ESTIMATING INTRINSIC DIMENSION VIA CLUSTERING

Brian Eriksson and Mark Crovella

Department of Computer Science, Boston University

ABSTRACT

Estimating the intrinsic dimension of a data set from pairwise distances is a critical issue for a wide range of disciplines, including genomics, finance, and networking. Current estimation techniques are agnostic to structure in the data, failing to exploit properties that can improve efficiency. In this paper, we present a methodology that uses inherent clustering present in data to efficiently and accurately estimate intrinsic dimension. Our experiments show that this approach has greater accuracy and better scalability than prior techniques, even when the data does not conform to an obvious clustering structure.

1. INTRODUCTION

Modern data analysis problems often rely on the study of objects observed in some high D-dimensional space, making direct analysis computationally intractable (the "curse of dimensionality"). However, frequently one finds that the data approximately lies on a lower-dimensonal manifold of, say, dimension d. This *intrinsic dimension* d represents the actual number of parameters needed to accurately approximate the data set. The intrinsic dimension can be much smaller than the observed dimension ($d \ll D$), allowing for tractable solutions to problems once the data is expressed in terms of d parameters. The property of low intrinsic dimension is commonly found in problems as diverse as genomics [1], network analysis [2], computational finance [3], and computer vision [4], to name only a few.

Estimating intrinsic dimension is a well-studied problem [5, 6, 7, 8, 9, 10]. However, despite the considerable prior work on estimating intrinsic dimension, prior methods have not sought to exploit structure in the data to decrease computational complexity. Examples of structure that could be used include clusters and hierarchy.

To exploit such structure in the data, we develop the CLUSTERDIMENSION algorithm, which efficiently calculates the intrinsic dimension of a data set using a particular kind of hierarchical clustering. We present sufficient conditions on data sets under which the output of CLUSTERDIMENSION will converge to the true intrinsic dimension. We show that CLUSTERDIMENSION has both decreased computational complexity and increased accuracy compared to state-of-the-art methods.

CLUSTERDIMENSION allows the analysis of data sets in which only distances (with or without metric embedding coordinates) can be observed. This makes it useful whether or not the data exhibits metric distance. Analysis of data sets which do not satisfy metric distance is important in gene microarray analysis [11] and Internet measurements [12].

2. CLUSTERING-BASED INTRINSIC DIMENSION ESTIMATION

To illustrate the intuition behind CLUSTERDIMENSION, we examine limitations of standard intrinsic dimension estimation techniques. Specifically, consider the performance of box-counting in Figure 1-(A). As seen in the figure, a fixed grid requires 6 boxes to cover this set of data points. However the best possible covering, as shown in Figure 1-(B), only requires 3 boxes of the same size. This inflation of the covering occurs because box-counting is agnostic to structure in the data.

To introduce CLUSTERDIMENSION we start with some definitions. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a collection of N items. Our observations consist of a set of values $\mathbf{D} = \{d_{ij}\}$ giving the distance between items i and j, from which we seek to determine the *intrinsic dimension* of $\mathbf{X}$.

Definition 1. *A **cluster** $\mathcal{C}$ is a subset of* $\mathbf{X}$. *A collection of clusters $\mathcal{T}$ is called a **hierarchical clustering of** $\mathbf{X}$ if $\cup_{\mathcal{C}_i \in \mathcal{T}} \mathcal{C}_i = \mathbf{X}$ and for any $\mathcal{C}_i, \mathcal{C}_j \in \mathcal{T}$, only one of the following is true **(i)** $\mathcal{C}_i \subset \mathcal{C}_j$, **(ii)** $\mathcal{C}_j \subset \mathcal{C}_i$, **(iii)** $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$.*

Given the set of pairwise distances $\mathbf{D} = \{d_{ij}\}$, CLUSTERDIMENSION starts by constructing a hierarchical clustering $\mathcal{T}$. A hierarchical clustering can be constructed using a variety of methodologies, including divisive [13] and agglomerative methods [14]; CLUSTERDIMENSION uses *minimum-linkage agglomerative clustering*. The approach starts by placing each object into a distinct cluster. It then iteratively finds the smallest cluster distance, and creates a new cluster containing those clusters. This process repeats until a single cluster is constructed which contains every object in $\mathbf{X}$.

