r) = 6$

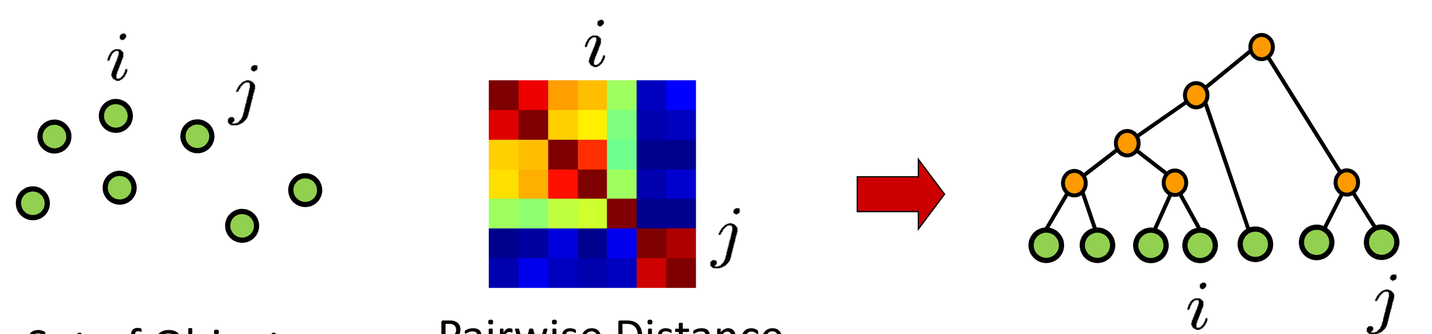
Additionally, often pairwise distances are observed, requiring embedding of the data for box counting techniques.

Can we estimate the intrinsic dimension using the inherent structure of the data?

Methodology

ClusterDimension Methodology:

We construct a Hierarchical Clustering of the set of items from pairwise distances.

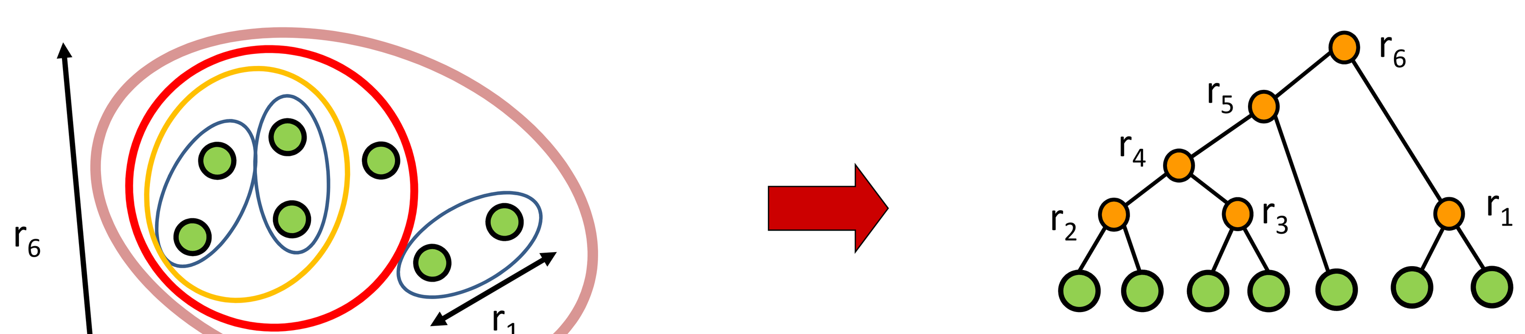


Set of Objects

Pairwise Distance Matrix

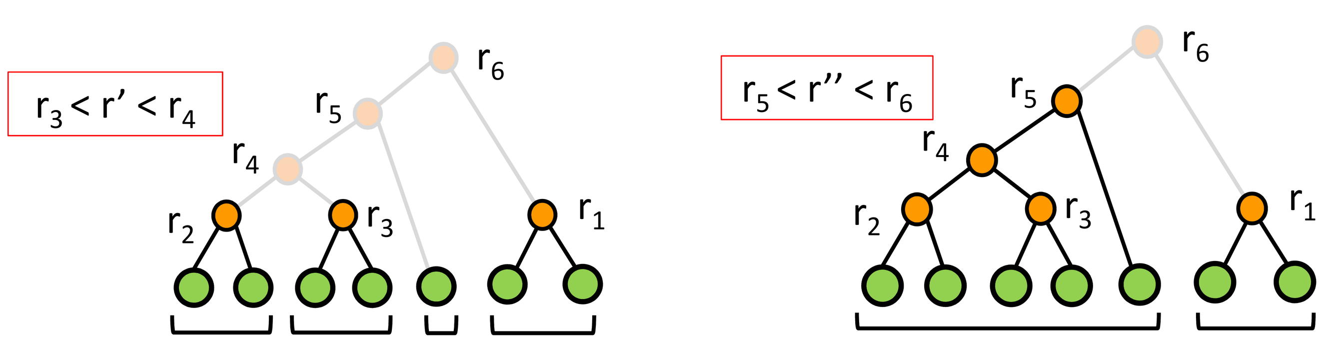
Hierarchical Clustering

This hierarchical clustering defines a nested covering of the objects. We annotate each cluster with the covering diameter.



