

# Estimation of Intrinsic Dimension via Clustering

Brian Eriksson  
Department of Computer Science  
Boston University  
eriksson@cs.bu.edu

Mark Crovella  
Department of Computer Science  
Boston University  
crovella@bu.edu

June 6, 2011

BU/CS Technical Report 2011-12

## Abstract

The problem of 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 the structure of the data, resulting in techniques that may be computationally intractable for large data sets. In this paper, we present a methodology that exploits the inherent clustering structure of data to efficiently estimate intrinsic dimension. Our experiments show that the clustering-based approach allows for more accurate intrinsic dimension estimation and decreased computational complexity compared with prior techniques, even when the data does not conform to an obvious clustering structure. Finally, we present results which shows the clustering-based estimation allows for a natural partitioning of the data points that lie on separate manifolds of different intrinsic dimension.

## 1 Introduction

Modern data analysis problems often rely on the study of objects observed in some high  $D$ -dimensional space. Due to the “curse of dimensionality”, the analysis of the data set in all  $D$  dimensions may be computationally intractable. Fortunately, we are often helped by the data lying on a manifold of *Intrinsic Dimension*,  $d$ , representing the true number of variables needed to describe the data set. Commonly, this intrinsic dimension is much smaller than the observed dimensions ( $d \ll D$ ), allowing for tractable solutions to problems that would be impossible in larger dimensions. This reduction in dimensionality is commonly found in problems as diverse as, genomics [1], Internet topology analysis [2], computational finance [3], and computer vision [4], to name only a few.

The estimation of the intrinsic dimension for a given data set is a well-studied problem examined previously in [5, 6, 7, 8, 9, 10]. While there has been a large amount of prior work on estimating the intrinsic dimension of a data set, none of the prior methodologies exploit the structure of the data to increase accuracy and decrease computational complexity. In this paper, we consider the general case where only distances, not metric embedding coordinates, can be observed. Dimension estimation using pairwise distances is critical for real-world data sets which do not satisfy metric distances, as commonly found in gene microarray analysis [11] and Internet measurements [12].

To exploit the structure of the data, we develop the CLUSTERDIMENSION algorithm, which uses pairwise distances to efficiently calculate the intrinsic dimension of a data set through the use of a modified hierarchical clustering methodology. We present sufficient conditions on data sets where the CLUSTERDIMENSION algorithm estimated intrinsic dimension converges to the true intrinsic dimension. We show the decreased computation complexity and demonstrate the increased accuracy of our estimation methodology on data sets with known intrinsic dimension.

Additionally, we show how our clustering-based techniques can be extended to cluster data points based on our dimension estimation. This allows for decomposition of a data set generated from a mixture of processes of different intrinsic dimension. Finally, we demonstrate how dimension-based clustering can be applied to outlier detection on real-world data.

The paper is organized as follows. The intrinsic dimension estimation problem and prior methods are introduced in Section 2. By exploiting the structure of data, in Section 3 we introduce our clustering-based intrinsic dimension

estimation approach. We then show how our clustering-based technique allows for a natural partitioning of data sets with respect to dimension in Section 4. Concluding remarks are made in Section 5.

## 2 The Intrinsic Dimension Problem and Prior Work

Let  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  be a collection of  $N$  items in  $\mathbb{R}^D$  space. We observe pairwise distances between items, the matrix  $\mathbf{D}$ , where  $d_{i,j}$  is the distance between items  $i$  and  $j$ . Using these pairwise distances, our goal will be to resolve the *intrinsic dimension* of the set of items. Informally, we will consider the intrinsic dimension,  $d$ , to be the largest dimension such that the data set sufficiently fills a region in  $d$ -dimensional space.

Standard approaches, such as Principal Component Analysis (PCA), allow for accurate intrinsic dimension estimation when the data is linearly embedded into a lower dimensional space. When the data is embedded onto a nonlinear manifold, these linear methods can dramatically overestimate the intrinsic dimension. To avoid this problem, nonlinear methods have been popularized in recent years, including ISOMAP [13] and Locally Linear Embeddings [14], although both resolve an integer intrinsic dimension.