We denote the hierarchical clustering obtained using this technique as $\widehat{\mathcal{T}}$. The clustering $\widehat{\mathcal{T}}$ is isomorphic to a tree in which each interior node corresponds to a cluster that contains all descendants of the node. (We often treat $\widehat{\mathcal{T}}$ as a tree in the rest of the paper.) Each interior node of $\widehat{\mathcal{T}}$ is annotated

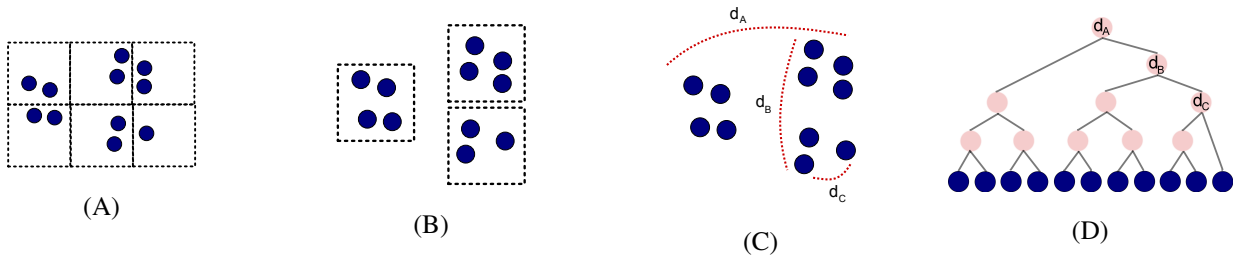

Fig. 1. (A) Six grid boxes covering a set of points; (B) The same set of points covered by only three boxes of the same size; (C) Points with selected pairwise distances labeled ; (D) Annotated hierarchical clustering. (For clarity, only distances d_A, d_B, d_C are shown.)

with the *maximum* distance between any pair of items in the corresponding cluster. An example of this annotated tree is shown in Figure 1-(D) for the example data in Figure 1-(C).

CLUSTERDIMENSION uses each annotated distance as an estimate for the minimum covering diameter for the items in the corresponding cluster. Given a threshold r, the strategy taken by CLUSTERDIMENSION is to remove all nodes in the tree whose parent has annotation less than r. The number of leaves found in the pruned tree, $\widehat{m}(r)$, is used as an estimate of the size of the minimum covering using balls of diameter r. In what follows we establish conditions under which this estimate is accurate.

2.1. Performance

Consider the following condition on the observed pairwise distances $\mathbf{D}$ with an associated hierarchical clustering $\mathcal{T}$ over a set of items $\mathbf{X}$.

Definition 2. *The triple* $(\mathbf{X}, \mathcal{T}, \mathbf{D})$ *satisfies the* ***Complete Linkage Condition*** *if for every set of three items* $\{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k\}$ *such that* $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}$ *and* $\mathbf{x}_k \notin \mathcal{C}$ *for some* $\mathcal{C} \in \mathcal{T}$*, the distances satisfy* $d_{ij} < \min(d_{ik}, d_{jk})$.

For data sets satisfying the Complete Linkage Condition, we now show that the number of clusters found for distance r is equal to the minimum number of balls of diameter r needed to cover $\mathbf{X}$.

Proposition 1. *If* $(\mathbf{X}, \mathcal{T}, \mathbf{D})$ *satisfies the Complete Linkage Condition, then using the annotated minimum-linkage tree approach, the estimated covering number is equal to the minimum covering number,* $\widehat{m}(r) = m(r)$*, for all values of* r.

Proof. Consider a violation of this proposition. In that case there exists an r such that the number of leaves in the collapsed tree is not equal to the minimal ball covering with diameter r, *i.e.,* $\widehat{m}(r) \neq m(r)$. By definition $\widehat{m}(r)$ cannot be less than $m(r)$, since each cluster with annotation r can be covered by a ball of diameter r. Hence, $\widehat{m}(r) > m(r)$, so it must be that there exist (at least) two items, $\mathbf{x}_i$ and $\mathbf{x}_j$ such that $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different clusters in $\mathcal{T}$ but are covered by the same ball. Then the distance between x_i, x_j, $d_{ij} < r$ which violates the Complete Linkage Condition. □

Next, we review a standard definition of ball-covering dimension:

Definition 3. *A point set* $\mathbf{X}$ *has ball-covering dimension* d *if and only if* $\lim_{r \to 0} \frac{\log m(r)}{\log r} = d$.

We can now state the following theorem establishing the performance of CLUSTERDIMENSION:

Theorem 2.1. *For a triple* $(\mathbf{X}, \mathcal{T}, \mathbf{D})$ *that satisfies the Complete Linkage condition, if* $\lim_{r \to 0} \frac{\log \widehat{m}(r)}{\log r} = d$ *then* $\mathbf{X}$ *has ball-covering dimension* d.

Proof. Follows directly from Proposition 1 and Definition 3. □

Theorem 2.1 shows that CLUSTERDIMENSION accurately estimates the intrinsic dimension of data sets that satisfy the Complete Linkage Condition.

