http://dx.doi.org/10.15388/NA.2014.4.4

Geodesic distances in the intrinsic dimensionality estimation using packing numbers

Rasa Karbauskaitė, Gintautas Dzemyda

Institute of Mathematics and Informatics, Vilnius University
Akademijos str. 4, LT-08663 Vilnius, Lithuania
rasa.karbauskaite@mii.vu.lt; gintautas.dzemyda@mii.vu.lt

Received: 18 December 2013 / **Revised:** 28 April 2014 / **Published online:** 25 August 2014

Abstract. Dimensionality reduction is a very important tool in data mining. An intrinsic dimensionality of a data set is a key parameter in many dimensionality reduction algorithms. When the intrinsic dimensionality of a data set is known, it is possible to reduce the dimensionality of the data without losing much information. To this end, it is reasonable to find out the intrinsic dimensionality of the data. In this paper, one of the global estimators of intrinsic dimensionality, the packing numbers estimator (PNE), is explored experimentally. We propose the modification of the PNE method that uses geodesic distances in order to improve the estimates of the intrinsic dimensionality by the PNE method.

Keywords: multidimensional data, intrinsic dimensionality, packing numbers estimator, manifold, degrees of freedom, image understanding, motion.

1 Introduction

In the real applications, we confront with data that are of a very high dimensionality. For example, in the image analysis each image is described by a large number of pixels of different colour. Another application, that produces data of a very high dimensionality, is analysis of DNA microarray data [1]. The analysis of high-dimensional data is usually challenging. The dimensionality reduction or visualization methods are recent techniques to discover knowledge hidden in multidimensional data sets [2]. Although data are considered in a high-dimensional space, in fact they are often either points of a nonlinear manifold of some lower dimensionality or the points close to that manifold. Thus, one of the major problems is to find the exact dimensionality of the manifold. Afterwards, it is reasonable to transfer the data points that lie on or near to this manifold into the space, the dimensionality of which is coincident with the manifold dimensionality. As a result, the dimensionality of the data set will be reduced to that of a manifold. Therefore, the problem is to disclose the manifold dimensionality, i.e. the intrinsic dimensionality (ID) of the analysed data.

The intrinsic dimensionality of a data set is usually defined as the minimal number of parameters or latent variables necessary to describe the data [3]. Latent variables are

still often called as degrees of freedom of a data set [3, 4]. Let the dimensionality of the analysed data be n. High-dimensional data sets can have meaningful low-dimensional structures hidden in the observation space, i.e. the data are of a low intrinsic dimensionality d ($d \ll n$).

Recently, a lot of manifold learning methods have been proposed to solve the problem of nonlinear dimensionality reduction. Important manifold learning algorithms include isometric feature mapping (ISOMAP) [4], locally linear embedding (LLE) [5, 6], Laplacian eigenmaps (LE) [7], Hessian LLE [8] etc. They all assume data to distribute on an intrinsically low-dimensional manifold and reduce the dimensionality of data by investigating the intrinsic structure of data. However, all manifold learning algorithms require the intrinsic dimensionality of the data as a key parameter for implementation. In recent years, ISOMAP and LLE have drawn great interests. They avoid nonlinear optimization and are simple to implement. However, both ISOMAP and LLE methods need the precise information of both the input parameters k for the neighbourhood identification and the intrinsic dimensionality d of the data set. The ways to select the value of the parameter k are proposed and investigated in [9–12]. If the intrinsic dimensionality d is set larger than what it really is, much redundant information will also be preserved; if it is set smaller, useful information of the data could be lost during the dimensionality reduction [13].

Due to the increased interest in dimensionality reduction and manifold learning, a lot of techniques have been proposed in order to estimate the intrinsic dimensionality of a data set [14–19] etc. Techniques for intrinsic dimensionality estimation can be divided into two main groups [20]: (1) estimators based on the analysis of local properties of the data (the correlation dimension estimator [21], the nearest neighbour dimension estimator [22–24], and the maximum likelihood estimator (MLE) [19]), and (2) estimators based on the analysis of global properties of the data (the eigenvalue-based estimator [23, 25], the packing numbers estimator (PNE) [18], and the geodesic minimum spanning tree estimator [17]). Local intrinsic dimensionality estimators are based on the idea that the number of data points, that are covered by a hypersphere of some radius $\tilde{r}$ around a given data point, grows in proportion to $\tilde{r}^d$, where d is the intrinsic dimensionality of the data manifold around that point. As a result, the intrinsic dimensionality d can be estimated by measuring the number of data points, covered by a hypersphere with a growing radius $\tilde{r}$. While the local estimators of the intrinsic dimensionality compute the average over the local estimates of intrinsic dimensionality, the global estimators consider the data as a whole when estimating the intrinsic dimensionality.

The packing number estimator (PNE) [18] has drawn great interests in the literature. Huge number of references may be found. However, to our knowledge, no further development of this estimator is done. Our research improves the quality of the estimates using packing numbers.

