Int. J. Appl. Math. Comput. Sci., 2015, Vol. 25, No. 4, 895–913
DOI: 10.1515/amcs-2015-0064

OPTIMIZATION OF THE MAXIMUM LIKELIHOOD ESTIMATOR FOR DETERMINING THE INTRINSIC DIMENSIONALITY OF HIGH–DIMENSIONAL DATA

RASA KARBAUSKAITĖ [a,*], GINTAUTAS DZEMYDA [a]

[a] Institute of Mathematics and Informatics
Vilnius University, Akademijos st. 4, 08663 Vilnius, Lithuania
e-mail: {rasa.karbauskaite,gintautas.dzemyda}@mii.vu.lt

One of the problems in the analysis of the set of images of a moving object is to evaluate the degree of freedom of motion and the angle of rotation. Here the intrinsic dimensionality of multidimensional data, characterizing the set of images, can be used. Usually, the image may be represented by a high-dimensional point whose dimensionality depends on the number of pixels in the image. The knowledge of the intrinsic dimensionality of a data set is very useful information in exploratory data analysis, because it is possible to reduce the dimensionality of the data without losing much information. In this paper, the maximum likelihood estimator (MLE) of the intrinsic dimensionality is explored experimentally. In contrast to the previous works, the radius of a hypersphere, which covers neighbours of the analysed points, is fixed instead of the number of the nearest neighbours in the MLE. A way of choosing the radius in this method is proposed. We explore which metric—Euclidean or geodesic—must be evaluated in the MLE algorithm in order to get the true estimate of the intrinsic dimensionality. The MLE method is examined using a number of artificial and real (images) data sets.

Keywords: multidimensional data, intrinsic dimensionality, maximum likelihood estimator, manifold learning methods, image understanding.

1. Introduction

Image analysis and understanding is a very challenging topic in exploratory data analysis. Recently, manifold learning methods (locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2003), isometric feature mapping (ISOMAP) (Tenenbaum *et al*., 2000), Laplacian eigenmaps (LEs) (Belkin and Niyogi, 2003), Hessian LLE (HLLE) (Donoho and Grimes, 2005), local tangent space analysis (LTSA) (Zhang and Zha, 2004), etc., see also the work of Lee and Verleysen (2007)) have been often applied in image processing. The practical value of these methods is shown in different applications such as face pose detection (Li *et al*., 2001; Hadid *et al*., 2002), face recognition (Yang, 2002; Zhang *et al*., 2004), the analysis of facial expressions (Chang *et al*., 2004; Elgammal and su Lee, 2004b), human motion data interpretation (Jenkins and Mataric, 2004), gait analysis (Elgammal and su Lee, 2004a; 2004b), wood texture analysis (Niskanen and Silven, 2003), and medical data analysis (Varini *et al*., 2004). The dimensionality of a manifold is very important in manifold learning. In this paper, the way how to determine the true value of the dimensionality is proposed.

In image analysis, we are confronted with data that are of a very high dimensionality, because each image is described by a large number of pixels of different colour. So, it is very difficult to understand these data. Although data are considered in a high-dimensional space, they are in fact either points of a nonlinear manifold of some lower dimensionality or 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 whose dimensionality 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 of the analysed data.

The intrinsic dimensionality of a data set is usually

*Corresponding author

defined as the minimal number of parameters or latent variables necessary to describe the data (Lee and Verleysen, 2007). Latent variables are still often called degrees of freedom of a data set (Tenenbaum *et al*., 2000; Lee and Verleysen, 2007). 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 low intrinsic dimensionality d ($d \ll n$). In more general terms, following Fukunaga (1982), a data set $X \subset \mathbb{R}^n$ is said to have the intrinsic dimensionality equal to d if its elements lie entirely within the d-dimensional subspace of $\mathbb{R}^n$ (where $d < n$) (Camastra, 2003).

Dimensionality reduction or visualization methods are recent techniques to discover knowledge hidden in multidimensional data sets (Shin and Park, 2011; Dzemyda *et al*., 2013; Kulczycki and Łukasik, 2014). Recently, a lot of manifold learning methods have been proposed to solve the problem of nonlinear dimensionality reduction. They all assume that data distribute on an intrinsically low-dimensional manifold and reduce the dimensionality of data by investigating their intrinsic structure. However, all manifold learning algorithms require the intrinsic dimensionality of data as a key parameter for implementation. In recent years, the ISOMAP and LLE have become of great interest. They avoid nonlinear optimization and are simple to implement. However, both ISOMAP and LLE methods need the precise information on both the input parameters k for the neighbourhood identification and the intrinsic dimensionality d of the data set. The ways of selecting the value of the parameter k are proposed and investigated by Kouropteva *et al.* (2002), Karbauskaite *et al.* (2007; 2008; 2010), Karbauskaitė and Dzemyda (2009) or Álvarez-Meza *et al*. (2011). If the intrinsic dimensionality d is set larger than 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 (Qiao and Zhang, 2009).

The term of a manifold is defined by Dzemyda *et al*. (2013) and Gong *et al*. (2014). A manifold is an abstract topological mathematical space in which the area of each point is similar to the Euclidean space; however, the global structure of a manifold is more complex. Therefore, operations performed on the manifold require choosing a metric. The minimum length curve over all possible smooth curves on the manifold between two points is called a geodesic, and the length of this curve stands for a geodesic distance; i.e., the geodesic metric measures lengths along the manifold, contrary to the Euclidean one, which measures lengths along the straight lines (Lee and Verleysen, 2007; Gong *et al*., 2014).