2.2. The CLUSTERDIMENSION Algorithm

Using the insights of Theorem 2.1, we introduce the CLUSTERDIMENSION algorithm. The strategy starts by obtaining an annotated minimum-linkage hierarchical clustering that conforms to the given set of pairwise distances, $\mathbf{D}$. The annotated clustering is then used to find $\widehat{m}(r)$ — the inferred minimum number of clusters corresponding to covering $\mathbf{X}$ with balls of size r. Finally, the the intrinsic dimension $\widehat{d}$ is estimated as the power law relationship between the observed values of $\widehat{m}(r)$ and r, such that $\widehat{m}(r) = r^{-\widehat{d}}$. A formal description of CLUSTERDIMENSION is given in Algorithm 1. Although we have only established the performance of CLUSTERDIMENSION under the Complete Linkage Condition, the experiments later will not impose this condition on the data sets observed, and will demonstrate the generality of our technique.

2.3. Implementation Details

When working with finite data, one obviously cannot resolve $\hat{m}(r)$ as $r \to 0$. Instead, one must examine the scaling of $\hat{m}(r)$ over an appropriate range of scales. This is a standard

Algorithm 1 - CLUSTERDIMENSION

Input:

A set of items: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$.

An $N \times N$ matrix of pairwise distances: $\mathbf{D} = \{d_{ij}\}$.

Minimum and maximum scales of interest: $r_{\min}$ and $r_{\max}$.

Scaling increment: Δr.

Main Body:

Using the pairwise distances $\mathbf{D}$, construct $\widehat{\mathcal{T}}$ using minimum linkage agglomerative clustering [14].

Annotate each interior node of $\widehat{\mathcal{T}}$ with the maximum distance between any two nodes in its cluster.

For $r = \{r_{\min}, r_{\min} + \Delta r, r_{\min} + 2\Delta r, ..., r_{\max}\}$

1. Prune $\widehat{\mathcal{T}}$ by removing all nodes whose parent has annotation $\leq r$.
2. Set $\hat{m}(r) =$ number of leaf nodes in pruned $\widehat{\mathcal{T}}$

Output:

Return the estimated intrinsic dimension, $\widehat{d}$, as the power-law scaling relationship between $\hat{m}(r)$ and r.

problem that is common to all techniques for estimating intrinsic dimension; addressing this problem optimally is outside the scope of this paper. For the purposes of evaluating CLUSTERDIMENSION in this paper, we adopt simple heuristics for determining the range of scales r, and we apply the same heuristics across all algorithms we compare. To determine the minimum scale of interest r_{min}, we choose the smallest value of r for which the median cluster size is greater than one (*i.e.,* the diameter where less than half of the found clusters are singletons). The maximum scale of interest r_{max} is taken to be the diameter of $\mathbf{X}$. These heuristics allow us to estimate intrinsic dimension despite the finiteness of the data set and without the need to manually inspect the data. The same range of scales is used in all dimension estimation methodologies that require scale examination (*e.g.,* box counting, MST, MLE, etc.) to provide a consistent and fair comparison.

Likewise, it is necessary to estimate the power-law scaling relationship between $\hat{m}(r)$ and r. For this purpose we fit a least-squares line to the points given by $(\log r, \log \hat{m}(r))$. Again, this approach is applied to all algorithms that require estimating a power-law relationship.

2.4. Complexity

Estimating the intrinsic dimension of large data sets in acceptable time requires a method with low computational complexity. Table 1 compares the computational complexity of CLUSTERDIMENSION with that of the most commonly used alternatives: a Maximum Likelihood technique [7], box counting [9], correlation dimension [6], a minimum spanning tree-based approach [8], and linear PCA [14].

The Table shows that there is no algorithm with lower computational complexity than CLUSTERDIMENSION. CLUSTERDIMENSION requires only $O\left(N^2\right)$ operations, as $O\left(N^2\right)$ operations are required to construct the minimum-linkage hierarchical clustering ([14]), and $O\left(N\right)$ operations are then required to prune the tree to obtain the number of leaf nodes for a given distance (as there are at most N interior nodes in the tree structure).

In evaluating computational complexity, it is important to consider the form in which the data is presented. CLUSTERDIMENSION only requires knowledge of inter-point distances; however, methods that rely on a linear embedding of the data (*e.g.,* PCA or box counting) are dependent on the embedding dimension of the data (d_ℓ) which for non-linear real-world data can potentially approach the size of the data set, N.