In this paper, we propose the modification of the PNE method that uses geodesic distances. We compare the application of both Euclidean and geodesic distances between data points in the PNE algorithm. The application of the geodesic distances in the intrinsic dimensionality estimation using packing numbers is grounded via the large experimental investigation.

2 Packing numbers estimator of the intrinsic dimensionality

Packing numbers estimator [18] is one of the global estimators of the intrinsic dimensionality.

At first, several definitions are given for clarity.

1. Given a metric space $\mathbb{R}^n$ with the distance metrics $d(\cdot,\cdot)$, a set $X \subset \mathbb{R}^n$ is said to be *r-separated*, if $d(X_i, X_j) \geqslant r$ for all distinct $X_i, X_j \in X$. In [18], r is called as a resolution.
2. The *r-packing number* $M(r)$ of a set $X \subset \mathbb{R}^n$ is defined as the maximum cardinality of an r-separated subset of X [18].

In [18], the intrinsic dimensionality of a set X has been suggested to be found by evaluating the limit

$$d = -\lim_{r\to 0} \frac{\log M(r)}{\log r}. \tag{1}$$

For a finite set, the zero limit cannot be achieved. Therefore, Kégl [18] suggests to redefine the intrinsic dimensionality in a scale-dependent manner. Let the analysed data set X consists of m n-dimensional points $X_i = (x_{i1}, \ldots, x_{in})$, $i = \overline{1,m}$ ($X_i \in \mathbb{R}^n$). Then the intrinsic dimensionality of the finite data set X is estimated by the formula

$$\hat{d} = -\frac{\log M(r_2) - \log M(r_1)}{\log r_2 - \log r_1}. \tag{2}$$

In order to find the r-packing number $M(r)$ for a finite data set $X = \{X_1, \ldots, X_m\}$, Kégl [18] applies the following approximation algorithm. Given a data set X, the algorithm starts with an empty set of centers C, and in an iteration over X it adds to C the data points that are at a distance of at least r from all the centers in C (lines 4 to 8 in Algorithm 1).

Algorithm 1. The PNE algorithm for the estimate $\hat{d}$ of the intrinsic dimensionality of the data set X [18].

Input: $X, m, r_1, r_2, \varepsilon$
Output: $\hat{d}$

1: **for** $l \longleftarrow 1, \infty$ **do**
2: Permute X randomly
3: **for** $k \longleftarrow 1, 2$ **do**
4: $C \longleftarrow \varnothing$
5: **for** $i \longleftarrow 1, m$ **do**
6: **for** $j \longleftarrow 1, |C|$ **do**
7: **if** $d(X[i], C[j]) < r_k$ **then** $j \longleftarrow m+1$
8: **if** $j < m+1$ **then** $C \longleftarrow C \cup \{X[i]\}$
9: $\hat{M}_k[l] \longleftarrow \log|C|$
10: $\hat{d} \longleftarrow -(\mu(\hat{M}_2) - \mu(\hat{M}_1))/(\log r_2 - \log r_1)$
11: **if** $l > 10$ **and** $1.65\sqrt{\sigma^2(\hat{M}_1) + \sigma^2(\hat{M}_2)}/(\sqrt{l}\,(\log r_2 - \log r_1)) < \hat{d}(1-\varepsilon)/2$ **then return** $\hat{d}$

The estimate $\hat{M}(r)$ is the cardinality of C after each point in X has been visited (line 9). According to Kégl, a good estimate for $\hat{d}$ can be obtained by using $\hat{M}(r)$ instead of $M(r)$. Because of the dependence of $\hat{M}(r)$ on the order of the data points at which they are visited, the variance of $\hat{M}(r)$ can distort the dimension estimate $\hat{d}$. To eliminate this variance, the procedure is repeated several times with random permutations of the data (lines 1, 2), and the estimate $\hat{d}$ is computed by using the average $\mu(\cdot)$ of logarithms of the packing numbers (line 10). The number of repetitions depends on r_1, r_2, and a preset parameter ε that determines the accuracy of the final estimate (set to 99% in all experiments). The complete algorithm is given formally in Algorithm 1.

3 Data sets

The following data sets were used in the experiments:

- 1000 three-dimensional data points ($m = 1000,\ n = 3$) that lie on a nonlinear two-dimensional S-shaped manifold (Fig. 1a).
- 1000 three-dimensional data points ($m = 1000,\ n = 3$) that lie on a nonlinear two-dimensional 8-shaped manifold (Fig. 1b). The components (x_1, x_2, x_3) of these data are calculated by the parametric equations below:

$$\begin{aligned} x_1 &= \cos(v), \\ x_2 &= \sin(v)\cos(v), \\ x_3 &= u, \end{aligned}$$

  where $v \in [2\pi/m; 2/\pi]$, $u \in (0; 5)$, m is the number of data points.
- 1000 three-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear one-dimensional manifold, i.e. a helix (Fig. 1c). The components (x_1, x_2, x_3) of these data are calculated by the parametric equations below:

$$\begin{aligned} x_1 &= \big(2 + \cos(8t)\big)\cos(t), \\ x_2 &= \big(2 + \cos(8t)\big)\sin(t), \\ x_3 &= \sin(8t), \end{aligned}$$

  where $t \in [2\pi/m; 2\pi]$, m is the number of data points.
- 1801 three-dimensional data points ($m = 1801,\ n = 3$) that lie on a nonlinear one-dimensional manifold, i.e. a spiral (Fig. 1d). The components (x_1, x_2, x_3) of these data are calculated by the parametric equations below:

$$\begin{aligned} x_1 &= 100\cos(t), \\ x_2 &= 100\sin(t), \\ x_3 &= t, \end{aligned}$$

  where $t \in [0; 10\pi]$.

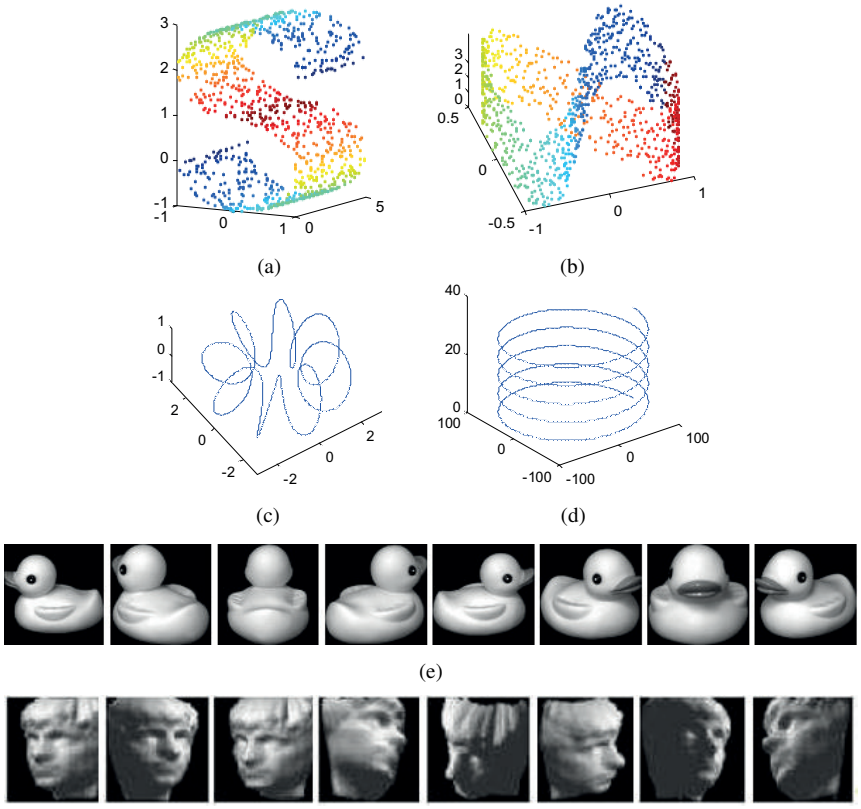


Fig. 1. Data sets of manifolds.

- *A data set of uncoloured (greyscale) pictures of a rotated duckling* [26] (samples of pictures are shown in Fig. 1e). The data are comprised of uncolored pictures of the same object (a duckling), obtained by a gradually rotated duckling at the 360° angle. The number of pictures (data points) is $m = 72$. The images have 128×128 greyscale pixels, therefore the dimensionality of points, characterizing each picture in a multidimensional space, is $n = 16384$, and the colour value is from the interval $[0, 255]$. Source database: `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`.
- *A data set of uncoloured (greyscale) images of a person's face* [4] (samples of images are shown in Fig. 1f). The data consist of many photos of the same person face observed in different poses (left-and-right pose, up-and-down pose) and lighting conditions, in no particular order. The number of photos (data points) is $m = 698$. The images have 64×64 greyscale pixels, therefore the dimensionality of points that characterize each photo in a multidimensional space is $n = 4096$. Source database: `http://isomap.stanford.edu/datasets.html`.

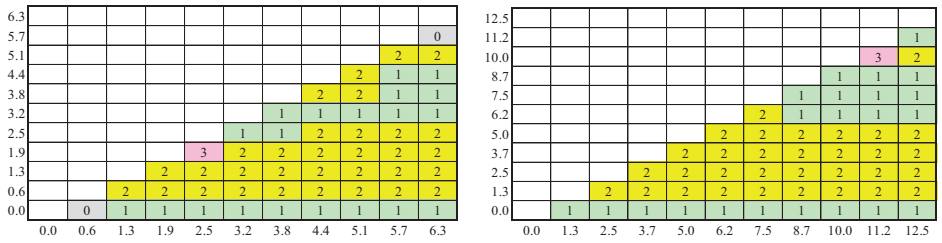
4 Experimental exploration of the packing numbers estimator (PNE)

In this section, the PNE method is investigated experimentally with various artificial and real data sets. The data points lie on manifolds, the dimensionality of which is known in advance. Therefore, we will be able to establish precisely, whether the estimate of the intrinsic data dimensionality, obtained by the PNE, is true. The aim of these investigations is to find out, which distances (Euclidean or geodesic) are better to be used in the PNE algorithm, while estimating the similarity between data points.

