Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2015, Article ID 837185, 10 pages
http://dx.doi.org/10.1155/2015/837185

Research Article

Dimension Estimation Using Weighted Correlation Dimension Method

Yuanhong Liu,[1,2] Zhiwei Yu,[1] Ming Zeng,[1] and Shun Wang[1]

[1]*Space Control and Inertial Technology Research Center, Harbin Institute of Technology, Harbin 150001, China*
[2]*School of Information and Electrical Engineering, Northeast Petroleum University, Daqing 163318, China*

Correspondence should be addressed to Yuanhong Liu; 39522496@qq.com

Received 24 September 2014; Revised 29 December 2014; Accepted 29 December 2014

Academic Editor: Luca Guerrini

Dimension reduction is an important tool for feature extraction and has been widely used in many fields including image processing, discrete-time systems, and fault diagnosis. As a key parameter of the dimension reduction, intrinsic dimension represents the smallest number of variables which is used to describe a complete dataset. Among all the dimension estimation methods, correlation dimension (CD) method is one of the most popular ones, which always assumes that the effect of every point on the intrinsic dimension estimation is identical. However, it is different when the distribution of a dataset is nonuniform. Intrinsic dimension estimated by the high density area is more reliable than the ones estimated by the low density or boundary area. In this paper, a novel weighted correlation dimension (WCD) approach is proposed. The vertex degree of an undirected graph is invoked to measure the contribution of each point to the intrinsic dimension estimation. In order to improve the adaptability of WCD estimation, k-means clustering algorithm is adopted to adaptively select the linear portion of the log-log sequence $(\log \delta_k, \log C(n, \delta_k))$. Various factors that affect the performance of WCD are studied. Experiments on synthetic and real datasets show the validity and the advantages of the development of technique.

1. Introduction

Many engineering applications are difficult to be analyzed by traditional methods owing to the existence of high dimensional signals, such as face recognition [1–3], nonlinear dynamic systems [4, 5], and fault diagnosis. Therefore, a qualified dimension reduction for the high dimension signals is necessary before further proceeding.

Currently, considerable attention has been paid to the dimension reduction and many techniques have been reported [6, 7]. They can be toughly divided into two groups: linear methods and nonlinear methods. Principal component analysis (PCA) [8], local discriminant analysis (LDA), local preserving projections (LPP), and multidimensional scaling (MDS) are the classical linear methods, in which the original space is uniformly assumed to be linear and the raw data can be directly mapped into a lower dimension space. Classical nonlinear methods such as isometric mapping (Isomap), locally linear embedding (LLE) [9], Laplacian eigenmaps (LE), local tangent space alignment (LTSA), Hessian locally linear embedding (HLLE), and diffusion maps (DM) all regard the dataset as being locally homeomorphic to R^n and the local geometric approximation of the high dimensional space is preserved in low one.

For dimension reduction, one key is to choose proper intrinsic dimension. The lower intrinsic dimension estimation may lose significant information, whereas the higher one may leave too much redundant information, increasing amount of calculation and obscuring the important features. Recently, intrinsic dimension estimation methods have attracted plenty of concerns [10–16]. Usually they can be categorized into three classes, projection approach [17], probabilistic approach, and geometric approach. For projection approach, the first step is to extract a low-dimensional representation from a high-dimensional space; then the representation is analyzed and the dimension is estimated

by PCA, factor analysis, or MDS. The classical probabilistic approach is maximum likelihood estimate (MLE) [18], which estimates the probability distribution of a dataset first, and then the intrinsic dimension is estimated by maximum likelihood method. The accuracy of intrinsic dimension completely depends on the estimation of the probability distribution. The geometric approach includes geodesic minimal spanning tree (GMST) and fractal method. GMST simply constructs a minimal spanning tree sequence [19] using geodesic edge matrix and estimates the intrinsic dimension by the overall lengths of MST. GMST is a global method which does not require estimating the multivariate density of the dataset, but the drawback of GMST is the restriction to isometric embeddings. Fractal dimension [20, 21] is a statistical index of complexity of a dataset, which is commonly calculated by box-counting method [22–24] and CD method [25, 26].

In this paper, a WCD method is presented to improve the accuracy of CD method. The remainder of this paper is organized as follows. Section 2 presents a review of previous work on dimension estimation. In Section 3, theoretical analysis of WCD estimation is conducted. Section 4 thoroughly analyzes the influence of various factors on WCD by experiments. In Section 5, experiments on synthetic and real world datasets are used to confirm the effectiveness of WCD. Finally, conclusion is drawn in Section 6.

2. Previous Work on Dimension Estimation

Informally, intrinsic dimension of a dataset is the minimum number of independent variables that can completely describe a dataset and it can be used to measure complexity of a dataset. The smaller intrinsic dimension indicates a simpler dataset and vice versa. The accurate estimator of intrinsic dimension is useful to improve the performance of dimension reduction methods and to extract features.

A detailed review of intrinsic dimension estimation methods can be found in [16], which summarised almost all the typical intrinsic dimension estimation methods so far, including Fukunaga-Olsen's method, near neighbor methods, TRN-based methods, projection techniques, multidimensional scaling methods, and fractal-based methods. Recently some new intrinsic dimension estimation methods have been presented, such as minimal cover method [27], axiomatic method [28], packing number method [29], and expected absolute projection (EAP) method [30]. Each method has its own characteristic and, therefore, can only suit different datasets.

Fractal methods are a powerful tool to estimate the intrinsic dimension. Among the existing fractal methods, Hausdorff dimension method, box-counting dimension method, and CD method are the most representative ones. Further research on the fractal methods refers to [31].

Hausdorff dimension is the basis of fractal dimension, which is derived from Hausdorff measure. To proceed further, the Hausdorff measure [32] is firstly introduced.