The simplest manifolds are a line and a circle that are one-dimensional. A plane and the surface of a ball, a torus are two-dimensional manifolds, etc. The area of each point on the one-dimensional manifold is similar to a line segment. The area of each point on the two-dimensional manifold is similar to a flat region. A simple example is given in Fig. 1. Data points of a two-dimensional manifold (Fig. 1(a)) are embedded in three dimensions in three different ways: a linear embedding (plane), Fig. 1(b), an S-shape, Fig. 1(c), and a "Swiss roll", Fig. 1(d).

In practice, more complicated examples of data manifolds are met in image processing. Each picture is digitized; i.e., a data point consists of colour parameters of pixels, and, therefore, it is of very large dimension. A question arises: Is the dimensionality of these data really so large or maybe data points lie on a manifold of much lower dimensionality? Data are often comprised of pictures of the same object, by turning the object gradually at a certain angle, or taking a picture of the object at different moments, etc. In this way, the points slightly differ from one another, making up a certain manifold. Detailed examples are given by Tenenbaum *et al*. (2000), Kouropteva *et al*. (2002), Saul and Roweis (2003) (face analysis) and Karbauskaitė *et al*. (2007) (comparison of pictures of an object rotated at different angles). It is often very important to understand and analyse these pictures in terms of their variability, for example, to view how a position of a human being, facial expression or a turn of the same object are changing (Weinberger and Saul, 2006). It is useful when identifying an unknown position of an object if we have a set of pictures of the object in different positions.

In the work of Levina *et al*. (2007), the estimated intrinsic dimensionality is applied to a real problem, i.e., to the issue of determining the number of pure components in a mixture from Raman spectroscopy data. The authors show how the estimate of the intrinsic dimensionality corresponds to the number of pure components. Having an accurate estimate of the number of pure components, it saves time in component extraction, etc. Another possible application is given by Karbauskaitė *et al*. (2011) as well as Karbauskaitė and Dzemyda (2014) to find the number of degrees of freedom of motion of the object in a set of pictures.

Due to an 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 (Camastra, 2003; Brand, 2003; Costa and Hero, 2004; Kégl, 2003; Hein and Audibert, 2005; Levina and Bickel, 2005; Weinberger and Saul, 2006; Qiao and Zhang, 2009; Yata and Aoshima, 2010; Mo and Huang, 2012; Fan *et al*., 2013; Einbeck and Kalantan, 2013; He *et al*., 2014).

Techniques for intrinsic dimensionality estimation can be divided into two main groups (van der Maaten, 2007; Einbeck and Kalantan, 2013): (1) estimators based on the analysis of local properties of the data (the correlation dimension estimator (Grassberger and

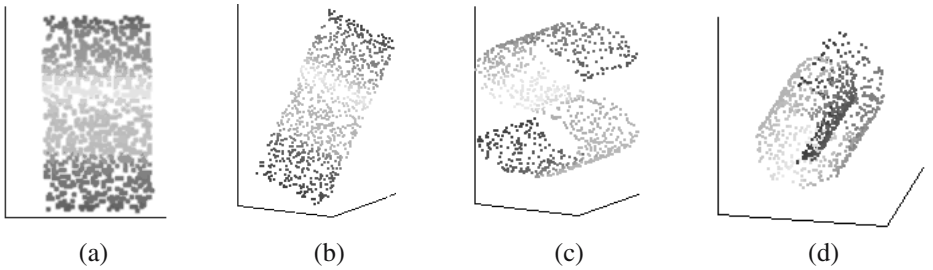

Fig. 1. Data points of a two-dimensional manifold (a) embedded in three dimensions: linear embedding (plane) (b), S-shape (c), "Swiss roll" (d).

Procaccia, 1983), the nearest neighbour dimension estimator (Verveer and Duin, 1995; Camastra, 2003; Carter *et al*., 2010), the maximum likelihood estimator (MLE) (Levina and Bickel, 2005), etc.), and (2) estimators based on the analysis of global properties of the data (the eigenvalue-based estimator (Fukunaga and Olsen, 1971; Camastra, 2003), the packing numbers estimator (PNE) (Kégl, 2003), and the geodesic minimal spanning tree (GMST) estimator (Costa and Hero, 2004), etc.). Local intrinsic dimensionality estimators are based on the idea that the number of data points that are covered by a hypersphere of some radius r around a given data point grows in proportion to 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 r. While the local estimators of the intrinsic dimensionality compute the average over the local estimates of the intrinsic dimensionality, the global estimators consider the data as a whole when estimating the intrinsic dimensionality.

Multiple novel applications in local intrinsic dimensionality estimation (anomalous activity in router networks, data clustering, image segmentation, etc.) are presented by Carter *et al*. (2010). The authors show the advantage of local intrinsic dimensionality estimation compared with the global one. They show that, by estimating the dimensionality locally, they are able to extend the use of dimensionality estimation in many applications that are not possible in global estimation.

Intrinsic dimension estimation methods can be categorized in another way, too, for example, into projection techniques and geometric approaches (Qiao and Zhang, 2009; Yata and Aoshima, 2010; Fan *et al*., 2013). Projection techniques project the data onto a low-dimensional space. The intrinsic dimensionality may be estimated by comparing the projections to the space of varying dimensionality with the initial data set. Such methods include: principal component analysis (PCA) and various PCA modifications, multidimensional scaling methods (MDS), nonlinear manifold learning methods (LLE, ISOMAP, LE, HLLE, etc.) (Camastra, 2003; Lee and Verleysen, 2007). Geometric techniques find the intrinsic dimensionality by investigating the geometric structure of data. Geometric methods are mostly based on fractal dimensions (box-counting dimension or capacity dimension (Camastra, 2003), correlation dimension (Grassberger and Procaccia, 1983), packing dimension (PNE) (Kégl, 2003), etc.) or nearest neighbour distances (Fukunaga–Olsen's algorithm, the near neighbour algorithm, topology representing network based methods (TRN) (Camastra, 2003), the incising balls method (Qiao and Zhang, 2009), the k-NNG method (Costa and Hero, 2005), the geodesic minimal spanning tree (GMST) (Costa and Hero, 2004), and the maximum likelihood estimator (Levina and Bickel, 2005), etc.). A good survey of intrinsic dimension estimation methods is given by Camastra (2003).