Table 1. Computational Complexity of Intrinsic Dimension Estimation Algorithms (for N items with linear embedding dimension d_ℓ)

Dimension Estimation Method	Computational Complexity
CLUSTERDIMENSION	$O\left(N^2\right)$
Maximum Likelihood [7]	$O\left(N^2\right)$
Box Counting [9]	$O\left(d_\ell N^2\right)$
Correlation Dimension [6]	$O\left(N^2\right)$
Minimum Spanning Trees [8]	$O\left(N^2 \log N\right)$
PCA [14]	$O\left(d_\ell N^2\right)$

2.5. Accuracy

To evaluate the accuracy of each of the dimension estimation techniques in Table 1, we apply each technique to a collection of fractals with known intrinsic dimension. We use the Koch Curve, the Sierpinski Triangle, and the Sierpinski Carpet. Use of these fractals allows us to validate our dimension estimation techniques on data with known ground-truth non-integer intrinsic dimension, and the choice of these fractals spans a range of true dimensions. To demonstrate the performance on high-dimensional data, we also test on a 1-D line in 100 dimensional space.

Table 2 shows the estimated intrinsic dimension over 10 random realizations of each of the data sets where each realization samples 750 points at random. The fractal data sets are constructed to a depth of 50 self-similar iterations. For all data sets, the pairwise distance between two points is the Euclidean distance. The table presents the results in terms of both the average intrinsic dimension using the various esti-

mation techniques, and the root mean squared error (RMSE) over all realizations for the estimated intrinsic dimension. We find that CLUSTERDIMENSION gives the best estimate over all methods for two of the three fractals, and the RMSE of CLUSTERDIMENSION is consistently among the smallest for all four data sets.

Table 2. Intrinsic Dimension Estimation Root Mean Squared Error (RMSE) for self-similar fractals (Koch Curve, Sierpinski Triangle, and Sierpinski Carpet) in 2-D space and a 1-D line in 100-D space from 750 uniformly sampled points and error results across 10 realizations.

Method	Koch Curve $d = 1.26$	Sierp. Triangle $d = 1.58$	Sierp. Carpet $d = 1.89$	1-D Line $d = 1$
CLUSTER DIMENSION	0.030	0.083	0.062	0.041
MLE	0.062	0.065	0.245	0.043
Box Count.	0.172	0.273	0.444	0.035
Correlation	0.280	0.250	0.491	0.191
MST	0.248	0.221	0.091	0.131
PCA	0.738	0.416	0.107	0

3. CONCLUSIONS

Estimating the intrinsic dimension of data is critical for a wide range of real world problems. In this paper, we present a new approach to intrinsic dimension estimation, requiring only pairwise distances between data items. This new approach uses clustering rather than geometric embedding, which affords both low complexity and improved performance compared to state of the art alternatives. Experiments on both synthetic and real-world data show the improvements of these techniques over prior methods. In future work we look to examine multi-class classification using our dimension-based clustering, and intrinsic dimension estimation using incomplete observations of pairwise distance.

4. REFERENCES

[1] H. Lahdesmaki, O. Y. Harja, W. Zhang, and I. Shmulevich, "Intrinsic Dimensionality in Gene Expression Analysis," in *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics (GENSiPS)*, 2005, vol. 2.

[2] B. Abrahao and R. Kleinberg, "On the Internet Delay Space Dimensionality," in *Proceedings of ACM Internet Measurement Conference (IMC)*, 2008, pp. 157–168.

[3] M. Verleysen, E. de Bodt, and A. Lendasse, "Forecasting Financial Time Series through Intrinsic Dimension Estimation and Non-Linear Data Projection," in *Proceedings of International Work-Conference on Artificial and Natural Neural Networks (IWANN)*, Alicante, Spain, June 1999.

[4] K.M. Carter, R. Raich, and A.O. Hero, "On Local Intrinsic Dimension Estimation and Its Applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 650 –663, February 2010.

[5] F. Hausdorff, "Dimension Und Ausseres Mass," *Mathematics Annalen*, vol. 79, 1919.

[6] Peter Grassberger and Itamar Procaccia, "Characterization of strange attractors," *Physical Review Letters A*, vol. 50, no. 5, pp. 346–349, January 1983.

[7] L. Elizaveta and P. J. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 777–784.

[8] V.J. Martinez, R. Dominguez Tenreiro, and L. J. Roy, "Hausdorff Dimension from the Minimal Spanning Tree," in *Physical Review E*, January 1993, vol. 47, pp. 735–738.

[9] B. B. Mandelbrot, "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," in *Science*, 1967, vol. 156, pp. 636–638.

[10] J. Theiler., "Estimating Fractal Dimension," in *Journal of the Optical Society of America*, 1990, vol. 7, pp. 1055–1073.

[11] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering Short Time Series Gene Expression Data," in *Bioinformatics*, 2005, vol. 21, pp. 159–168.

[12] Han Zheng, Eng Keong Lua, Marcelo Pias, and Timothy G. Griffin, "Internet routing policies and round-trip-times," in *Proceedings of the Passive and Active Measurement Workshop (PAM)*, 2005.

[13] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, "Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities," in *Proceedings of AISTATS 2011*, April 2011.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.