Definition 1 (Hausdorff measure). Let (X, ρ) be a metric space. For any subset $U \subset X$, one defines a nonnegative function

$$H_\delta^D(X) = \inf\left\{\sum_{i\in\mathbb{N}} \operatorname{diam}(U_i)^D : X \subseteq \bigcup_{i\in\mathbb{N}} U_i;\ U_i \text{ open}, \operatorname{diam}(U_i) < \delta,\ \forall i \in \mathbb{N}\right\}, \tag{1}$$

where $\operatorname{diam}(U) = \sup\{\rho(x, y) : x, y \in U\}$ represents diameter of subset U. D dimension Hausdorff measure of X can be defined as

$$H^D(X) = \lim_{\delta\to 0} H_\delta^D(X). \tag{2}$$

Definition 2 (Hausdorff dimension). Hausdorff dimension of a set X in a metric space (X, ρ) is

$$\begin{aligned}\dim_H(X) &= \inf\left\{D : H^D(X) = 0\right\}\\ &= \sup\left\{D : H^D(X) = \infty\right\}.\end{aligned} \tag{3}$$

Hence, Hausdorff dimension D is a critical value of Hausdorff measure from ∞ to 0. Hausdorff dimension presents a perfect theoretical framework for dimension estimation, from which many new fractal dimension estimation methods can be derived. But Hausdorff dimension is difficult for dimension estimation in practice. The box-counting dimension derived from Hausdorff dimension simplifies calculation complexity of Hausdorff dimension.

Definition 3 (box-counting dimension). For a totally bounded set X in a metric space, let $N_\delta(X)$ be the minimal number of balls with scale δ that cover X. The box-counting dimension is then [33]

$$\dim_{\mathrm{BC}}(X) = \lim_{\delta\to 0} \frac{\log N_\delta(X)}{-\log\delta}, \tag{4}$$

and the necessary condition for the existence of limit is that $N_\delta(S)$ is proportional to δ:

$$N_\delta(X) = c \cdot \delta^{(-D)}, \tag{5}$$

where c is a constant. Take the logarithm on (5)

$$\log N_\delta(X) = \log c - D\log\delta. \tag{6}$$

The box-counting dimension D can be expressed as

$$D = \frac{\log c}{\log\delta} - \frac{\log N_\delta(X)}{\log\delta}, \tag{7}$$

and according to (7), in order to obtain a good estimate of D, $\log c/\log\delta$ must approach 0. In practice, affected by sample size or the value of δ, $\log c/\log\delta$ cannot be completely eliminated. Usually, box-counting dimension is determined by calculating a slope of the linear part of curve fitted by $\log N_\delta(X)$ versus $\log\delta$.

Although box-counting method is simpler in calculation compared with Hausdorff method, it still has more computation complexity than CD method [32]. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a dataset, $X \in R^{D\times n}$. Correlation integral $C(n, \delta)$ [34] can be defined as

$$C(n,\delta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I\left(\left\|x_i - x_j\right\| < \delta\right), \quad (8)$$

where $\|x_i - x_j\|$ can be any metric between data points x_i and x_j. $I(\cdot)$ is Heaviside function, which is 1 if the condition is met and otherwise 0. $C(n, \delta)$ is a statistical average of distances less than δ. It can also be written

$$C(n,\delta) = \frac{\text{distances less than } \delta}{\text{distances altogether}}. \quad (9)$$

The CD is defined as

$$D = \lim_{\delta \to 0} \frac{\log C(n,\delta)}{-\log \delta}, \quad (10)$$

although (7) and (10) are the same form of the formula, their calculation process is completely different. The numerator of CD method represents a global bulk with scale δ; however, the numerator of box-counting method stands for the minimum number of hyperspheres with scale δ that covers the dataset. Note that (10) cannot be directly applied to obtain CD in practice. A commonly used scheme is to calculate the slope of a curve, which indicates the relationship of $\log C(n, \delta)$ and $\log \delta$. Let $(\log C(n, \log \delta_1), \log \delta_1)$ and $(\log C(n, \delta_2), \log \delta_2)$ denote any two points of curve, respectively; the slope is then defined as

$$D = \frac{\log C(n,\delta_2) - \log C(n,\delta_1)}{\log \delta_2 - \log \delta_1}, \quad (11)$$

and the accuracy of CD method is much dependent on the choice of δ_1 and δ_2. To get high accurate CD, the linear portion of the log-log $(\log \delta_k, \log C(n, \delta_k))$ sequence is selected and a new straight line is then fitted by the linear portion.

3. Theoretical WCD Estimation

3.1. Analysis of WCD Estimation. From a geometric point of view, an object's bulk is directly related to the dimension power of its scale δ [31]. For example, a straight line length is one power of scale. The area of a circle is two powers of scale. The relationship between the bulk and the δ can be described as

$$\text{bulk} \sim \delta^{\text{dimension}}, \quad (12)$$

where the bulk can be any metric like a volume, area, or mass. Although many notions of bulk are possible, a good quantity for bulk function $\beta_{X_j}(\delta)$ is defined in CD method [31]:

$$\beta_{X_j}(\delta) \approx \frac{1}{n-1} \sum_{i=1, i\neq j}^{n} I\left(\left\|x_i - x_j\right\| < \delta\right), \quad (13)$$

and (13) indicates that the local bulk is denoted by the number of points falling into the hypersphere with scale δ at center x_j. It is noted that $i = j$ should be excluded, which implies that the denominator is $n - 1$ rather than n. Since $\beta_{X_j}(\delta)$ is a local bulk, some averaging method should be used for the global bulk. In CD method, the algebraic average is used:

$$C(n,\delta) = \frac{1}{n} \sum_{j=1}^{n} \beta_{X_j}(\delta), \quad (14)$$

where $C(n, \delta)$ is correlation integral, that is, global bulk.