In this paper, the maximum likelihood estimator of the intrinsic dimensionality d is analysed. The way of choosing the parameter r is proposed in this method.

2. Maximum likelihood estimator of intrinsic dimensionality

The maximum likelihood estimator (Levina and Bickel, 2005) is one of the local estimators of the intrinsic dimensionality. Similarly to the correlation dimension and the nearest neighbour dimension estimator, the maximum likelihood estimator of the intrinsic dimensionality estimates the number of data points covered by a hypersphere with a growing radius r. In contrast to the former two techniques, the maximum likelihood estimator does so by modelling the number of data points inside the hypersphere as a Poisson process.

A detailed algorithm of the MLE is provided by Levina and Bickel (2005). The idea is given in the sequel.

The analysed data set X consists of m n-dimensional points $X_i = (x_{i1}, \ldots, x_{in})$, $i = \overline{1,m}$ ($X_i \in \mathbb{R}^n$). The MLE finds the intrinsic dimensionality d_{MLE} of the data set X.

In the MLE algorithm, it is necessary to look for the neighbouring data points. The search for the neighbours

of each point X_i can be organized in two ways: (i) by the fixed number k of the nearest points from X_i, starting from the closest point to the k-th point according to the distance, (ii) by all the points within some fixed radius r of a hypersphere, whose center is the point X_i. In the works of Tenenbaum *et al*. (2000), Levina and Bickel (2005), as well as Karbauskaitė *et al*. (2011), the k points, obtained in the first case, are called the k nearest neighbours of X_i.

Levina and Bickel (2005) provide a formula to estimate the intrinsic dimensionality of the data point X_i:

$$d_r\left(X_i\right)=\left[\frac{1}{N(r,X_i)}\sum_{j=1}^{N(r,\,X_i)}\log\frac{r}{T_j(X_i)}\right]^{-1}, \quad (1)$$

where $T_j(X_i)$ is the radius of the smallest hypersphere that is centred at X_i and contains j nearest neighbouring data points, i.e., $T_j(X_i)$ is the Euclidean distance $d(X_i,X_{ij})$ from the point X_i to the j-th nearest neighbour X_{ij} within the hypersphere centred at X_i; $N(r,X_i)$ counts the data points that are within the distance r from X_i, i.e., it is the number of data points among X_s, $s=\overline{1,m}$, $i\neq s$, that are covered by a hypersphere with the centre X_i and the radius r.

However, according to the authors, in practice, it is more convenient to fix the number k of the nearest neighbours rather than the radius r of the hypersphere. The maximum likelihood estimator given in the formula (1) then becomes

$$d_k\left(X_i\right)=\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(X_i)}{T_j(X_i)}\right]^{-1}, \quad (2)$$

where $T_k(X_i)$ represents the radius of the smallest hypersphere with the centre X_i that covers k neighbouring data points. Levina and Bickel (2005) show that one could divide by $k-2$, rather than $k-1$, to make the estimator asymptotically unbiased:

$$d_k\left(X_i\right)=\left[\frac{1}{k-2}\sum_{j=1}^{k-1}\log\frac{T_k(X_i)}{T_j(X_i)}\right]^{-1}. \quad (3)$$

It is clear from the above equations that the estimate depends on the parameter k (or the radius r of the hypersphere), and it also depends on the point X_i. Sometimes, the intrinsic dimensionality varies as a function of the location (X_i) in the data set, as well as the scale (k or r). Thus, it is a good idea to have estimates of the intrinsic dimensionality at different locations and scales.

Levina and Bickel (2005) assume that all the data points come from the same manifold, and therefore they average the estimated dimensions over all the observations (m is the number of data points):

$$d_k=\frac{1}{m}\sum_{i=1}^{m}d_k\left(X_i\right). \quad (4)$$

The choice of k affects the estimate. In general, for the MLE to work well, the hypersphere should be small and, at the same time, contain rather many points. Levina and Bickel (2005) choose the value of the parameter k automatically: in some heuristic way they simply average over a range of small to moderate values $k=k_1,\ldots,k_2$ to get the final estimate,

$$\widehat{d}_{\mathrm{MLE}}=\frac{1}{k_2-k_1+1}\sum_{k=k_1}^{k_2}d_k. \quad (5)$$

According to the experimental investigations, Levina and Bickel (2005) recommend $k_1=10$ and $k_2=20$. However, these estimates are valid for some fixed data sets only.

In the work of Karbauskaitė *et al*. (2011), the way of achieving good estimates of the intrinsic dimensionality by the MLE method is proposed. Since it is not known how to choose the values of the parameters k_1 and k_2 in the general case, by analysing the MLE algorithm, the authors use only one control parameter k, the number of the nearest neighbours for each data point, instead of two control parameters k_1 and k_2. The MLE algorithm is explored by evaluating two types of metrics: Euclidean and geodesic. In both the cases, the values d_k (4) of the MLE are calculated with different values k of the nearest neighbours. In such a way, dependences of the estimate of the intrinsic dimensionality of data on the number k of the nearest neighbours are obtained. The authors choose such a value d_k of the MLE that remains stable in a long interval of k. But this method requires a human participation in making a decision.

Levina *et al*. (2007) suggest to select the value of k equal to 20 on the basis of a dataset with the known number of pure components in the mixture of Raman spectroscopy data.