In this paper, we consider the more general case in which the intrinsic dimension is a non-integer, or *fractal*, dimension. The most accurate measure of fractal dimension, the *Hausdorff Dimension* [5], examines the relationship between the size of the smallest ball covering of the data set,  $m(r)$ , and the ball radius,  $r$ . Unfortunately, finding the best placement of the ball covering is combinatorial in the number of items in the data set, making finding the true Hausdorff dimension of a data set intractable for sets of any realistic size.

To approximate bounds on the minimal ball covering of a data set (and therefore the Hausdorff Dimension), the *Box Counting Dimension* is often used. In this approach, a grid is laid on the data set, with separation between each grid point being of size  $r$ , and we consider the number of grid boxes necessary to completely cover the data. Using box counting, the intrinsic dimension is estimated as the power law relationship between the number of observed grid boxes containing data points and the grid length ( $r$ ). The main limitation with the box counting approach is the discrepancy between the box counting number and the true minimal covering size as the ambient dimensionality of the data set ( $D$ ) grows [15].

Given issues with both the Hausdorff covering and the box counting approach, commonly the *Correlation Dimension* [6] is used to approximate the fractal dimension of a dataset. This approach is dependent on finding the number of nearest-neighbors for each item within a radius of size  $r$ . Although this estimator has less computational complexity than the box counting approach, under some circumstances it has been shown to have drastically higher error [16].

One of the more recent intrinsic dimension approaches, a Maximum Likelihood approach [7], estimates intrinsic dimensionality by approximating the number of nearest neighbors for each data point by a Poisson point process and resolving the most likely dimension given this parametric model. Empirical results have shown lower bias and variance for intrinsic dimension estimation using this MLE approach over the Correlation dimension methodology.

Finally, the most similar prior work with respect to the methodology discussed in this paper is the Minimum Spanning Tree based approach in [8]. In this work, the intrinsic dimension of the data set is estimated by first constructing a minimum spanning tree given the pairwise distances, and then resolving a ball covering on the graph structure. In contrast with this approach, our cluster-based approach will have lower computational complexity ( $O(N^2)$  vs.  $O(N^2 \log N)$  for the spanning tree methodology), and we extend our approach to naturally partition data in terms of intrinsic dimension.

Our dimension estimation approach is dependent on the construction of a hierarchical clustering based on the observed distances. While recent results on hierarchical clustering [17] have shown only  $O(N \log N)$  operations are needed to resolve the tree structure (as opposed to  $O(N^2)$  for standard agglomerative methods, [18]), this requires an additional restrictive condition on how the pairwise distances conform to the natural set of clusters. To make our approach general, we will be agnostic to the method used to construct the hierarchical clustering of the data. This opens our procedure to potentially exploit recent work on efficient (e.g., [17]), or robust (e.g., [19]) hierarchical clustering methods.

## 3 Clustering-Based Intrinsic Dimension Estimation

Several limitations occur with standard intrinsic dimension estimation techniques. Specifically, consider an example of the performance of a box counting-based approach in Figure 1-(A). As seen in the figure, a fixed grid requires 6 boxes to cover this set of data points. This is in contrast to the best possible covering of this small example, as seen

in Figure 1-(B), which only requires 3 boxes of the same size. This inflation of the covering is the result of the box counting technique being agnostic to the structure of the data.