For the uniform dataset, a good result can be obtained by algebraic average for correlation integral $C(n, \delta)$. However, for the nonuniform dataset, it is unreasonable to treat every point equally due to the fact that the local bulk $\beta_{X_j}(\delta)$ is different at different point. Here, a developed weighted bulk approach could be considered for global bulk, that is, treating each local bulk with different weights for global bulk; then the global bulk can be described as

$$C(n,\delta) = \sum_{j=1}^{n} W(j)\, \beta_{X_j}(\delta), \quad (15)$$

where W is the weighted vector.

Local bulk calculated at three cases including high dense points, sparse points, and boundary points is shown in Figure 1. Without considering the noise points, it is obvious that the local bulks estimated at high dense area are more reliable than the other two cases. It is natural for us to increase the weights of high dense area and simultaneously decrease the ones of low dense area and boundary area for dimension estimation. So accurate estimation of the data distribution is important, and there are many methods estimating the distribution of dataset, such as the probability distribution estimation methods and the boundary detection methods. In this paper, the vertex degree of an undirected graph is used to measure the distribution of a dataset, upon which a novel and simple WCD method is then proposed to improve the performance of CD method. If the vertex degree is big, the area around the vertex is dense; otherwise it is a sparse point or a boundary point. Moreover, vertex degree can reflect the credibility of the local bulk estimated. It is reasonable to regard the vertex degree as a weight of the local bulk. Twenty points are marked by vertex degree method in the dataset in Figure 2, in which ten squares represent the biggest vertexes degree and ten circles indicate the smallest ones. We can see that the density area and the sparse or boundary points are distinguished correctly. Therefore, the WCD method is more accurate for the intrinsic dimension than CD method. The specific description of WCD method is shown in Algorithm 1.

3.2. Selecting the Linear Portion of the log-log Sequence. Selecting different portion of the log-log sequence to calculate the slope will lead to different precision of CD estimation. A log-log plot drawn by the log-log sequence $(\log \delta_k, \log C(n, \delta_k))$ is shown in Figure 3 and it can be divided into three portions, the low portion, the middle portion, and the upper portion. In the low portion, the scale δ of the

Input: Signal dataset X.
Output: Intrinsic dimension D.
(1) Normalize the dataset X between 0 and 1, then the distance matrix W_1 can be constructed by $W_1(j,i) = \|x_i - x_j\|$.
(2) Construct the similarity matrix $W_2(j,i) = \exp(-||x_i - x_j||/2\theta^2)$. Where θ is the variance of the dataset. The vertex degree, that is the weighted vector is defined as $W(j) = \sum_{i=1}^{n} W_2(j,i)$.
(3) The scale sequences $(\delta_1, \delta_2, \ldots, \delta_m)$ are computed by $\delta_k = \min(W_1) + k((\max(W_1) - \min(W_1))/m)$, $k = 1, 2, \ldots, m$. Where m is the number of the scale δ.
(4) Compute the local bulk $\beta_j(\delta_k)$ at point x_j. $\beta_{X_j}(\delta_k) \approx (1/(n-1)) \sum_{i=1, i\neq j}^{n} I(\|x_i - x_j\| < \delta)$, $j = 1, 2, \ldots, n$.
(5) In the scale δ_k, the global bulk is computed $C(n, \delta_k) = \sum_{j=1}^{n} W(j)\beta_j(\delta_k)$.
(6) The linear part of the log-log sequence $(\log \delta_k, \log C(n, \delta_k))$ is selected by *k-means* method and a curve is fitted using linear part by the linear least square method.
(7) Correlation dimension is calculated by the slope of the curve. $D = (\log C(n, \delta_2) - \log C(n, \delta_1))/(\log \delta_2 - \log \delta_1)$.

Algorithm 1: The calculating procedure of WCD.

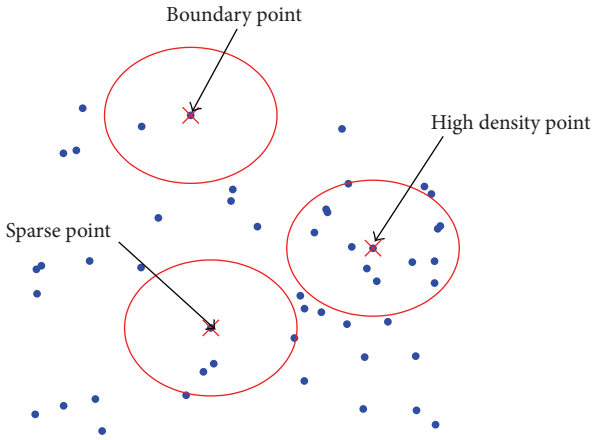

Figure 1: Different local bulks at three different points.

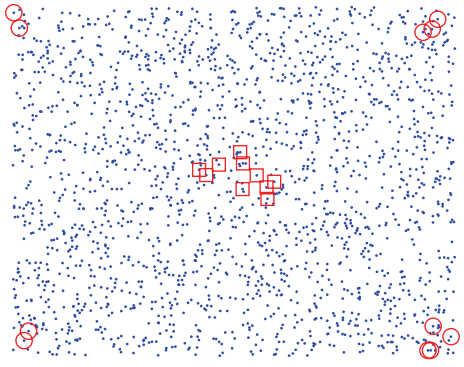

Figure 2: The indication of falling into the circle at different location.