This problem is considered by Carter *et al*. (2010) as well. The authors state that one of the keys to local dimensionality estimation is defining the value of k. There must be a significant number of samples in order to obtain a proper estimate, but it is also important to keep a small sample size as to (ideally) only include samples that lie on the same manifold. Although the authors agree that a more definitive method of choosing k is necessary, Carter *et al*. (2010) arbitrarily choose k, based on the size of the data set.

3. Choice of the parameter r in the MLE

Levina and Bickel (2005) state that a more convenient way to estimate the intrinsic dimensionality of data is to fix

the number k of neighbours instead of the radius r of the hypersphere. As far as we know, everyone who has been investigating the MLE until now has used the formula (3), too. For our investigations, we use the formula (1), i.e., we fix the radius r of the hypersphere instead of the number k of neighbours. Then, we suggest averaging the estimated dimensions over all the m data points and get the final estimate:

$$\widehat{d}_{\mathrm{MLE}}=\frac{1}{m}\sum_{i=1}^{m} d_r\left(X_i\right). \tag{6}$$

In (6), the value of $\widehat{d}_{\mathrm{MLE}}$ is a real number. Assuming that the intrinsic dimensionality of a data set is an integer number, the value of $\widehat{d}_{\mathrm{MLE}}$ is rounded to the nearest integer. We denote this integer value by d_{MLE}.

Significant features of our contribution are as follows:

(i) We propose an automatic way to select the value of the parameter, i.e., the radius r of the hypersphere in the formula (1).

(ii) We show that the geodesic distances between data points must be used instead of the Euclidean ones when estimating the intrinsic dimensionality.

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 on the manifold. The search for the neighbours of each point X_i can be organized in two ways: (i) by the fixed number k_{geod} of the nearest points from X_i, (ii) by all the points within some fixed radius of a hypersphere whose center is the point X_i. When the neighbours are found, a weighed neighbourhood 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.

Let the radius r of the hypersphere be equal to the average distance calculated in the following way:

1. The distances $d(X_i, X_j)$, $i < j$, between all the data points X_i, $i = \overline{1, m}$, are calculated.

2. The distances $d(X_i, X_j)$ are distributed into l intervals $A_1, \ldots, A_l$.

3. A histogram is drawn with reference to the intervals $A_1, \ldots, A_l$ (the abscissa axis corresponds to distances, the ordinate axis corresponds to the frequency—the number of distances that fall in each interval).

4. The middle point $d(j)$, $j = \overline{1, l}$, of each interval is chosen and the number $f(j)$, $j = \overline{1, l}$, of distances in each interval is fixed.

5. The average distance between data points is given by the expected value of $D = \{d\left(j\right), j = \overline{1, l}\}$; i.e., the value of the parameter r is calculated by the formula

$$r = E\left(D\right) = \sum_{j=1}^{l} d\left(j\right) p(j), \tag{7}$$

where $p\left(j\right)$, $j = \overline{1, l}$, is a frequency estimator of probability given by the formula

$$p(j) = \frac{f(j)}{\sum_{z=1}^{l} f(z)}. \tag{8}$$

We chose the average pairwise distance as an estimate of r. This idea comes from probability theory and statistics. The expected value of a random variable is the integral of a random variable with respect to its probability measure. The expected value of a random variable is the average of all the values it can take, and thus the expected value is what one expects to happen on average. If the values of a random variable are not equally probable, then the simple average must be replaced by the weighted average, which takes into account the fact that some values are more likely than others. In our case, the weighted average is defined by the formula (7).

The algorithm proposed above has the parameter r whose value is calculated automatically by the formula (7), as well as two other parameters l and k_{geod} that should be set manually. We took $l = 100$ in the experiments. The value of l is chosen rather large because we do not try to optimize it, but we seek a more exact result. Moreover, the larger value of l would increase computing expenses. Since k_{geod} defines the number of the nearest neighbours used to construct a weighted graph while looking for geodesic distances, it is reasonable to pick the value of this parameter rather small, because too large a value of k_{geod} may lead to an inaccurate estimate of the intrinsic dimensionality. The structure of a nonlinear manifold is ignored when k_{geod} is too large; i.e., the nearest neighbours of a point may be points that are distant on the manifold. If the value of k_{geod} is too small, then the neighbourhood graph is not a connected one; i.e., there are some points between which there is no path in the graph. As a result, geodesic distances between all the data points cannot be calculated and the estimate d_{MLE} is not obtained. The experiments and detailed recommendations for selecting the value of k_{geod} are presented in Section 5.

In the next section, it is shown that good results are obtained if the geodesic distances are used and the value of the parameter r is calculated according to the formula (7).

An advantage of this algorithm, as compared with that described by Karbauskaitė *et al*. (2011), is that there is no need to have dependences of the estimate of the intrinsic dimensionality on the parameter, because we obtain the value of the parameter r automatically. These dependences (Figs. 3, 5, 7, 9, 11, 13, 15) are drawn here to illustrate the place of the results (the value r of the average distance and the intrinsic dimensionality corresponding to this value r) among other possible values of distances only.