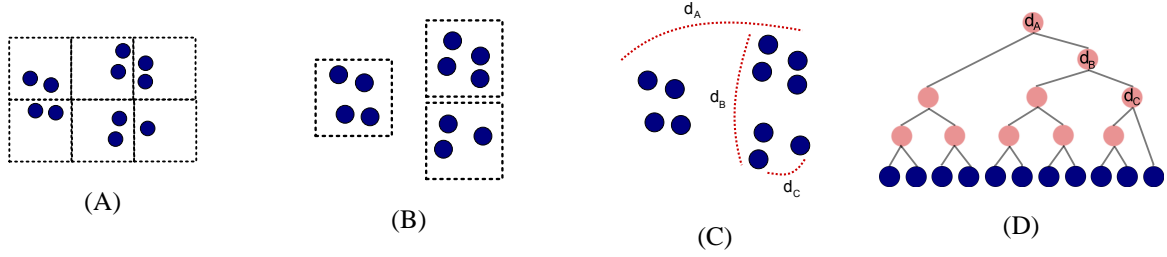


Figure 1: (A) Six grid boxes covering a set of points; (B) The same set of points covered by only three boxes of the same size; (C) Observed pairwise distances; (D) Annotated hierarchical clustering. (For clarity, only distances  $d_A, d_B, d_C$  are shown.)

Using observed pairwise distances, we look to exploit the inherent clustering structure in the data. We construct a *hierarchical clustering* using pairwise distances. Formally defining hierarchical clustering as:

**Definition 1.** A *cluster*  $\mathcal{C}$  is defined as any subset of  $\mathbf{X}$ . A collection of clusters  $\mathcal{T}$  is called a **hierarchical clustering** 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)  $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ .

The hierarchical clustering  $\mathcal{T}$  has the form of a tree, where each interior node corresponds to a particular cluster. While the hierarchical clustering can be constructed using a variety of methodologies, such as divisive [17] or agglomerative methods [18], we will be agnostic to the particular methodology used to cluster. Regardless of the clustering technique used, we will require annotating the tree structure with the maximum distance found in each cluster, as seen in Figure 1-(D). This will allow us to examine the resolved tree structure and intrinsic dimension, as these annotated distances will be an estimate for the minimum covering diameter for clusters found in the hierarchy.

### 3.1 The CLUSTERDIMENSION Algorithm

Using this hierarchical clustering-based methodology, we introduce the CLUSTERDIMENSION methodology in Algorithm 1. This methodology consists of obtaining a hierarchical clustering that conforms to the data set with annotated interior distances, an example of which is seen in Figure 1-(D). We exploit these annotated distances to resolve the intrinsic dimension of data by merging all the interior nodes in the tree with annotated distance less than a threshold,  $r$ , thus forming a collapsed subtree structure. We then evaluate the number of leaves found in the collapsed subtree,  $\hat{m}(r)$ . We estimate the intrinsic dimension,  $\hat{d}$ , as the power law relationship between the observed values of  $\hat{m}(r)$  and  $r$ , such that  $\hat{m}(r) = r^{-\hat{d}}$ .

Due to the use of real world, discrete data sets, we cannot resolve the cluster covering as  $r \rightarrow 0$ . Instead, we must only examine the covering characteristics for a feasible range of covering sizes. We limit  $r$  to be greater than the value  $\max_{i=\{1,2,\dots,N\}} d_i^{(K)}$ , where  $d_i^{(K)}$  is the  $K$ -th nearest neighbor distance to item  $i$ . This will allow us to characterize the data set intrinsic dimension while mitigating the influence of discrete data.

#### 3.1.1 CLUSTERDIMENSION Algorithm Analysis

Critical to the dimension estimation performance of the CLUSTERDIMENSION algorithm is the relationship between the estimated number of clusters for a given diameter,  $\hat{m}(r)$ , compared with the minimum number of clusters possible for a given diameter,  $m(r)$ .

We consider the following condition on the observed pairwise distances,  $\mathbf{D}$ , conforming to an underlying tree structure,  $\mathcal{T}$ .

**Definition 2.** The triple  $(\mathbf{X}, \mathcal{T}, \mathbf{D})$  satisfies the **Complete Linkage Condition** if for every set of three items  $\{x_i, x_j, x_k\}$  such that  $x_i, x_j \in \mathcal{C}$  and  $x_k \notin \mathcal{C}$ , for some  $\mathcal{C} \in \mathcal{T}$ , the distances satisfy,  $d_{i,j} < \max(d_{i,k}, d_{j,k})$ .