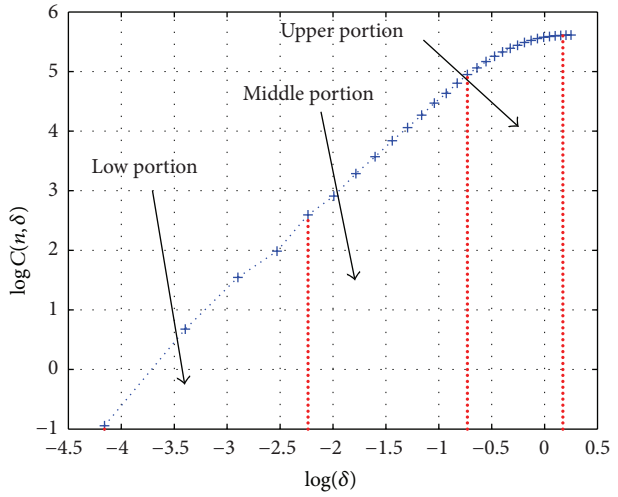

Figure 3: log-log plot for computation of CD.

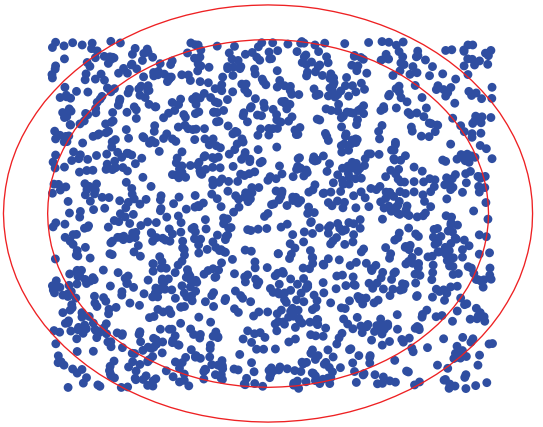

Figure 4: Bending explanation for the upper portion.

hyperspheres is small and only few points fall into the hyperspheres. So very small noise points can cause great error, which is the reason that the low portion occurs fluctuating phenomenon. Besides, in the upper portion, where the scales δ of the hyperspheres are larger than a specific value, the number falling into the hyperspheres will not increase. The scattering plot of the dataset is shown in Figure 4. This is the reason that the upper portion bends down and approaches a plateau. Usually, the middle portion is linear which is perfect to estimate CD of a dataset. In order to minimize the error caused by nonlinearity, we should choose small points from the log-log sequence $(\log \delta_k, \log C(n, \delta_k))$ and try our best

to choose the linear portion of the sequence. However, to maximize our sample size, we want to include as many points as possible. How can we accurately choose the linear points from the log-log sequence? For the obvious characteristics of the three portions of the sequence, we can use the k-means clustering method to decide which pairs of the log-log sequence should be used for CD estimation. k-means clustering method aims to partition the log-log sequence into three categories by minimizing the objective function

$$\arg\min_{S} \sum_{i=1}^{c} \sum_{y_j \in s_i} \left\| y_j - \mu_i \right\|^2, \tag{16}$$

where y_i is the pair in the sequence $(\log \delta_k, \log C(n, \delta_k))$. $c = 3$ represents three categories, including the low portion, the middle portion, and the upper portion, respectively. S_1, S_2, S_3 are the number of the three categories. μ_i is the mean of c_i. Hence, those points that belong to c_2 are chosen to fit a curve by the least squares method and used to estimate the CD. The most important factor of the k-means method is the initial value of μ_i. In this paper, the curve is divided equally into three portions and the mean of each portion represents the initial value of μ_i.

3.3. Complexity Analysis of WCD Method. In this section, the computational complexity of WCD method is investigated and compared with CD method. From the whole calculation process, we can see that the local bulk of WCD method costs more calculations than that of CD method. For the analysis, we assume that the sample size is n. The calculation of a local bulk $\beta_j(\delta)$ at point x_j with scale δ requires $n - 1$ operations and the complexity is $O(n - 1)$. There are n local bulks that should be calculated, so all of the complexity is $O((n - 1)^n)$. However, the CD method is only $O((n - 1)!)$. In addition, compared with CD method, vertex degree need be calculated in WCD method and the complexity cannot be ignored, when the sample size is huge. All these seem that the computational complexity of WCD method is much higher than CD method. But actually, it is unnecessary to calculate all local bulks of the dataset for WCD method. We can only use very few points to estimate the local bulks and can also get a high accuracy result. The computational complexity of WCD method is almost the same as CD method and this can be proved by the following experiments.

4. Experimental Study

There are many factors affecting the results using WCD method, including the sample size, the intrinsic dimension, selecting of linear portion of log-log sequence, number of local bulks used for correlation integral $C(n, \delta)$, and selecting scales. In our experiments, samples with different dimensions and sample sizes are generated by MATLAB randn function. Each sample is independent of Gauss distribution. The performance of WCD method is compared with CD method and the various factors are analyzed. Correlation dimensions are depicted in Figures 5(a), 5(b), and 5(c) for both WCD and CD methods, respectively, with three different sample sizes. Specifically, only sample sizes of 100, 200, and 500 and intrinsic dimensions of 3, 5, and 8 are used to plot. It is similar for other sample sizes and intrinsic dimensions. For each plot in Figure 5, the horizontal axis indicates the number of local bulks β, whose maximum value is the same as the sample size. The vertical axis represents the actual and the estimated values (via the WCD method and the CD method) of the intrinsic dimension. Each horizontal green line represents the actual intrinsic dimension for reference. Each red dot denotes the intrinsic dimension estimated by the WCD method. Each black asterisk denotes the intrinsic dimension estimated by the CD method.