The geodesic distance is the length of the shortest path between two points along the surface of a manifold. Here the Euclidean distances are used when calculating the length of the shortest path. In order to compute the geodesic distances between the points $X_1, X_2, \ldots, X_m$, it is necessary to set some number of the nearest points (neighbours) of each point X_i. The search for the neighbours of each point X_i can be organized in two ways: (1) by the fixed number k_{geod} of the nearest points from X_i, (2) by all the points within some fixed radius of a hypershere, the center of which is the point X_i. When the neighbours are defined, a weighted graph over the points is constructed: each point X_i is connected with its neighbours; the weights of edges are Euclidean distances between the point X_i and its neighbours. Using one of the algorithms for the shortest path distance in the graph, the shortest path lengths between the pairs of all points are computed. These lengths are estimates of the geodesic distances between the points.

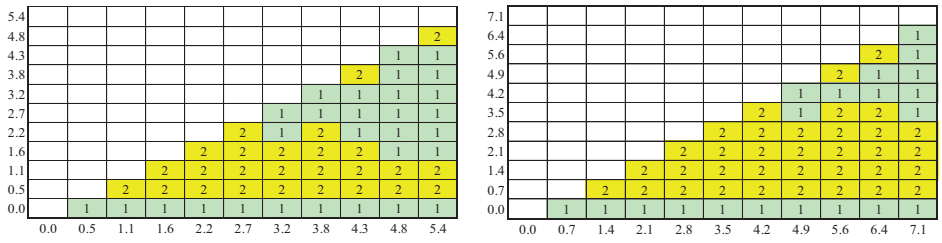
In [18], the dimensionality estimate $\hat{d}$ is measured on consecutive pairs of a sequence $r_1, \ldots, r_s$ of resolutions, and the estimate is plotted halfway between the two parameters, i.e. $\hat{d}(r_i, r_{i+1})$ is plotted at $(r_i + r_{i+1})/2$. In our paper, the values of r_1 and r_2 ($r_1 < r_2$) are chosen in another way: the minimum distance and the maximum distance between all the data points are found; then all the cases of r_1 and r_2 from the minimum distance to the maximum one with a step equal to (*maximum distance-minimum distance*)/10 are fixed. In this case, the total number of combinations of r_1 and r_2 ($r_1 < r_2$) is equal to 55 ($11 \times (11-1)/2 = 55$). The PNE algorithm is explored by evaluating two types of distances: Euclidean and geodesic. In both cases, the PNE values $\hat{d}$, defined by formula (2), are calculated with all the combinations of r_1 and r_2. In such a way, dependences of the estimate of intrinsic dimensionality of the data on the parameters r_1 and r_2 are obtained. As the estimate of the intrinsic dimensionality, we choose such a PNE value $\hat{d}$ that remains the same in the largest area of the values of parameters r_1 and r_2. Further this area will be called by a dominant area and the value of $\hat{d}$ in this area will be called by a dominant value. Let the numbers of combinations of r_1 and r_2, that fall in the dominant area in the case of Euclidean and geodesic distances, be denoted as v_{e} and v_{g}, respectively.

The first investigation is performed with the points of the two-dimensional S-shaped manifold ($m = 1000$, $n = 3$), see Fig. 1a. We can see from Fig. 2 that the dominant value is equal to the true intrinsic dimensionality, i.e. $\hat{d} = 2$, when both Euclidean and geodesic distances are used. However, the dominant area is a bit larger when using the geodesic distances in the PNE algorithm: $v_{\mathrm{e}} = 27$ and $v_{\mathrm{g}} = 31$. The next investigation is performed with the points of the two-dimensional 8-shaped manifold ($m = 1000$, $n = 3$), see Fig. 1b. At first sight, it is difficult to say which area is dominant when the Euclidean distances are used (Fig. 3a). However, $v_{\mathrm{e}} = 24$ as $\hat{d} = 2$ and $v_{\mathrm{e}} = 19$ as



(a) (b)

Fig. 2. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: the S-shaped manifold.



(a) (b)

Fig. 3. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: the 8-shaped manifold.

$\hat{d} = 1$. It leads to the conclusion that the dominant value is $\hat{d} = 2$. In the case of geodesic distances, it is obvious that the dominant value is $\hat{d} = 2$ ($v_{\mathrm{g}} = 33$). The investigations performed with the two-dimensional S-shaped and 8-shaped manifolds show that the dominant values of the estimate of the intrinsic dimensionality are coincident with the true intrinsic dimensionality of these data in the case of both distances, i.e. $\hat{d} = d = 2$.

