Estimation of Intrinsic Dimension using Supervised Parameter Selection Method

Haiquan Qiu¹* and Shulun Yang²*

¹School of Mathematics and Statistics, Xidian University, 266 Xinglong Section of Xifeng Road, Xi’an, P. R. China.
²School of Information Science and Technology, Northwest University, 1 Xuefu Ave., Guodu Education and Hi-Tech Industries Zone, Chang’an District, Xi’an, P. R. China.
Email: qiuhaiquan0902@163.com; 2533354922@qq.com

Abstract. In this paper, we propose a new method for estimating the intrinsic dimension of datasets. The new method uses the local information of different scales of the sample points (by adjacency matrix) to estimate the intrinsic dimension. The only parameter used in the new method is the scaling ratio \( k \), which determines the adjacency matrix of different scales. We propose a parameter selection method based on the difference of estimated dimension and the classification accuracy of projection data. Experiments on real datasets demonstrate the effectiveness of the proposed method.

1. Introduction
In practical problems, high-dimensional data usually has a low-dimensional structure, or the data is located on a low-dimensional manifold. The dimension of the low-dimensional manifold is called the intrinsic dimension of the dataset. The early works of intrinsic dimension estimation began in 1960s (Trunk [1], Bennett [2]). The existing estimation methods of intrinsic dimension can be divided into two categories: direct estimation and indirect estimation.

The direct estimation methods take into account the structure of the dataset itself. Such as PCA (Principle Component Analysis)[3], Kégle’s algorithm[4], Correlation Dimension based methods([5], [6], [7]), Graph Distance-based algorithm[8], DANCo[9]. PCA first calculates the covariance matrix between the features, and then calculates the eigenvalues of the covariance matrix and normalizes them. The number of eigenvalues that greater than a given threshold \( \theta \) is the intrinsic dimension of the dataset. The main drawback of PCA is the choose of an appropriate value for the threshold \( \theta \), and PCA usually overestimate. The initial correlation dimension of a dataset is defined as

\[
d_c = \lim_{\varepsilon \to 0} \frac{\ln C(\varepsilon)}{\ln \varepsilon}, \quad C(\varepsilon) = \lim_{n \to \infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{1}_{d(X_i, X_j) \leq \varepsilon},
\]

where \( n \) represents the number of data points, and \( d_c \) represents the correlation dimension, \( \mathbf{1}_{d(X_i, X_j) \leq \varepsilon} \) represents the indicator function. The estimation methods based on correlation dimension are based on formula (1) or its deformation. The estimated dimension obtained by these methods is usually the lower bound of the correlation dimension. DANCo[9] exploits the information both by the normalized nearest neighbor distances and by the angles computed on couples of neighboring points for intrinsic dimension estimation, but the estimation of the actual data set is not good.

Dimensionality reduction methods usually belong to the indirect dimension estimation method. These methods determine the intrinsic dimension of the dataset by using the effect of dimensionality reduction.
reduction. MDS (Multidimensional scaling methods) tend to preserve the distances between data points, the relevant techniques cf. (Bennett [2], Sammon [10], Demartines [11]). However, MDS techniques for intrinsic dimension estimation can become infeasible when dataset dimensionality is high [12]. ISOMAP (Isometric feature Mapping) [13] is a widely used method of indirect intrinsic dimension estimation based on manifold learning and MDS. The performance of ISOMAP is strongly influenced by the size of neighborhood, so the size control parameter K (or $\varepsilon$) is crucial. Other indirect methods, such as LLE (Local Linear Embedding) [14] and Laplacian Eigenmaps [15] are also classical method.

The review of intrinsic dimension estimation can be found in [12, 18]. Most intrinsic dimension estimation methods contain parameters to be determined, and the appropriateness of parameter selection directly affects the estimation effect. Usually, the parameter is specified or obtained by heuristic method. Samko [16] discussed the optimal parameter selection for ISOMAP. Karbauskaitė [17] proposed a method for select the number of neighbors of each point for the locally linear embedding algorithm.

In this paper, we propose a new estimation method of intrinsic dimension based on correlation dimension. In order to get the exact intrinsic dimension of labelled datasets, we use the classification accuracy of projection datasets in low dimensional space to determine the optimal parameters.