Where: $r_6 > r_5 > r_4 > r_3 > r_2 > r_1$

By pruning this tree structure, we estimate the minimum covering, *m(r)*, for a specified covering diameter, *r*.



m(r') = 4

m(r'') = 2

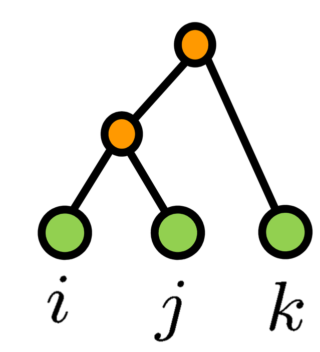
Using values of *m(r) and r*, we estimate the intrinsic dimension.

Consider pairwise distances that conform to a clustering structure.

Complete linkage condition:

Given three objects (i,j,k) where:

$\{i,j\} \in \mathcal{C}$ and $k \notin \mathcal{C}$



➡ $d_{i,j} < \min\left(d_{i,k}, d_{j,k}\right)$

If this condition holds, the estimated dimension using ClusterDimension converges to the true intrinsic dimension.

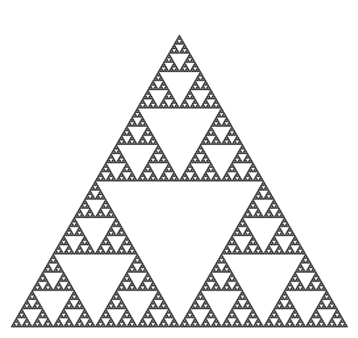
$$\lim_{r \to 0} \frac{-\log m(r)}{\log r} = d$$

Computational Complexity Comparison

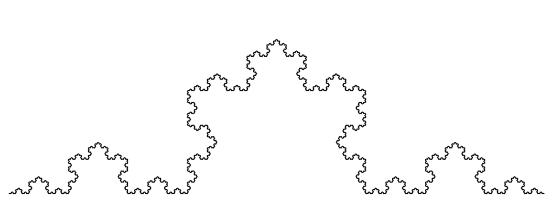
Dimension Estimation Method	Computational Complexity
ClusterDimension	$O(N^2)$
Maximum Likelihood [1]	$O(N^2)$
Box Counting [2]	$O(d_L N^2)$
Correlation Dimension [3]	$O(N^2)$
Minimum Spanning Tree [4]	$O(N^2 \log N)$
PCA	$O(d_L N^2)$

Experiments

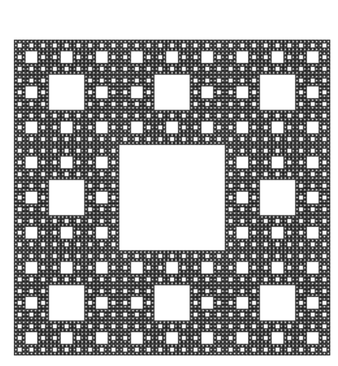
Validation is performed with respect to fractals where the non-integer intrinsic dimension is known.



Sierpinski Triangle
d = 1.58



Koch Curve
d = 1.26



Sierpinski Carpet
d = 1.89

Intrinsic Dimension Estimation RMSE
(averaged across 10 realization of 750 sampled points)

Dimension Estimation Method	Koch Curve	Sierpinski Triangle	Sierpinski Carpet	1-D Line
ClusterDimension	0.030	0.083	0.062	0.041
Maximum Likelihood [1]	0.062	0.065	0.245	0.043
Box Counting [2]	0.172	0.273	0.444	0.350
Correlation Dimension [3]	0.280	0.250	0.491	0.191
Minimum Spanning Tree [4]	0.248	0.221	0.091	0.131
PCA	0.738	0.416	0.107	0

References

[1] - L. Elizaveta, et. al., "Maximum Likelihood Estimation of Intrinsic Dimension," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 777–784.
[2] - B. B. Mandelbrot, "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," in *Science*, 1967, vol. 156, pp. 636–638.
[3] - Peter Grassberger, et. al., "Characterization of Strange Attractors," *Physical Review Letters A*, vol. 50, no. 5, pp. 346–349, January 1983.
[4] - V.J. Martinez, et. al., "Hausdorff Dimension from the Minimal Spanning Tree," in *Physical Review E*, January 1993, vol. 47, pp. 735–738.