4. 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. 2(a)).
- 1000 three-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear two-dimensional 8-shaped manifold (Fig. 2(b)). 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.
- 1801 three-dimensional data points ($m = 1801$, $n = 3$) that lie on a nonlinear two-dimensional manifold—right helicoid (Fig. 2(c)). The components $(x_1,\ x_2,\ x_3)$ of these data are calculated by the parametric equations below:

$$\begin{aligned} x_1 &= u\cos(v),\\ x_2 &= u\sin(v),\\ x_3 &= 0.5v, \end{aligned}$$

  where $u, v \in [0, 10\pi]$.
- 1801 three-dimensional data points ($m = 1801$, $n = 3$) that lie on a nonlinear one-dimensional manifold—spiral (Fig. 2(d)). 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 $t \in [0, 10\pi]$.
- 1000 three-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear one-dimensional manifold—helix (Fig. 2(e)). 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.
- *A data set of uncoloured (greyscale) pictures of a rotated duckling* (Nene *et al*., 1996) (samples of pictures are shown in Fig. 2(f)). The data are comprised of uncoloured 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 intensity value (shade of grey) is from the interval $[0, 255]$ (source database: `www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`. The coloured analogue of the set of rotating duckling is presented at `www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`). The number of pictures (data points) is $m = 72$. The images have 128×128 colour pixels, therefore the dimensionality of points characterizing each picture in a multidimensional space is three times larger compared with the greyscale pictures, i.e., $n = 49152$, and the colour value is from the interval [0, 255].
- *Data sets of coloured pictures of rotated objects* (Nene *et al*., 1996). The data are comprised of coloured pictures of the same object, obtained by gradually rotating it at the 180° angle. Each picture is digitized; i.e., a data point consists of colour parameters of pixels, and, therefore, it is of very large dimensionality. The number of pictures (data points) is $m = 35$. The images have 128×128 colour pixels, therefore $n = 49152$. At `www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`, 100 data sets of this type are stored.
- *A data set of uncoloured (greyscale) images of a person's face* (Tenenbaum *et al*., 2000) (samples of images are shown in Fig. 2(h)). The data consist of many photos of the same person's face observed in different poses (left-and-right pose, up-and-down

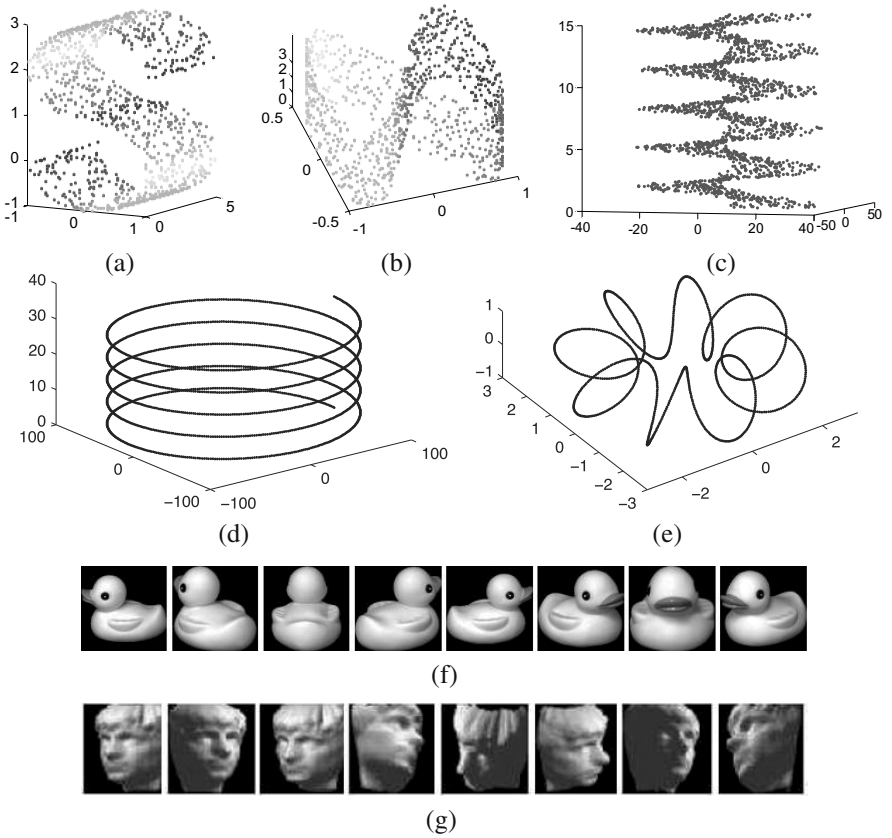

Fig. 2. Data sets of manifolds: S-shaped manifold (a), 8-shaped manifold (b), right helicoid (c), spiral (d), helix (e), pictures of a rotated duckling (f), images of a person's face (g).

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: `isomap.stanford.edu/datasets.html`).

5. Experimental exploration of the MLE

In this section, the MLE method is investigated experimentally with various artificial and real data sets. The analysed data points of artificial data sets (Figs. 2(a)–(e)) lie on manifolds, whose dimensionality is known in advance. Therefore, we will be able to establish precisely whether the estimate of the intrinsic data dimensionality obtained by the MLE is true. As a result, we will be able to disclose the relation between the intrinsic dimensionality of the data set of images and the number of degrees of freedom of a possible motion of the object.

The aim of these investigations is to find out (i) which distances (Euclidean or geodesic) are better to be used in the MLE algorithm while estimating the similarity between data points and (ii) how to select the value of the parameter r in order to get the true estimate of the intrinsic dimensionality of data using the MLE.