After investigating a helix ($m = 1000,\ n = 3$, Fig. 1c) that is the one-dimensional manifold, it became clear that it is reasonable to evaluate geodesic distances instead of Euclidean ones in the PNE algorithm in order to get the true estimates of the intrinsic dimensionality d of the data. Figure 4 indicates that the dominant value is $\hat{d} = 2 \neq d$ in the case of Euclidean distances ($v_{\mathrm{e}} = 25$), and the dominant value is $\hat{d} = 1 = d$, in the case of geodesic distances ($v_{\mathrm{g}} = 49$).

Next, a spiral (Fig. 1d) which is the one-dimensional manifold ($m = 1801$, $n = 3$) was investigated. In this case, the difference between the use of the Euclidean and geodesic distances in the PNE algorithm is not considerable (Fig. 5). In both cases, the dominant value of the estimate of the intrinsic dimensionality is $\hat{d} = 1$ which is coincident with the true intrinsic dimensionality of these data, i.e. $\hat{d} = d = 1$. However, the dominant area is a bit larger when the geodesic distances are used, i.e. $v_{\mathrm{g}} = 54$ and $v_{\mathrm{e}} = 49$.

One of the practical fields, where high-dimensional data appear, is the analysis of images. Let us have a set of images of some moving object. There are many investigations of such sets [4, 27, 28] etc. Each image is described by the number of pixels of different colour. The dimensionality of such a data set is equal to the number of pixels in the

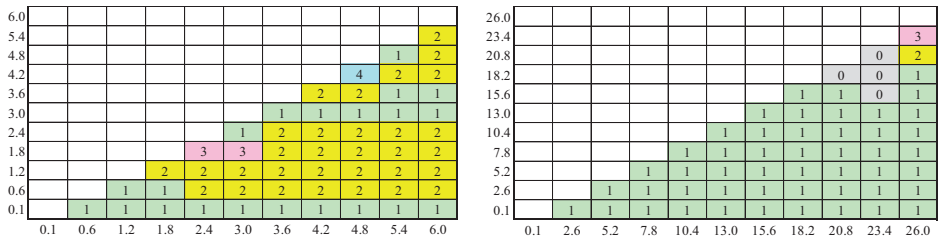


Fig. 4. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: the helix.

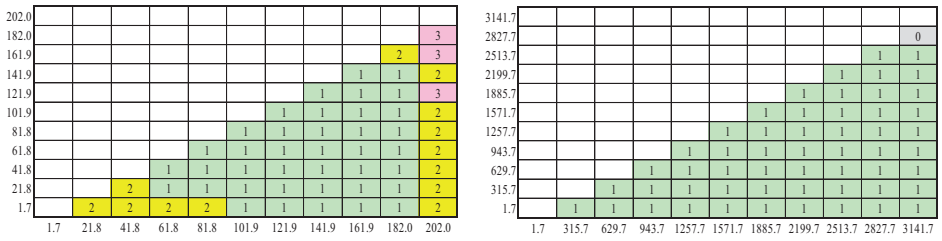
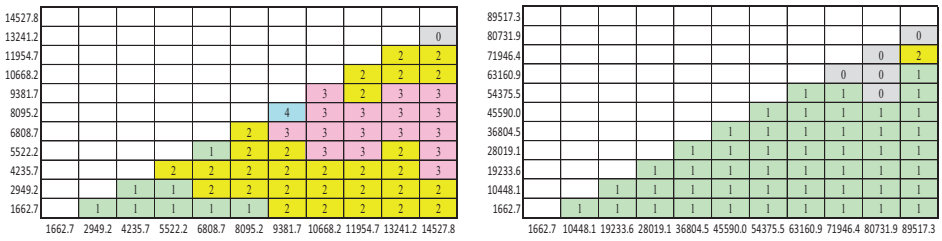


Fig. 5. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: the spiral.

greyscale case or even it is three times larger than the number of pixels in the coloured case. So, the dimensionality of these data is very large. Since the intrinsic dimensionality of a data set is defined as the minimal number of latent variables or features, necessary to describe the data [3], we hypothesize that there are latent variables or features that characterize the motion of the object in the images and their number is the same as the number of degrees of freedom of a possible motion of the object. Therefore, the minimal possible intrinsic dimensionality of a data set of images should be equal to the number of degrees of freedom of a possible motion of the object.