It can be well observed from Figure 5 that the intrinsic dimensions calculated by the WCD method are more accurate than the ones by the CD method. However, the front part of the curves plotted by the WCD method fluctuates frequently. This is because there are few local bulks used for intrinsic dimension estimation which lead to the fact that the result is instability. In addition, all the curves plotted by the WCD method slop downward with the number of local bulks increasing, but they still can converge to a good value. In general, the front part of the curves plotted by the WCD method is more precise than the latter part. The loss of precision is mainly caused by the data distribution. The high dense area is chosen first to calculate the local bulks by the vertex degree method, leading to the high accuracy. However, with more sparse points or boundary points being used to calculate the local bulks, the accuracy will be lost. Hence, it is inferred that the number of the points used to calculate the local bulks is one of the main factors to the intrinsic dimension estimation. This also verifies the effectiveness of our developed methods of using small high dense points to estimate the intrinsic dimension by the WCD method. Examining the curves estimated by both methods, when the samples size is fixed, the accuracy will gradually reduce with the increase of actual intrinsic dimension. The main reason is that the dataset becomes more and more sparse with the increasing intrinsic dimension in the same sample size. Observing the curves in Figures 5(a), 5(b), and 5(c), respectively, it can be seen that the accuracy of both methods tends to improve, along with the increasing sample sizes in the same actual intrinsic dimension. This is because the dataset will become dense with the increase of the sample sizes. Additionally, the selection of scales δ is also an important factor of affecting the performance of the intrinsic dimension. The smaller scales δ will be easily susceptible to noise; however, the larger scales will result in saturation phenomenon, in which the correlation integral $C(n, \delta)$ will not change with the increasing scales δ. In addition, abundance scales will inevitably increase the computational cost and the smaller number one will reduce the precision.

For the purpose of analyzing the calculation speed, we generate three dimension datasets with sample sizes from 100 to 4000 by MATLAB randn function and estimate intrinsic dimension by these four methods. The computation time of all four methods is shown in Figure 6. It reveals that the GMST method costs the most computation time, while WCD method, MLE method, and CD method cost almost the same

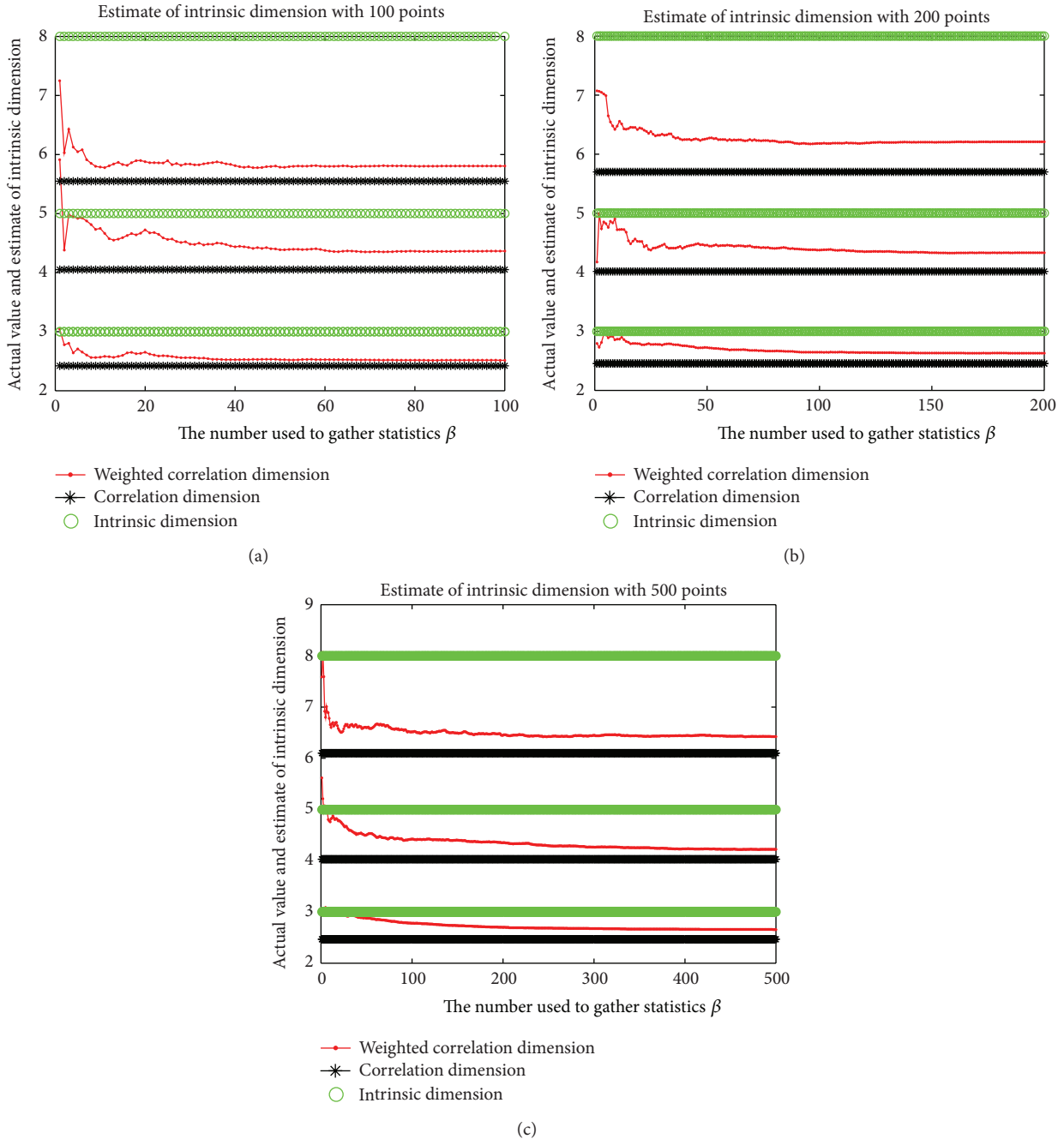

Figure 5: Estimated and actual intrinsic dimension for datasets on different sample size.

calculation time. But the computation speed of WCD method will obviously slow down with the increase of the local bulks.

5. Empirical Results

In order to validate the proposed method, WCD method is used to estimate the intrinsic dimension of two kinds of datasets (the synthetic datasets and the real world datasets). Moreover, the comparisons with geodesic minimum spanning tree (GMST), correlation dimension (CD), and maximum likelihood estimation (MLE) are also performed to further the advantage of our developed findings in this paper.