5.1. Analysis of artificial data sets. The first investigation is performed with the points of the two-dimensional S-shaped manifold ($m = 1000, n = 3$), (Fig. 2(a)). First, after estimating the distances (Euclidean or geodesic ($k_{\mathrm{geod}} = 5$)) between all the data points, dependences of the estimate d_{MLE} on those distances, i.e., the possible values of the parameter r, are calculated. In Fig. 3, when the value of the distance varies from the least to the largest one, the estimate of the intrinsic dimensionality obtained by the MLE acquires two values: 1 or 2. This means that the value of d_{MLE} depends on the distance. That is valid in both cases: (a) the Euclidean distance, (b) geodesic distance. In Fig. 4, histograms of the distribution of distance values ((a) the Euclidean and (b) the geodesic) between the points of the S-shaped manifold are shown. The frequency of various distances is different. The average distance (the value of the control parameter r) is calculated by the formula (7). In the case of the Euclidean distances, $r = 2.81$, $d_{\mathrm{MLE}} = 2$, and, in the case of the geodesic distances $r = 4.42$,

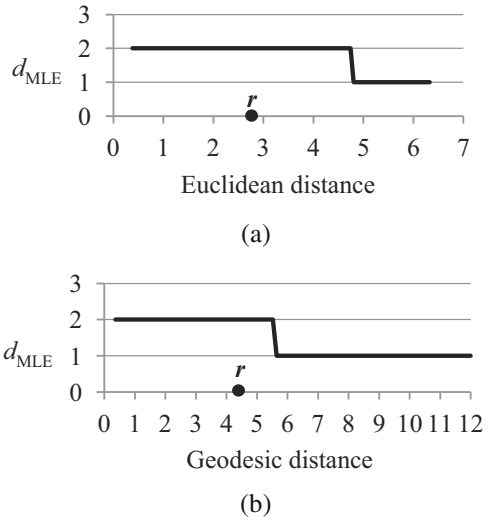

Fig. 3. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic, $k_{\mathrm{geod}} = 5$ (b)); data set: the S-shaped manifold.

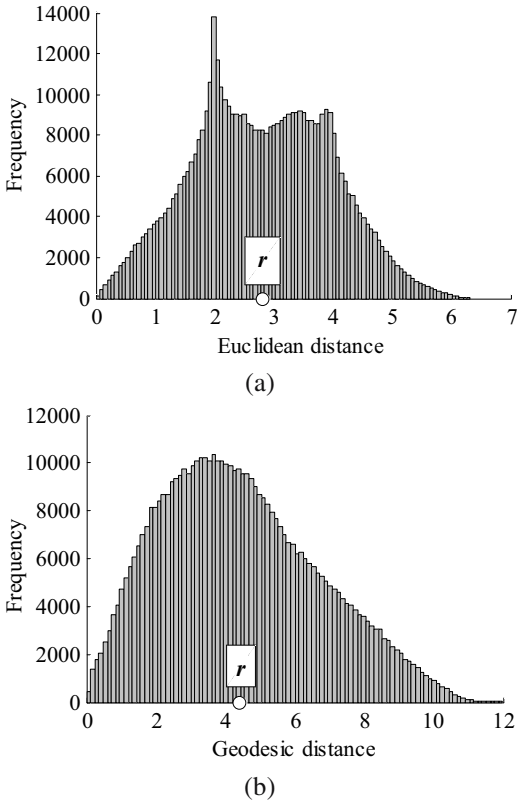

Fig. 4. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 5$) (b) distances between the data points of the S-shaped manifold.

$d_{\mathrm{MLE}} = 2$. So, in both cases, the intrinsic dimensionality of data points of the two-dimensional S-shaped manifold is evaluated truly by the MLE.

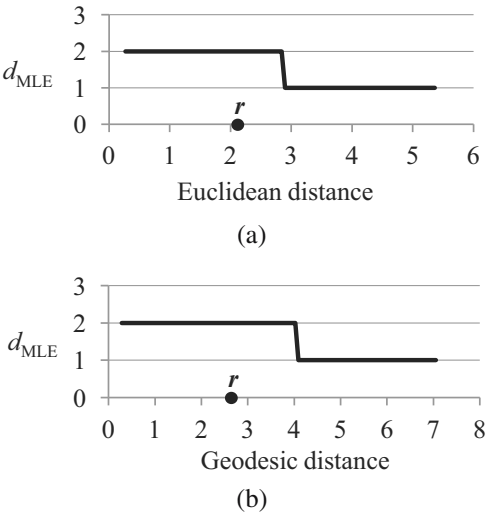

Fig. 5. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic, $k_{\mathrm{geod}} = 5$ (b)); data set: the 8-shaped manifold.

The second investigation is performed with the points of the two-dimensional 8-shaped manifold ($m = 1000$, $n = 3$); see Fig. 2(b). The third investigation is performed with the points of the two-dimensional manifold—helicoid ($m = 1801$, $n = 3$); see Fig. 2(c). The fourth investigation is performed with the points ($m = 1801$, $n = 3$) of a spiral that is a one-dimensional manifold (Fig. 2(d)). The results are shown in Figs. 5–10 and Table 1. As in the first experiment, in these three experiments we see that the intrinsic dimensionality of the data sets is evaluated exactly by the MLE, using both the Euclidean and geodesic distances.

The fifth investigation is performed with the points ($m = 1000, n = 3$) of the helix that is a one-dimensional manifold (Fig. 2e). The results are given in Figs. 11, 12 and Table 1. Figure 11(a) shows that the estimate of the intrinsic dimensionality obtained by the MLE while evaluating the Euclidean distances has the values $\{1,\ 2,\ 6\}$. However, if the geodesic distances ($k_{\mathrm{geod}} = 2$) are used, the obtained value of the estimate is a single one and $d_{\mathrm{MLE}} = 1$. In Fig. 12, histograms of the distribution of the distance values ((a) Euclidean, (b) geodesic) between the points of the helix are shown. We can see that the geodesic distances are distributed almost uniformly. However, this cannot be said about the Euclidean distances. The value of the average distance, i.e., the value of the control parameter r, is calculated by the formula (7). In the case of the Euclidean distances, $r = 2.92$, $d_{\mathrm{MLE}} = 2$, and, in the case of the geodesic distances, $r = 13.01$, $d_{\mathrm{MLE}} = 1$. Consequently, we have the case where the value of d_{MLE} is false if the Euclidean distances are used. However, the intrinsic dimensionality of the helix is evaluated exactly by the