Given this Complete Linkage condition [18], we can state the following proposition with respect to the CLUSTERDIMENSION algorithm.

---

**Algorithm 1** - CLUSTERDIMENSION( $\mathbf{X}, \mathbf{D}$ )

---

**Input:**

1. A set of items,  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ .
2. An  $N \times N$  matrix of pairwise distances,  $\mathbf{D} = \{d_{i,j}\}$ .
3. Set the maximum distance length,  $r_{\max} = \max_{i,j} d_{i,j}$ .
4. Set the minimum distance length,  $r_{\min} = \max_{i=\{1,2,\dots,N\}} d_i^{(K)}$ , where  $d_i^{(K)}$  is the  $K$ -th nearest neighbor distance to item  $i$ .
5. Choose an increment  $\Delta r$  that divides  $(r_{\min}, r_{\max})$  into a sufficient number of intervals.

**Main Body:**

Using the pairwise distances,  $\mathbf{D}$ , estimate the clustering hierarchy,  $\widehat{\mathcal{T}}$  (e.g., using agglomerative clustering [18]) with annotated interior distances.

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

1. Prune  $\widehat{\mathcal{T}}$  by merging pairs of leaf nodes whose parent node has annotation  $\leq r$ . Repeat until no further merger is possible.
2. Set  $\hat{m}(r) =$  number of leaf nodes in pruned  $\widehat{\mathcal{T}}$

**Output:**

Return the estimated intrinsic dimension as the slope of the least-squares fit to  $\log \hat{m}(r)$  versus  $\log r$ .

---

**Proposition 1.** *If  $(\mathbf{X}, \mathcal{T}, \mathbf{D})$  satisfies the Complete Linkage Condition, then using the CLUSTERDIMENSION algorithm,  $\hat{m}(r) = m(r)$  for all values of  $r$  considered.*

*Proof.* Under the Complete Linkage condition, the estimated tree structure using the CLUSTERDIMENSION algorithm,  $\widehat{\mathcal{T}}$ , will be equivalent to the true tree structure,  $\mathcal{T}$  using off-the-shelf agglomerative clustering methods for hierarchical clustering [17].

Consider a violation of this proposition, where a cluster diameter value,  $r$ , exists such that our estimated number of clusters using CLUSTERDIMENSION is greater than the minimum possible number of clusters,  $\hat{m}(r) > m(r)$ . Therefore, some alternative choice of  $m(r)$  clusters can be found such that fewer than  $\hat{m}(r)$  clusters are needed to cover the data set. But, by the Complete Linkage condition, there exists distance of at least  $> r$  between all pairs of  $\hat{m}(r)$  clusters. Therefore, to use fewer than  $\hat{m}(r)$  clusters would require a diameter larger than  $r$ . Thus, we require  $\hat{m}(r) = m(r)$  for any value of  $r$ .  $\square$

We can now state the following Theorem with respect to the estimated intrinsic dimension using the CLUSTERDIMENSION algorithm under data sets that satisfy the Complete Linkage condition.

**Theorem 3.1.** *For a triple  $(\mathbf{X}, \mathcal{T}, \mathbf{D})$  that satisfies the Complete Linkage condition, using the CLUSTERDIMENSION algorithm the estimated intrinsic dimension of the data set ( $\hat{d}$ ) converges to the true intrinsic dimension,  $\hat{d} \rightarrow d$ , as the size of the data set,  $N$  grows.*

*Proof.* The proof of this theorem is obvious given Proposition 1.  $\square$

Although we only prove performance of the CLUSTERDIMENSION algorithm under the Complete Linkage condition, our experiments later in this section will not require this condition on the data sets observed and will demonstrate the generality of our technique.