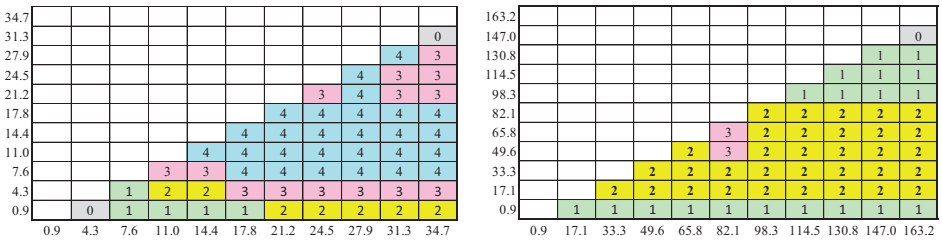
The high-dimensional data, obtained from the set of images (greyscale pictures of a rotated duckling and photos of the same person face observed in different poses), are investigated. Since a duckling was gradually rotated at a certain angle on the same plane, i.e. without turning the object itself, these data have only one degree of freedom, i.e. the minimal intrinsic dimensionality of these data may be equal even to 1. The person's face analysed in [4] has two directions of motion (two poses): left-and-right pose and up-and-down pose. Therefore, the high-dimensional data, corresponding to these pictures, have two degrees of freedom, i.e. the minimal possible intrinsic dimensionality of these data should be equal to 2.

The results of investigation with the high-dimensional data points, corresponding to real pictures of a rotated duckling ($m = 72$, $n = 16384$, Fig. 1e), are given in Fig. 6. Taking into consideration that the duckling in the pictures has only one degree of freedom, an assumption can be made that the intrinsic dimensionality of these data is 1. In this case,



(a) (b)

Fig. 6. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: pictures of the rotated duckling.



(a) (b)

Fig. 7. Estimate $\hat{d}$ of the intrinsic dimensionality depending on distances: (a) Euclidean, (b) geodesic. Data set: photos of a face.

we notice the advantage of the geodesic distances. If the Euclidean distances are used, the dominant value $\hat{d}$ is false, i.e. $\hat{d} = 2$, $v_{\mathrm{e}} = 23$. However, the intrinsic dimensionality of these data is evaluated truly by PNE, if the geodesic distances between the data points are used, i.e. $\hat{d} = 1$, $v_{\mathrm{g}} = 49$.

The last investigation is performed with the high-dimensional data points ($m = 698$, $n = 4096$), corresponding to the photos of the face of the same person observed in different poses from different lighting direction (Fig. 1f). We can see the results in Fig. 7a. If the Euclidean distances are used, the dominant value is $\hat{d} = 4$, $v_{\mathrm{e}} = 27$. However, $\hat{d} = 2$ dominates ($v_{\mathrm{g}} = 33$) if the geodesic distances are used in the PNE method. Referring to the hypothesis that the minimal intrinsic dimensionality of these data is 2, we see the advantage of the geodesic distance again.

The intrinsic dimensionality of the face data set [4] is analysed in several papers. It is shown in [4] that the intrinsic dimensionality of this data set is 3. Levina and Bickel [19] state that the estimated dimensionality of about 4 is very reasonable. In [27], the estimated dimensionality is equal to 2, when geodesic distances are used in the MLE algorithm, and it is equal to 4 or 5, when Euclidean distances are used in MLE. The question arises which estimated dimensionality can be taken as the true intrinsic dimensionality?

In order to answer this question and verify whether our hypothesis on the minimal intrinsic dimensionality is true in this case, let us analyse the face database [4] in detail. At first, the 4096-dimensional data points are projected on the 5-dimensional space by the ISOMAP method [4]. ISOMAP is used in the investigation, because recently it is one of

the most popular manifold learning methods. So we get a matrix of size [698 x 5]. The rows of this matrix correspond to the objects $Y_1, Y_2, \ldots, Y_m$, $m = 698$, and the columns correspond to the features $y_1, y_2, \ldots, y_{n^*}$, $n^* = 5$, that characterize the objects. Then the covariance matrix C of the features is obtained:

$$C = \begin{pmatrix} 1538.8 & 0 & 0 & 0 & 0 \\ 0 & 419.3 & 0 & 0 & 0 \\ 0 & 0 & 276.3 & 0 & 0 \\ 0 & 0 & 0 & 86.8 & 0 \\ 0 & 0 & 0 & 0 & 79.1 \end{pmatrix}. \tag{3}$$

It is obvious from this covariance matrix that all the 5 features y_k and y_l are not correlated, because their covariance coefficient is equal to zero: $c_{kl} = c_{lk} = 0, k \neq l$. The covariance coefficient c_{kk}, $k = \overline{1, n^*}$, is the variance of features y_k. So, we see from (3) that the variances of the first three features are much larger than others. The variances of the fourth and fifth features are the least ones and differ very little from each other. It means that three features are the main ones. A question arises which features from y_1, y_2, y_3 correspond to the left-and-right pose, the up-and-down pose, and to the lighting direction. In order to answer this question, we visualized these features pairwise on the plane (see Figs. 8–10). Figures 8–10 show that the feature y_1 corresponds to the left-and-right pose, the feature y_2 corresponds to the up-and-down pose, and the feature y_3 corresponds to the lighting direction. Summarizing everything, it is obvious that the