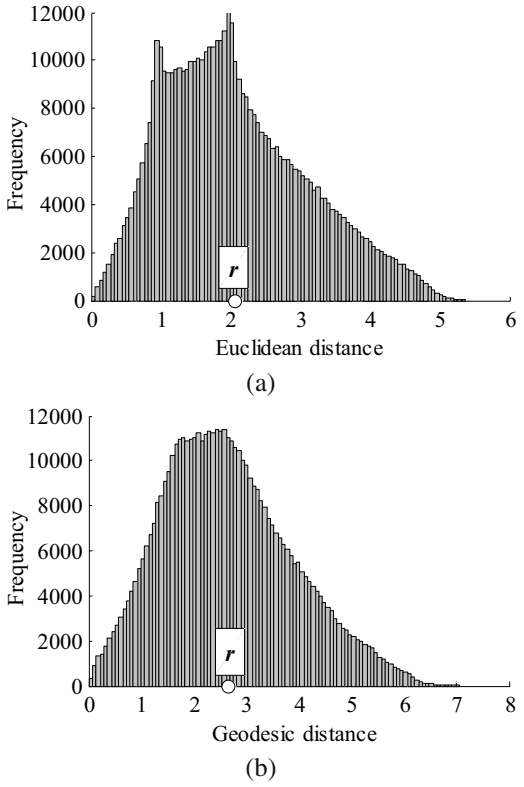

Fig. 6. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 5$) (b) distances between the data points of the 8-shaped manifold.

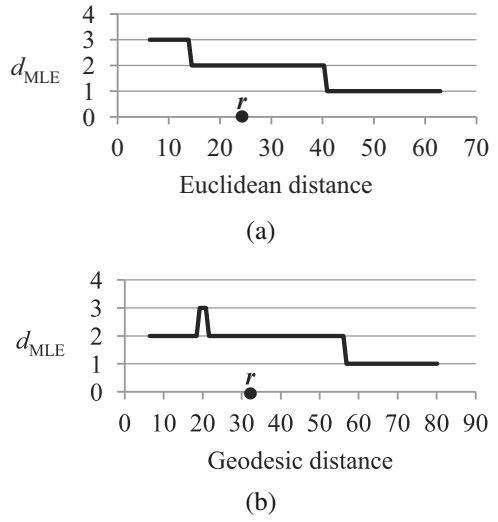

Fig. 7. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a) and geodesic, $k_{\mathrm{geod}} = 5$ (b)); data set: the helicoid.

MLE if the geodesic distances are evaluated.

5.2. Analysis of images. A challenging idea is to apply the manifold learning methods to high-dimensional data.

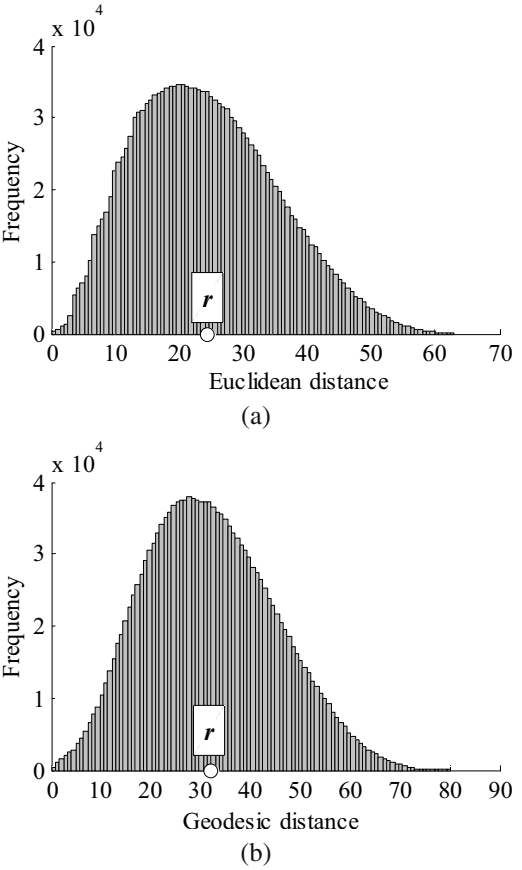

Fig. 8. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 5$) (b) distances between the data points of the helicoid.

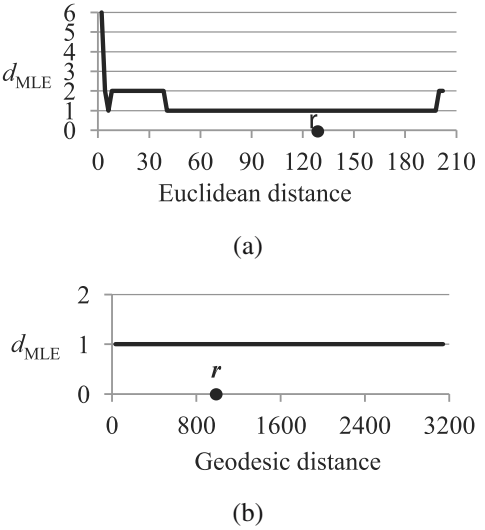

Fig. 9. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic, $k_{\mathrm{geod}} = 2$ (b)); data set: the spiral.