### 3.1.2 Computational Complexity Analysis

Critical to any intrinsic dimension estimation methodology is the necessity to estimate the dimension in feasible time. In Table 1, we review the computational complexity of each methodology. As seen in the table, we find that no

other methodology has lower computation complexity compared with the CLUSTERDIMENSION methodology. Prior methods that rely on a linear embedding of the data (*e.g.*, PCA and Box Counting approaches) are dependent on the linear embedding dimension of the data ( $d_\ell$ ) which for non-linear real-world data can approach the size of the data set,  $N$ .

Table 1: Computational Complexity of Intrinsic Dimension Estimation Algorithms (for  $N$  items with linear embedding dimension  $d_\ell$ )

Dimension Estimation Method	Computational Complexity
CLUSTERDIMENSION	$O(N^2)$
Maximum Likelihood [7]	$O(N^2)$
Box Counting [9]	$O(d_\ell N^2)$
Correlation Dimension [6]	$O(N^2)$
Minimum Spanning Trees [8]	$O(N^2 \log N)$
PCA [18]	$O(d_\ell N^2)$

### 3.2 Synthetic Datasets - Dimension Estimation

Performance of the dimension estimation techniques will be considered on a collection of fractals with known intrinsic dimension. Here we will consider a Koch Curve, the Sierpinski Triangle, and the Sierpinski Carpet as seen in Figure 2. This choice of synthetic fractals allows us to validate our dimension estimation techniques on data with known ground-truth intrinsic dimension. We compare the performance of the CLUSTERDIMENSION methodology with prior work on intrinsic dimension estimation using a Maximum Likelihood technique [7], box counting [9], the correlation dimension [6], a minimum spanning tree-based approach [8], and linear PCA [18].

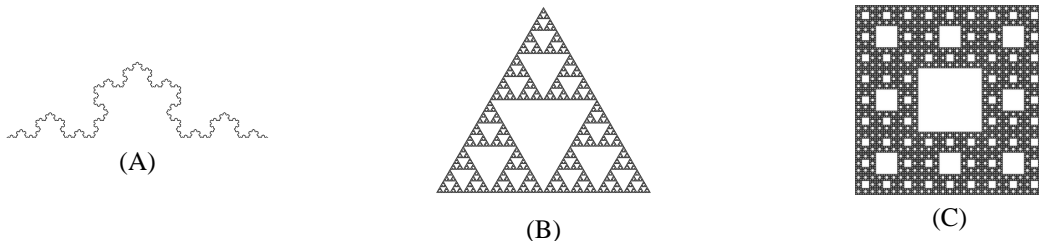


Figure 2: Fractal data sets : (A) - Koch Curve ( $d = 1.262$ ); (B) - Sierpinski Triangle ( $d = 1.584$ ); and (C) - Sierpinski Carpet ( $d = 1.893$ ).

In Table 2 we find the estimated intrinsic dimension over 10 random realizations of each of the fractal data sets, where each realization samples 750 points sampled at-random from the self-similar fractals generated with a depth of 50 self-similar iterations. The pairwise distances are found with respect to the Euclidean norm between each pair of points. We present the results in terms of both the average intrinsic dimension using the various estimation techniques, the standard deviation of the estimated intrinsic dimension ( $\sigma_d$ ), and the root mean squared error (RMSE) for the estimated intrinsic dimension. As seen in the table, the CLUSTERDIMENSION algorithm has the smallest RMSE for two of the three data sets (*i.e.*, for the Koch Curve and Sierpinski Triangle). It is only for the Sierpinski Carpet data set that our estimated dimension deviates from the true dimension slightly more than the Minimum Spanning Tree methodology, although our approach results in a significantly lower standard deviation ( $\sigma_d$ ) between the multiple estimates of intrinsic dimension. While this collection of fractals is highly structured, in the next section we demonstrate the performance of the CLUSTERDIMENSION algorithm on unstructured data.