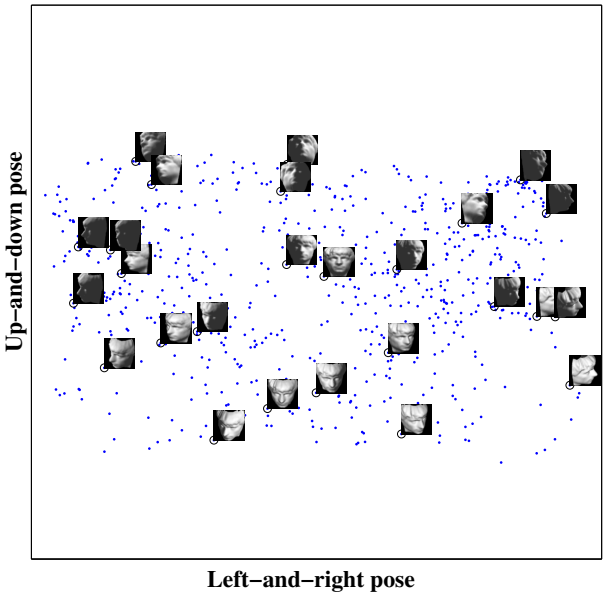


Fig. 8. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: left-and-right pose, up-and-down pose.

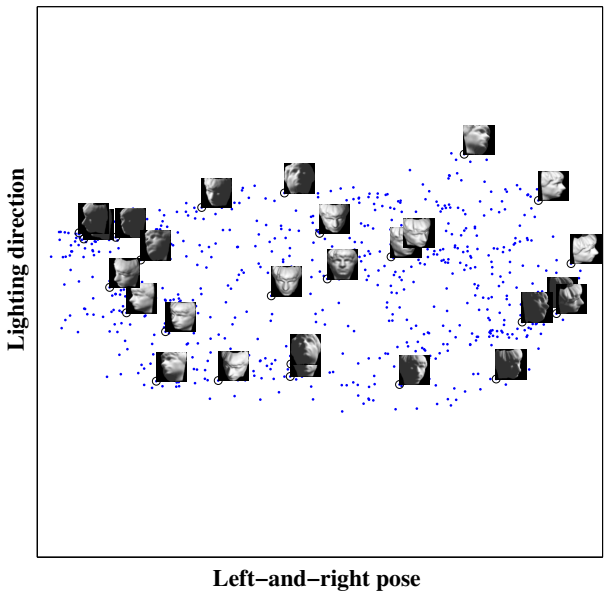


Fig. 9. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: left-and-right pose, lighting direction.

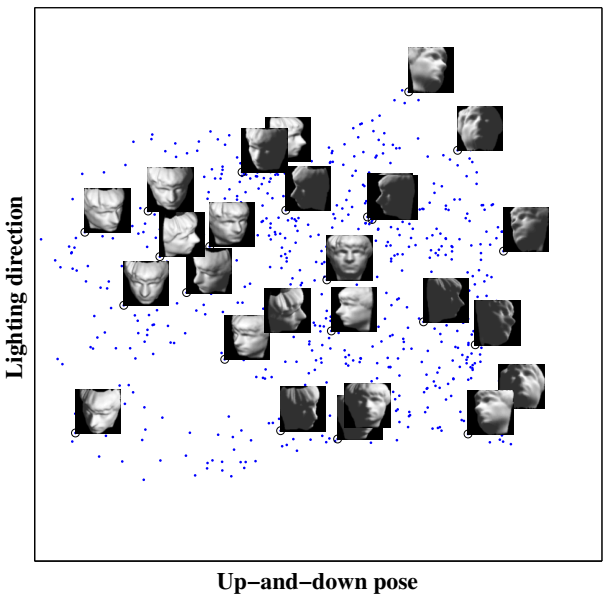


Fig. 10. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: up-and-down pose, lighting direction.

first two features, i.e. both poses – left-and-right and up-and-down – are more essential than the third feature regarding the lighting direction.

Since the face database consists of images of an artificial face under three changing conditions: vertical and horizontal orientation and illumination (lighting direction), it is possible to assume that the intrinsic dimensionality of this data set should be 3. However, the person's face has two directions of motion (two poses): left-and-right pose and up-and-down pose. So the minimal intrinsic dimensionality of these data can be assigned to 2, if we assume that the intrinsic dimensionality of the data set of images were equal to the number of degrees of freedom of a possible motion of the object in the image.

So, after such a discussion, we dare say, that in the last investigation, a false result is obtained by PNE, if the Euclidean distances are used. However, the minimal intrinsic dimensionality of these data is evaluated well, if the geodesic distances between the data points are calculated.

5 Conclusions

Real-life data are often hardly understandable because of their high dimensionality. While analysing these data, we often have to reduce their dimensionality so that to preserve as much information on the analysed data set as possible. To this end, it is reasonable to find out the intrinsic dimensionality of the data. Several methods for estimating the intrinsic dimensionality are proposed in the literature.

In this paper, we have analysed one of the global estimators for intrinsic dimensionality – the packing numbers estimator (PNE). We have shown that, in order to get true estimates by PNE, it is necessary to evaluate the geodesic distances between data points in this algorithm. If the Euclidean distances are used in PNE, we can get false estimates of the intrinsic dimensionality. In this paper, in the experiments, no false results are obtained in the PNE modification that uses the geodesic distances.