2. Estimation of Intrinsic Dimension

Suppose the sample data points $X_1, ..., X_n \in \mathbb{R}^D$, independently taken from distribution $F$, where $D \in \mathbb{N}$ is the ambient dimension. The metric $r$ is selected on $\mathbb{R}^D$ and the adjacency matrix is defined as follow.

**Definition 1.** Let $\varepsilon(i) = \min_{j \neq i} r(X_i, X_j)$, $i = 1, ..., n$, the adjacency matrix $A_{\varepsilon}(k \geq 1)$ is defined as

$$A_{\varepsilon}(k \geq 1) = \begin{cases} 1 & i \neq j, i, j = 1, ..., n, \\ 0 & i = j \end{cases}$$

It is sure that $A_{\varepsilon}$ is not symmetric matrix. If the sample points are regarded as vertices of a graph, then $A_{\varepsilon}$ is a special adjacency matrix of the graph. Let $B_k (k \geq 1)$ denotes the degrees of vertices in the graph, such that $B_k = (B_{k1}, ..., B_{kn})$

$$B_{ki} = \sum_{j=1}^{n} A_{\varepsilon}(j,i), \quad i = 1, ..., n.$$  

We define the ratio of $k\varepsilon$-neighborhood and $\varepsilon$-neighborhood of point $X_i$, as

$$q_{\varepsilon,i} = \frac{B_{k,i}}{B_{i,i}} = \frac{\sum_{j=1}^{n} A_{\varepsilon}(j,i)}{\sum_{j=1}^{n} A_{\varepsilon}(i,j)}, \quad i = 1, ..., n, (k > 1)$$

Assume that the dataset is located on a $d$ dimension manifold embedded in $D$ dimensional space. Consider the volume of sphere in the $d$ dimensional space with a radius of $\varepsilon$ (The metric $r$ used in this paper is the Euclidean distance).

$$\mu\{V_{\varepsilon}(d)\} = \left(\frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}\right)^{\frac{1}{2}} \varepsilon^d.$$  

In order to get an estimation function of $d$, we consider different neighborhood size with a ratio of $d$. For suitable scaling ratio $k$, the volume ratio of the two hyper spheres is $k^d$. Combined with the above conclusion, we have

$$q_{\varepsilon} = \mathbb{E}q_{\varepsilon,i} \approx k^d, \quad k > 1.$$
Assume the probability density function \( f(x) \) is smooth, and then \( f(x) \) is locally uniform where the probability is not zero [19]. This means that there is always a neighborhood for any point in the dataset, where the sample points can be regarded as uniformly distributed. Suppose \( \hat{q}_{n,k} \) is an estimator of ratio \( q_e \). Then we can get the estimation of intrinsic dimension \( d \).

\[
\hat{d}_{n,k} = \frac{\log \hat{q}_{n,k}}{\log k}, \quad k > 1.
\]  

(7)

For simplicity, we choose the sample mean as an unbiased estimator of \( q_e \).

\[
\hat{q}_{n,k} = \frac{1}{n} \sum_{i=1}^{n} B_{i,j}, \quad k > 1.
\]  

(8)

Although the above estimator is unbiased, because the sample size is limited and the spatial structure of the data is unknown, the scaling ratio \( k \) has a significant impact on the estimation of \( d \). In the next section, we will discuss how to select the appropriate \( k \).

3. **Supervised Parameter Selection Method**

In this section, we discuss the framework of parameter selection in datasets with class labels. Suppose we have a projection operator \( T \) that can project high dimensional data into a low dimensional space.

\[
\forall X \in \mathbb{R}^d, \exists Y \in \mathbb{R}^d, \text{ such that } TX = Y.
\]  

(9)

Then \( Y_1, \ldots, Y_e \in \mathbb{R}^d \) are the projections of \( X_1, \ldots, X_e \in \mathbb{R}^d \) in \( d \) dimensional space.