## 4 Intrinsic Dimension-Based Clustering

Often real world data is generated from multiple processes, each of which could have a different intrinsic dimension. We now look to the problem of classifying which observed items were generated from a particular manifold. To

Table 2: Intrinsic Dimension Estimation for self-similar fractals (Koch Curve, Sierpinski Triangle, and Sierpinski Carpet) for 750 uniformly sampled points in 2-D space.

Method	Fractal Data Set								
	Koch Curve ( $d = 1.262$ )			Sierpinski Triangle ( $d = 1.584$ )			Sierpinski Carpet ( $d = 1.893$ )		
	RMSE	Mean	$\sigma_d$	RMSE	Mean	$\sigma_d$	RMSE	Mean	$\sigma_d$
CLUSTER DIMENSION	0.045	1.219	0.040	0.022	1.604	0.040	0.103	1.791	0.026
MLE [7]	0.068	1.194	0.008	0.057	1.527	0.006	0.245	1.648	0.008
Box Count. [9]	0.272	0.991	0.028	0.334	1.251	0.028	0.548	1.347	0.049
Correlation [6]	0.373	0.889	0.007	0.283	1.301	0.005	0.555	1.338	0.007
MST [8]	0.134	1.382	0.119	0.100	1.646	0.194	0.072	1.942	0.150
PCA [18]	0.738	2	0	0.416	2	0	0.107	2	0

generalize our approach, we will only consider the problem of resolving a data set drawn from a mixture of two manifolds of different intrinsic dimension.<sup>1</sup>

In contrast to prior approaches, our clustering-based methodology gives a natural partitioning of the data using dimension estimation. While prior approaches (such as box counting) would result in a combinatorial time problem, the key insight here is that our hierarchical clustering-based method automatically returns at most  $O(N)$  division points in the data set that would result in valid dimension-based clusters (*i.e.*, each interior node in the tree structure). In order to determine the interior tree node that best partitions the data set, we estimate the intrinsic dimension at each interior node (related to subset of items  $\mathcal{C}$ , with estimated *interior dimension*  $d_{in}$ ) and the fractal dimension of the dataset with the specified subset of items removed (the subset of items  $\mathbf{X} - \mathcal{C}$  with estimated *exterior dimension*  $d_{out}$ ). An example of this partitioning and intrinsic dimension estimation is found in Figure 3.



Figure 3: Examples of partitioning based on hierarchical clustering tree structure and the resulting intrinsic dimension estimation.

Our intuition is that the interior node related to the largest difference between the interior and exterior dimension is the best choice for partitioning the data set into two clusters. The full DIMENSIONPARTITION methodology is found in Algorithm 2, and builds off of the CLUSTERDIMENSION algorithm.

## 4.1 Intrinsic Dimension-Based Clustering Experiments

We test this approach on the synthetic data sets seen in Figure 4. This collection consists of three data sets with points sampled at-random from both a one-dimensional line manifold and a two-dimensional manifold (Figures 4-(A-C)), and a dataset consists of points sampled at-random from a three-dimensional swiss roll manifold intersected with a one-dimensional line manifold (Figure 4-(D)). We compare our clustering methodology with two prior approaches. The first is a standard unsupervised K-Means clustering approach [18], where we specify the algorithm to resolve two clusters in the data set. The second methodology is a modified version of the Fractal Clustering algorithm [20]. To give the Fractal Clustering approach every opportunity, we initialize this algorithm with a small subset of items with oracle classification knowledge.

<sup>1</sup>Although the methodology presented here could easily be extended to multi-class problems, we reserve this for future work.

---

**Algorithm 2** - DIMENSIONPARTITION( $\mathbf{X}, \mathbf{D}$ )

---

**Main Body:**

1. Using pairwise distances  $\mathbf{D}$ , estimate the hierarchical clustering tree  $\mathcal{T}$ .
2. Estimate the intrinsic dimension for each interior node,  $i$ , of the clustering tree  $\mathcal{T}$ . Resolving the interior dimension  $d_{in}(i)$  and the exterior dimension  $d_{out}(i)$  using the CLUSTERDIMENSION algorithm.
3. Find the interior node point with largest intrinsic dimension difference  $\hat{i} = \arg \max_i |d_{in}(i) - d_{out}(i)|$ .