The efficiency of PNE is disclosed in the images analysis. The experiments with the sets of images of the moving object showed that there are latent variables or features that characterize the motion of the object in the images and their number is the same as the number of degrees of freedom of a possible motion of the object. It is shown in this paper that the minimal possible intrinsic dimensionality of a data set of images is equal to the number of degrees of freedom of a possible motion of the object.

References

1. E. Kriukienė, V. Labrie, T. Khare, G. Urbanavičiūtė, A. Lapinaitė, K. Koncevičius, D. Li, T. Wang, S. Pai, C. Ptak, J. Gordevičius, S. Wang, A. Petronis, S. Klimašauskas, DNA unmethylome profiling by covalent capture of CpG sites, *Nat. Commun.*, **4**, Article No. 2190, 2013.
2. G. Dzemyda, O. Kurasova, J. Žilinskas, *Multidimensional Data Visualization: Methods and Applications*, Springer Optimization and Its Applications, Vol 75, Springer, 2013.
3. J.A. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction*, Springer, New York, 2007.

4. J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**(5500):2319–2323, 2000.

5. S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, **290**(5500):2323–2326, 2000.

6. L.K. Saul, S.T. Roweis, Think globally, fit locally: Unsupervised learning of low dimensional manifolds, *J. Mach. Learn. Res.*, **4**:119–155, 2003.

7. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.), Neural Computation*, **15**(6):1373–1396, 2003.

8. D.L. Donoho, C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, *Proc. Natl. Acad. Sci. USA*, **102**(21):7426–7431, 2005.

9. R. Karbauskaitė, O. Kurasova, G. Dzemyda, Selection of the number of neighbours of each data point for the locally linear embedding algorithm, *Information Technology and Control*, **36**(4):359–364, 2007.

10. R. Karbauskaitė, G. Dzemyda, V. Marcinkevičius, Selecting a regularization parameter in the locally linear embedding algorithm, in: *Proceedings of the 20th International EURO Mini Conference "Continuous optimization and knowledge-based technologies" (Neringa, Lithuania, May 20–23, 2008)*, Technika, Vilnius, 2008, pp. 59–64.

11. R. Karbauskaitė, G. Dzemyda, Topology preservation measures in the visualization of manifold-type multidimensional data, *Informatica*, **20**(2):235–254, 2009.

12. R. Karbauskaitė, G. Dzemyda, V. Marcinkevičius, Dependence of locally linear embedding on the regularization parameter, *Top*, **18**(2):354–376, 2010.

13. M. Fan, H. Qiao, B. Zhang, Intrinsic dimension estimation of manifolds by incising balls, *Pattern Recognition*, **42**:780–787, 2009.

14. K.Q. Weinberger, L.K. Saul, Unsupervised learning of image manifolds by semidefinite programming, *Int. J. Comput. Vis.*, **70**(1):77–90, 2006.

15. M. Brand, Charting a manifold, in: S. Becker, S. Thrun, K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, 2003, pp. 985–992.

16. F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**(10):1404–1407, 2002.

17. J.A. Costa, A.O. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, *IEEE Trans. Signal Process.*, **52**(8):2210–2221, 2004.

18. B. Kégl, Intrinsic dimension estimation using packing numbers, in: S. Becker, S. Thrun, K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, 2003, pp. 697–704.

19. E. Levina, P.J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, MIT Press, Cambridge, 2005, pp. 777–784.

20. L.J.P. van der Maaten, An introduction to dimensionality reduction using MATLAB, Tech. Report MICC 07-07, Maastricht University, 2007.

21. P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, *Physica D*, **9**(1–2):189–208, 1983.

22. P. Verveer, R. Duin, An evaluation of intrinsic dimensionality estimators, *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**(1):81–86, 1995.

23. F. Camastra, Data dimensionality estimation methods: A survey, *Pattern Recognition*, **36**:2945–2954, 2003.

24. K.M. Carter, R. Raich, A.O. Hero, On local intrinsic dimension estimation and its applications, *IEEE Trans. Signal Process.*, **58**(2):650–663, 2010.

25. K. Fukunaga, D. Olsen, An algorithm for finding intrinsic dimensionality of data, *IEEE Trans. Comput.*, **20**:176–183, 1971.

26. S.A. Nene, S.K. Nayar, H. Murase, Columbia object image library (COIL-20), Tech. Report CUCS-005-96, Columbia University, 1996.

27. R. Karbauskaitė, G. Dzemyda, E. Mazėtis, Geodesic distances in the maximum likelihood estimator of intrinsic dimensionality, *Nonlinear Anal. Model. Control*, **16**(4):387–402, 2011.

28. V. Raudonis, A. Paulauskaitė-Tarasevičienė, L. Kižauskienė, The gaze tracking system with natural head motion compensation, *Informatica*, **23**(1):105–124, 2012.