Suppose \( H \) is any classifier, we consider the classification accuracy of the projection \( Y_1, \ldots, Y_e \in \mathbb{R}^d \) under this classifier, denoted as \( \text{Acc}(d) \). As we assumed in the first section, the intrinsic dimension of a dataset is the dimension of the manifold on which the data located. If the projection operator is well defined, the classification accuracy of \( Y_1, \ldots, Y_e \in \mathbb{R}^d \) is no difference with that of \( X_1, \ldots, X_e \in \mathbb{R}^d \). However, due to the existence of noise, the actual classification accuracy of \( Y_1, \ldots, Y_e \in \mathbb{R}^d \) may be slightly lower than that of the original data (Benchmark classification accuracy, denoted as \( \text{Acc}(D) \)).

We consider the optimal scaling ratio \( k \) from two aspects. First, the change of \( d \) with respect to \( k \) should be small. Second, the classification accuracy after projection to \( d \)-dimensional is close to the benchmark classification accuracy. We use the first-order difference to describe the rate of change of \( d \) with \( k \), and the absolute deviation to describe the degree of deviation of classification accuracy.

Then the scaling ratio selection model can be described as follows.

\[
k = \arg \min_k \alpha \left( \frac{\Delta d_k}{\Delta k} \right)^2 + \beta(\text{Acc}(d_k) - \text{Acc}(D))^2.
\]  

(10)

Here \( \alpha \) and \( \beta \) are two positive numbers used to normalized the weights of two parts, do not require \( \alpha + \beta = 1 \). The estimated dimension \( \hat{d}_{n,k} \) obtained by (7) is generally not an integer, so a rounding operation is required in dimension reduction.

4. **Experiments**

We verified the proposed intrinsic dimension estimation method and parameter selection framework by two datasets, the Epilepsy data from University of Bonn (http://epileptologieonn.de/cms/front_content.php?idcat=193) and the MNIST dataset (http://yann.lecun.com/exdb/mnist/).

First, we determine the range of \( k \) by the relationship between the minimum distance and the maximum distance between sample points, and then take the values of \( k \) by a fixed length. Second, the estimated dimension corresponding to different \( k \) values is calculated according to the new method.
The projection operator used in this paper is classical Multidimensional Scaling method (MDS). As mentioned in the first section, MDS tends to preserve the distances between data points. We choose different classifiers for different datasets.

Finally, according to the parameter selection model (10), the best parameter \( k \) is determined, and the corresponding dimension estimation value is the intrinsic dimension of the data set.

4.1. Epilepsy Data

The dataset consists of brain signals from five healthy people and five epilepsy patients. It contains five datasets, which are represented by capital letters Z, O, N, F, and S, respectively. The first two sets of data were collected from healthy people and labelled as negative, and the last three data were collected from epilepsy patients and labelled as positive. There are 100 classification segments in each dataset, that is, 5 datasets have a total of 500 data. Each segment is 23.6s, and each segment has 4097 samples, that is, each data is 4097 dimensions. We select the first 70 data of each dataset as the training data, and the last 30 data as the test data.

The estimated dimension corresponding to different \( k \) values is shown in Figure 1. The estimated dimension decreases monotonically as the scaling ratio increases. It is difficult to use this result to determine the intrinsic dimension of data sets.

In this example, we use Bagged Trees as the classifier. The benchmark classification accuracy of the Epilepsy data is 83.7%.

![Figure 1. Estimated intrinsic dimension under different scaling ratios of Epilepsy data.](image-url)
According to parameter selection model (10), we take $\alpha=0.11624$, $\beta=21.836$, then the optimal scaling ratio is $k=1.78$, and the corresponding intrinsic dimension is 9.554 (Rounded to 10).

4.2. MNIST Data

MNIST is a well-known handwritten digits dataset. In this paper, we only consider digits 2 and 9, select 700 handwritten digits from each training set and 300 in each test set.

The estimated dimension corresponding to different $k$ values is shown in Figure 3. The estimated dimension increases first and then decreases with the change of scaling ratio $k$. Intuitively, the intrinsic dimension of the dataset is about 13.5.

In this example, we use Subspace KNN as the classifier. The benchmark classification accuracy of the MNIST data is 99.4% (Consider only the training set).
Figure 4. Classification accuracy of MNIST data under different projection dimensions.