5.1. Synthetic Datasets. In this subsection, two synthetic datasets (Koch curve and S-curve) are firstly investigated. The sample sizes of the two datasets are 2000, respectively, and plots are shown in Figures 7 and 8. The dimensions estimated by all methods are listed in Table 1. Koch curve originates from a line whose middle segment is repeatedly replaced by an equilateral triangle. If we use a tool whose dimension is less than 1 to measure Koch curve, its Hausdorff measure is

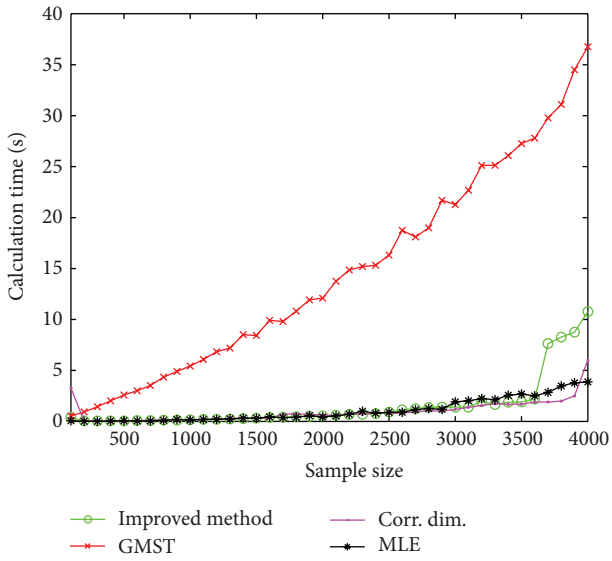

Figure 6: Calculation time comparison.

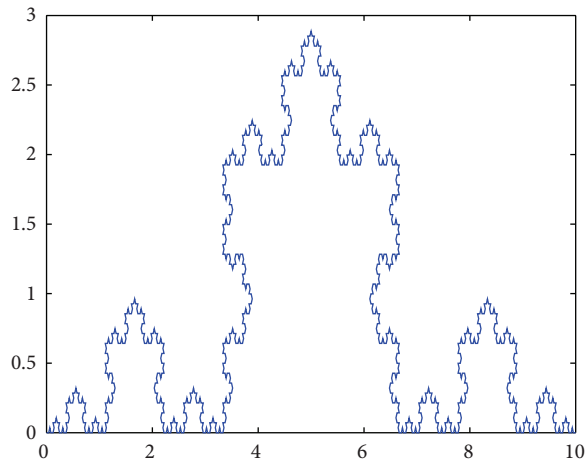

Figure 7: Koch curve dataset.

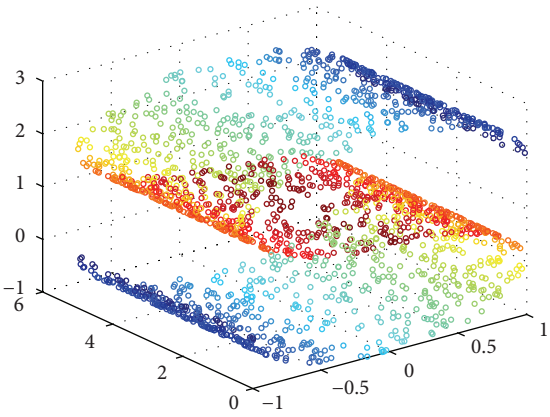

Figure 8: S-curve dataset.

Table 1: Intrinsic dimension estimation of synthetic datasets with different methods.

Dataset	Data dim.	Sample size	Improved corr. dim.	Corr. dim.	MLE	GMST
Koch curve	1-2	1025	1.3801	1.1206	1.7481	1.3424
S-curve	2	2000	2.0005	1.9567	1.9865	2.0994

inf. If we use two dimensions to measure it, its Hausdorff measure is 0. So the intrinsic dimension of Koch curve is between 1 and 2 and the dimension estimated by the four methods falls into this range. Moreover, the data points in S-curve dataset are contained in a curved surface in three-dimensional space, so the intrinsic dimension of S-curve dataset is 2. The obtained results show that all the considered methods have high accuracy, in which the developed one in this paper is the most optimal.

5.2. Real Datasets. Following a similar process in 5.1, another three real datasets (the laser generated data, the Ikeda map, and the *Hénon* map) will be analyzed in this subsection, where the specific explanations of the considered real datasets are illustrated as follows.

5.2.1. Laser Generated Data. The data were recorded from a far-infrared-laser in a chaotic state [4], formed by 1000 samples, and the attractor dimension is approximately 2.26. The plot is shown in Figure 9.

5.2.2. Ikeda Map. Ikeda map [31] is a complex map, which is defined by

$$z_{(n+1)} = a + Rz_{(n)} \exp\left(i\left(\phi - \frac{p}{\left(1 + \left|z_{(n)}\right|^2\right)}\right)\right). \quad (17)$$

Ikeda map is derived from a model of the plane-wave interactivity field in an optical ring laser. It is iterated many times, and the points $[\mathrm{Re}(zn), \mathrm{Im}(zn)]$ are plotted for $n = 2000$. Here, $a = 1.0$, $R = 0.9$, $\phi = 0.4$, and $p = 6$. The intrinsic dimension of this attractor is approximately 1.7. The visualization of the map is shown in Figure 10.

5.2.3. Hénon Map. *Hénon* map [31] is usually cast as an equation of the form

$$X_{(n+1)} = 1.0 - ax_{(n)}^2 + Y_{(n)};$$
$$Y_{(n+1)} = bx_{(n)}, \quad (18)$$

with $a = 1.4$ and $b = 0.3$, and gives an attractor with intrinsic dimension of approximately 1.3. The plot of *Hénon* map dataset for $n = 2000$ is shown in Figure 11.