**Output:**

1. Return two clusters: the first cluster contains all points clustered by interior node  $\hat{i}$ ,  $\mathcal{C}_{\hat{i}}$ , and the second cluster containing all remaining items,  $\mathbf{X} - \mathcal{C}_{\hat{i}}$ .
- 

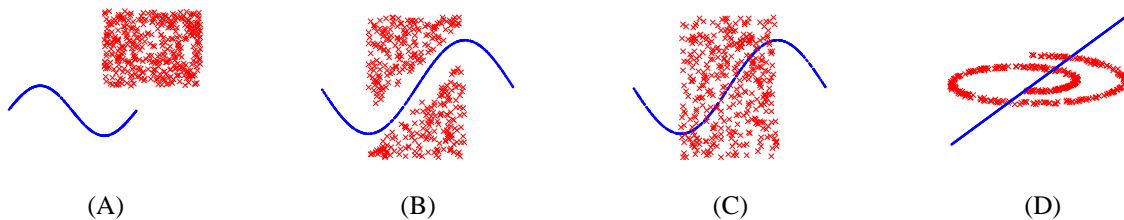


Figure 4: Data sets consisting of two manifolds of differing fractal dimension. (A) - Data set A; (B) - Data set B; (C) - Data set C; and (D) - Data set D (with a cross-section of three dimensional data shown here).

Table 3: Classification Error with respect to the four synthetic data sets with multiple manifolds (averaged across 10 realizations).

Method	Data Set			
	A	B	C	D
DIMENSIONPARTITION	0.33%	4.00%	8.00%	2.26%
K-Means [18]	1.27%	47.07%	48.80%	48.14%
Oracle-Based Fractal Clustering [20]	21.23%	20.69%	20.08%	22.12%

The classification error (*i.e.*, the percentage of items that are incorrectly classified) on these four data sets is found in Table 3. Compared with the two prior methods, the DIMENSIONPARTITION methodology resolves the two manifolds with a significantly lower error rate. Large error rates for the prior methods are seen in particular when there is no clear separation between the two manifolds (*i.e.*, data sets B, C, and D), even with the Fractal Clustering approach given every opportunity through ground-truth initialization. Examples of the classification performance using the DIMENSIONPARTITION method can be seen in Figure 5.

#### 4.1.1 Dimension-Based Outlier Detection Experiments

One application for dimension-based clustering is distinguishing between data that lies on a low dimensional manifold and high dimension outlier noise that corrupts a data set. To test our clustering approach, we use a set of real-world gene microarray data [21] and corrupt 30% of the genes with Gaussian white noise, with the remaining 70% of the genes uncorrupted. This simulates the occurrence of outliers commonly found in gene microarray experiments [1]. Across 10 random realizations, we test both the DIMENSIONPARTITION methodology and competing methodologies (*i.e.*, standard K-Means and Oracle-based Fractal Clustering) on the ability to distinguish between observed microarray genes and the noise corrupted genes (with standard deviation  $\sigma$ ), with the average classification error rates specified in Table 4. The table shows that our clustering-based approach is able to classify the outliers in the real-world data set with significantly fewer errors compared with both the K-Means or Oracle-Based Fractal Clustering approach. As expected, as the outlier noise power grows, the ability to classify outliers increases considerably.