According to parameter selection model (10), we take $\alpha=17.361$, $\beta=2.4901$, then the optimal scaling ratio is $k=1.21$, and the corresponding intrinsic dimension is $13.443$ (Rounded to 13).

Although the amount of data we select is smaller than that of the whole dataset, because we consider not only the data structure of the sample, but also the classification effect, the intrinsic dimension obtained is more convincing.

5. Conclusion
In this paper, the intrinsic dimension estimation problem is discussed. Based on correlation dimension, a novel estimating method is proposed. The new method contains only one parameter, the scaling ratio $k$. For datasets with class labels, we use the information provided by the estimator, and also consider the classification effect after projecting into low-dimensional space, constructing a supervised parametric selection model. Numerical experiments show that the method has satisfactory estimation ability.

Because the high-dimensional dataset usually have non-linear structure, and MDS is a linear dimension reduction method, in order to obtain higher classification accuracy, it is necessary to use non-linear dimension reduction methods. Future work will further study the influence of different projection operators and classifiers on dimension estimation, the convergence and stability of the new estimator also need further discussion.

6. Acknowledgments
This work was supported by National Natural Science Foundation of China (61573266).

7. References
[1] Trunk G V 1968 Statistical estimation of the intrinsic dimensionality of data collections Information and Control vol 12 pp 508-525
[2] Bennett R 1969 The intrinsic dimensionality of signal collections IEEE Trans. Inf. Theory vol 15 pp 517-525
[3] Zaki M J and Meira Jr W 2014 Data mining and analysis: fundamental concepts and algorithms Cambridge University Press p 184
[4] Kégle B 2003 Intrinsic dimension estimation using packing numbers Advances in neural information processing systems pp 697-704
[5] Grassberger P and Procaccia I 1983 Measuring the strangeness of strange attractors Phys D: Nonlinear Phenomena vol 9 pp 189-208
[6] Camastra F and Vinciarelli A 2002 Estimating the intrinsic dimension of data with a fractal-based method IEEE Trans. Pattern Anal. Mach. Intell. vol 24 pp 1404-07
[7] Serra P and Mandjes M 2017 Dimension Estimation Using Random Connection Models J. Mach. Learn. Res. vol 18 pp 5041-75
[8] Granata D and Carnevale V 2016 Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets Sci. Rep. vol 6 p 31377
[9] Ceruti C, Bassis S, Rozza A, Lombardi G, Casiraghi E and Campadelli P 2014 Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration Pattern Recogn. vol 47(8) pp 2569-81
[10] Sammon J W 1969 A nonlinear mapping for data structure analysis IEEE Trans. Comput. vol 100(5) pp 401-409
[11] Demartines P and Hérault J 1997 Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of datasets IEEE Trans. Neural Netw. vol 8(1) pp 148-154
[12] Camastra F and Staiano A 2016 Intrinsic dimension estimation: Advances and open problems Inf. Sci. vol 328 pp 26-41
[13] Tenenbaum J B, De Silva V and Langford J C 2000 A global geometric framework for nonlinear dimensionality reduction Science vol 290(5500) pp 2319-23
[14] Roweis S T and Saul L K 2000 Nonlinear dimensionality reduction by locally linear embedding Science vol 290(5500) pp 2323-26
[15] Belkin M and Niyogi P 2003 Laplacian eigenmaps for dimensionality reduction and data representation Neural Comput. vol 15(6) pp 1373-96
[16] Samko O, Marshall A D and Rosin P L 2006 Selection of the optimal parameter value for the Isomap algorithm Pattern Recognit. Lett. vol 27(9) pp 968-979
[17] Karbauskaië R, Kurasova O and Dzemyda G 2007 Selection of the number of neighbours of each data point for the locally linear embedding algorithm Inf Technol. Control. vol 36(4) pp 359-364
[18] Campadelli P, Casiraghi E, Ceruti C and Rozza A 2015 Intrinsic dimension estimation: Relevant techniques and a benchmark framework Math. Probl. Eng. vol 2015 id 759567
[19] Rozza A, Lombardi G, Ceruti C, Casiraghi E and Campadelli P 2012 Novel high intrinsic dimensionality estimators Mach. Learn. vol 89(1-2) pp 37-65