For estimating the intrinsic dimension of laser generated data, phase space is reconstructed by delay-time embedding technology. Although Takens has proved that original state space of a dynamical system will be reconstructed, as long

Table 2: Intrinsic dimension estimation of real datasets with different methods.

Dataset	Data dim.	Sample size	Improved corr. dim.	Corr. dim.	MLE	GMST
Laser generated data	2.06	1000	2.1027	1.9379	2.7124	1.9842
Ikeda map	1.7	2000	1.6889	1.6348	1.8010	1.8082
Hénon map	1.3	2000	1.3584	1.1989	1.5206	1.1962

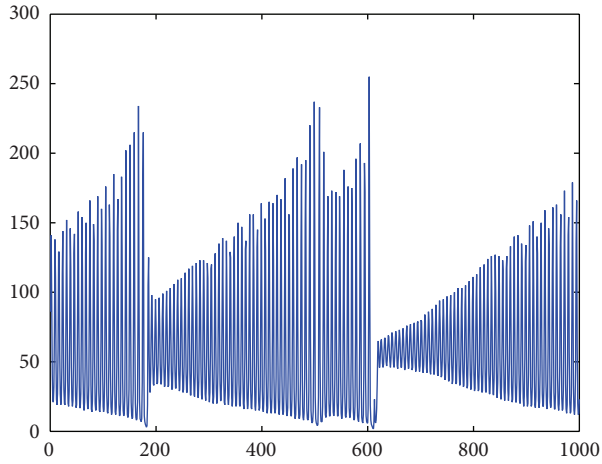

Figure 9: Dataset of laser generated data.

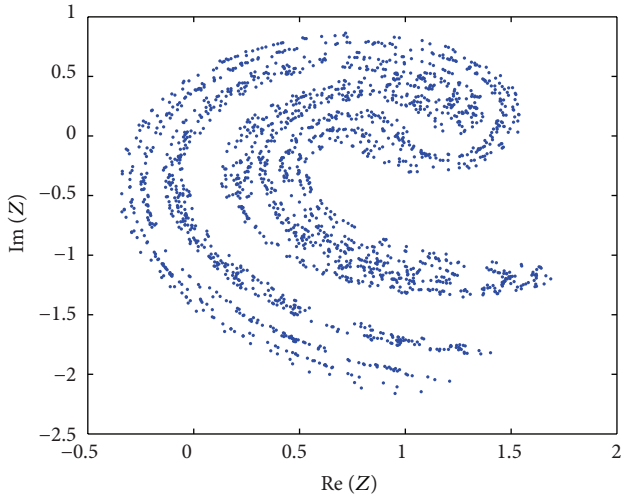

Figure 10: Dataset of Ikeda map.

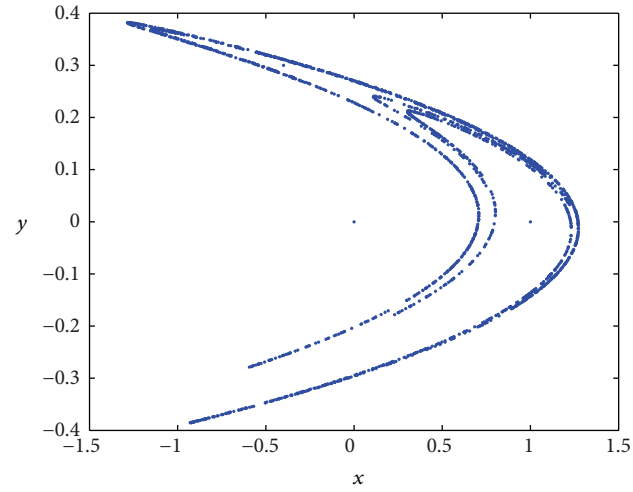

Figure 11: Dataset of *Hénon* map.

as $m > 2D + 1$, where m is the embedding dimension and D denotes the intrinsic dimension of the attractor, it is nontrivial to choose the embedding parameters. If the product $(m-1)\tau$ is too large, then the reconstructed vector will be effectively decorrelated in phase space, which lead to a larger dimension estimation. When the product $(m-1)\tau$ is too small, the reconstructed vector becomes effectively redundant, which will lead to a smaller dimension estimation. In order to compare the index with [4], we select embedding dimension $m = 5$, delay time $\tau = 10$. Furthermore, the dimension of Ikeda map and *Hénon* map is estimated directly by dimension estimation method, which avoids selecting m and τ. From Figures 10 and 11, we note that the thinner attractor is the lower dimension. The results are listed in Table 2, from which we can infer that the WCD method is also effective on the real datasets.

6. Conclusion

When the distribution of a dataset is nonuniform, the CD method for intrinsic dimension suffers from large bias. To address this issue, the WCD method has been proposed with an optimized weighted vector determined by the vertex degree. The influencing factors of the WCD method have also been comprehensively analyzed, including the sample size, the selecting of the linear portion of the log-log sequence, the number of local bulks used for correlation integral $C(n, \delta)$, and the selecting scales. The WCD method is validated by experiments on synthetic datasets and real world datasets.

Compared with the CD method, the main drawback of the WCD method is that the speed of the computation will slow down, when a lot of local bulks $\beta_X(\delta)$ are calculated. But the experiments indicate that it is unnecessary to calculate all the local bulks of the dataset and only a few points in the high dense area of the dataset used to calculate will also obtain a good result. From above experiments, it can be seen that the computational complexity of WCD method is almost the same as CD method, when the local bulks are less than 3500. Moreover, the density estimation of a dataset by vertex degree is only applicable to a single distribution. when the dataset is multiple distribution, WCD method will fail, which should be further studied.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] C. K. Loo, A. Samraj, and G. C. Lee, "Evaluation of methods for estimating fractal dimension in motor imagery-based brain computer interface," *Discrete Dynamics in Nature and Society*, vol. 2011, Article ID 724697, 8 pages, 2011.