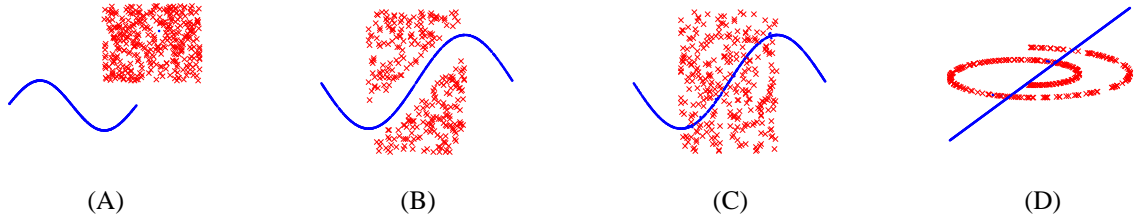


Figure 5: Estimated manifold classification using the DIMENSIONPARTITION methodology. (A) - Clusters estimated from data set A; (B) - Clusters estimated from data set B; (C) - Clusters estimated from data set C; (D) - Clusters estimated from data set D (with a cross-section of three dimensional data shown here).

Table 4: Classification error (averaged across 10 realizations) with respect to the real-world gene microarray data with 30% of the genes corrupted by additive Gaussian white noise with standard deviation,  $\sigma$ .

Method	Noise Standard Deviation			
	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 8$
DIMENSIONPARTITION	26.13%	18.82%	13.09%	11.05%
K-Means [18]	31.03%	30.78%	29.82%	30.16%
Oracle-Based Fractal Clustering [20]	35.66%	30.93%	33.13%	30.83%

## 5 Conclusions

The problem of estimating the intrinsic dimension from pairwise distances is a critical issue for a wide range of real world problems. By exploiting inherent clustering in the data, we developed an accurate and computationally efficient intrinsic dimension estimation methodology. In addition, our clustering-based methodology allows for a natural partitioning of data points that lie on separate manifolds. Experiments on both synthetic and real-world data shows the improvements of our techniques over prior methodologies. In future work we look to examine multi-class classification using our dimension-based clustering, and intrinsic dimension estimation using incomplete pairwise distances.

## 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)*, vol. 2, 2005.
- [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. Carter, R. Raich, and A. 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] P. Grassberger and I. 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. Martinez, R. D. Tenreiro, and L. J. Roy, “Hausdorff Dimension from the Minimal Spanning Tree,” in *Physical Review E*, vol. 47, no. 1, January 1993, pp. 735–738.
- [9] B. B. Mandelbrot, “How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension,” in *Science*, vol. 156, 1967, pp. 636–638.
- [10] J. Theiler., “Estimating Fractal Dimension,” in *Journal of the Optical Society of America*, vol. 7, 1990, pp. 1055–1073.
- [11] J. Ernst, G. J. Nau, and Z. Bar-Joseph, “Clustering Short Time Series Gene Expression Data,” in *Bioinformatics*, vol. 21, 2005, pp. 159–168.
- [12] H. Zheng, E. K. Lua, M. Pias, and T. G. Griffin, “Internet routing policies and round-trip-times,” in *Proceedings of the Passive and Active Measurement Workshop (PAM)*, 2005.
- [13] J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” in *Science*, vol. 290, Dec 2000, pp. 2319–2323.
- [14] S. Roweis and L. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” in *Science*, vol. 290, Dec 2000, pp. 2323–2326.
- [15] C.-T. C. J. Feng, W.-C. Lin, “Fractional Box-Counting Approach to Fractal Dimension Estimation,” in *International Conference on Pattern Recognition (ICPR)*, vol. 2, Vienna, Austria, August 1996, pp. 854–858.
- [16] P. J. Verveer and R. P. W. Duin, “An Evaluation of Intrinsic Dimensionality Estimators,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 81–86, January 1995.
- [17] 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.
- [18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.
- [19] M. Balcan and P. Gupta, “Robust Hierarchical Clustering,” in *Proceedings of the Conference on Learning Theory (COLT)*, July 2010.
- [20] D. Barbar and P. Chen, “Using Self-Similarity to Cluster Large Data Sets,” *Data Mining and Knowledge Discovery*, vol. 7, pp. 123–152, 2003.
- [21] J. DeRisi, V. Iyer, and P. Brown, “Exploring the metabolic and genetic control of gene expression on a genomic scale,” in *Science*, vol. 278, October 1997, pp. 680–686.