[2] A. T. B. Jin, P. Y. Han, and L. H. Siong, "Eigenvector weighting function in face recognition," *Discrete Dynamics in Nature and Society*, vol. 2011, Article ID 521935, 15 pages, 2011.

[3] S. Wang, T. Shi, M. Zeng, L. Zhang, F. E. Alsaadi, and T. Hayat, "New results on robust finite-time boundedness of uncertain switched neural networks with time-varying delays," *Neurocomputing*, vol. 151, part 1, pp. 522–530, 2015.

[4] F. Camastra and M. Filippone, "A comparative evaluation of nonlinear dynamics methods for time series prediction," *Neural Computing and Applications*, vol. 18, no. 8, pp. 1021–1029, 2009.

[5] J. Luo, G. Li, and H. Liu, "Linear control of fractional-order financial chaotic systems with input saturation," *Discrete Dynamics in Nature and Society*, vol. 2014, Article ID 802429, 8 pages, 2014.

[6] K. M. Carter, R. Raich, and I. Hero, "On local intrinsic dimension estimation and its applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 650–663, 2010.

[7] S. Yin, G. Wang, and X. Yang, "Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data," *International Journal of Systems Science*, vol. 45, no. 7, pp. 1375–1382, 2014.

[8] L. M. Elshenawy, S. Yin, A. S. Naik, and S. X. Ding, "Efficient recursive principal component analysis algorithms for process monitoring," *Industrial and Engineering Chemistry Research*, vol. 49, no. 1, pp. 252–259, 2010.

[9] E. E. Abusham and E. K. Wong, "Locally linear discriminate embedding for face recognition," *Discrete Dynamics in Nature and Society*, vol. 2009, Article ID 916382, 8 pages, 2009.

[10] S. Samudrala, K. Rajanr, and B. Ganapathysubramanian, "Data dimensionality reduction in materials science," in *Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application*, vol. 1, pp. 97–98, Elsevier Science, 2013.

[11] 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, 2010.

[12] L. Liao, Y. Zhang, S. J. Maybank, and Z. Liu, "Intrinsic dimension estimation via nearest constrained subspace classifier," *Pattern Recognition*, vol. 47, no. 3, pp. 1485–1493, 2014.

[13] R. Heylen and P. Scheunders, "Hyperspectral intrinsic dimensionality estimation with nearest-neighbor distance ratios," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 570–579, 2013.

[14] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6418–6428, 2014.

[15] S. Ding, S. Yin, K. Peng, H. Hao, and B. Shen, "A novel scheme for key performance indicator prediction and diagnosis with application to an industrial hot strip mill," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2239–2247, 2012.

[16] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern Recognition*, vol. 36, no. 12, pp. 2945–2954, 2003.

[17] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 779–785, 1994.

[18] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS '04)*, pp. 1092–1106, Vancouver, Canada, 2004.

[19] J. He, L. Ding, L. Jiang, Z. Li, and Q. Hu, "Intrinsic dimensionality estimation based on manifold assumption," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 740–747, 2014.

[20] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002.

[21] M. Sadikin and I. Wasito, "Fractal dimension as a data dimensionality reduction method for anomaly detection in time series," in *Proceedings of the 7th International Conference on Information & Communication Technologies (ICT '13)*, vol. 1, May 2013.

[22] Z. Feng and X. Sun, "Box-counting dimensions of fractal interpolation surfaces derived from fractal interpolation functions," *Journal of Mathematical Analysis and Applications*, vol. 412, no. 1, pp. 416–425, 2014.

[23] D. Sankar and T. Thomas, "Fractal features based on differential box counting method for the categorization of digital mammograms," *International Journal of Computer Information System and Industrial Management Applications*, vol. 2, pp. 11–19, 2010.

[24] Y.-C. Tzeng, K.-T. Fan, and K.-S. Chen, "A parallel differential box-counting algorithm applied to hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 2, pp. 272–276, 2012.

[25] A. Yarlagadda, J. V. R. Murthy, and M. H. M. Krishna prasad, "Estimating correlation dimension using multi layered grid and damped window model over data streams," *Procedia Technology*, vol. 10, pp. 797–804, 2013.

[26] A. R. Osborne and A. Provenzale, "Finite correlation dimension for stochastic systems with power-law spectra," *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 357–381, 1989.

[27] M. Fan, X. Zhang, S. Chen, H. Bao, and S. Maybank, "Dimension estimation of image manifolds by minimal cover approximation," *Neurocomputing*, vol. 105, no. 1, pp. 19–29, 2013.

[28] V. Pestov, "An axiomatic approach to intrinsic dimension of a dataset," *Neural Networks*, vol. 21, no. 2-3, pp. 204–213, 2008.

[29] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Proceedings of the 16th Annual Neural Information Processing Systems Conference (NIPS '02)*, pp. 681–688, December 2002.

[30] K. Johnsson, *Manifold dimension estimation for omics data analysis: current methods and a novel approach [M.S. thesis]*, Lund University, 2011.

[31] J. Theiler, "Estimating fractal dimension," *Journal of the Optical Society of America A*, vol. 7, no. 6, pp. 1055–1073, 1990.

[32] D. Schleicher, "Hausdorff dimension, its properties, and its surprises," *The American Mathematical Monthly*, vol. 114, no. 6, pp. 509–528, 2007.

[33] D. Mo and S. H. Huang, "Fractal-based intrinsic dimension estimation and its application in dimensionality reduction,"

IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 59–71, 2012.

[34] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," in *The Theory of Chaotic Attractors*, pp. 170–189, Springer, New York, NY, USA, 2004.