One of the fields where such data appear is the analysis of images. Let us have a set of images of some moving

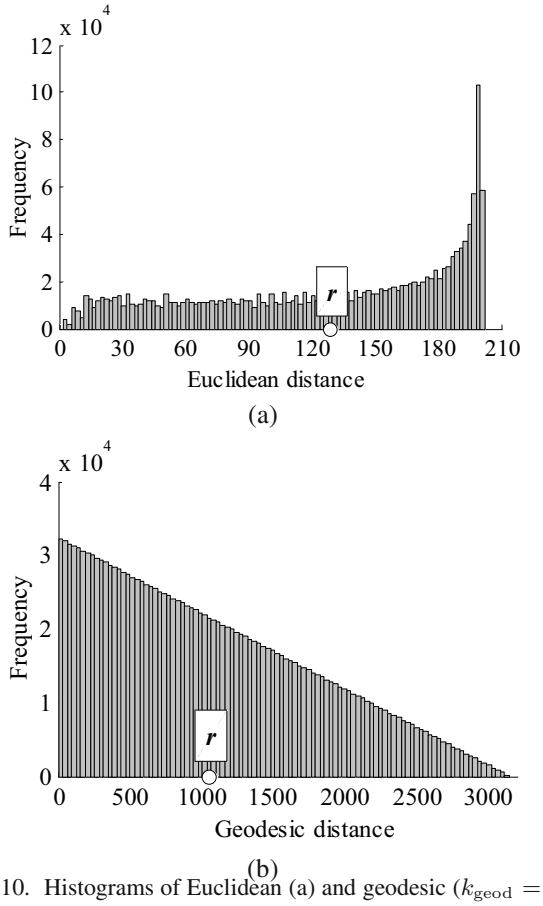

Fig. 10. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 2$) (b) distances between the data points of the spiral.

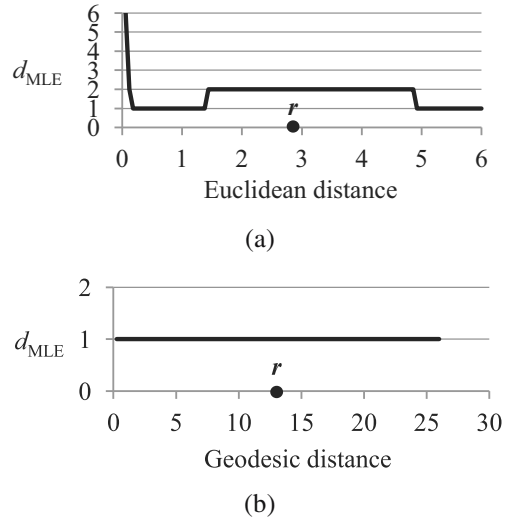

Fig. 11. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic, $k_{\mathrm{geod}} = 2$ (b)); data set: the helix.

object. 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 greyscale case, or

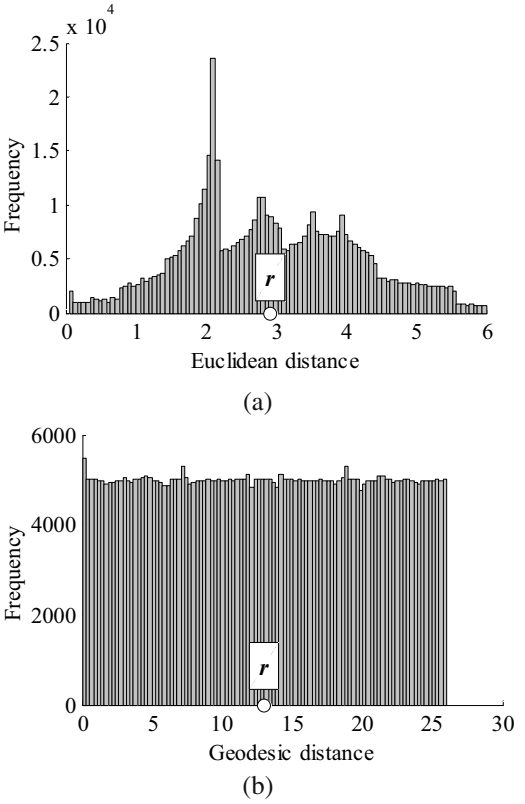

Fig. 12. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 2$) (b) distances between the data points of the helix.

it is even 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 (Lee and Verleysen, 2007), one can assume that there are latent variables or features that characterize the motion of the object in the images and their number is highly related to that 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. However, the true intrinsic dimensionality may be larger than the number of degrees of freedom of a possible motion of the object.

The high-dimensional data obtained from the set of images (greyscale or coloured pictures of a rotated duckling and photos of the same person's face observed in different poses) are investigated. Since a duckling was gradually rotated at a certain angle on the same plane, that is, 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 by Tenenbaum *et al.* (2000), has two directions of motion (two poses): the

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 the investigation with the high-dimensional data points ($m = 72$, $n = 16384$), corresponding to real pictures of a rotated duckling (Fig. 2(f)), are given in Figs. 13 and 14. Like in the previous investigations, two dependences of the estimate d_{MLE} on the distances that are possible values of r are calculated. Figure 13 shows that the estimate of the intrinsic dimensionality obtained by the MLE, acquires various values in both cases: (a) $d_{\mathrm{MLE}} \in \{2, 3, 4, 5, 6\}$ (Euclidean case (Fig. 13(a))) and (b) $d_{\mathrm{MLE}} \in \{1, 2, 3\}$ (geodesic case, $k_{\mathrm{geod}} = 3$ (Fig. 13(b))). Thus, the value of d_{MLE} strongly depends on the chosen value of r. In Fig. 14, histograms of the distribution of the distance values ((a) Euclidean, (b) geodesic) between the points of the data are shown. The value of the control parameter r (average distance) is calculated by the formula (7). In the case of the Euclidean distances, $r = 10243$, $d_{\mathrm{MLE}} = 3$, and, in the case of the geodesic distances, $r = 35798$, $d_{\mathrm{MLE}} = 1$.

Let us analyse the colour pictures of a rotated duckling instead of greyscale ones and measure the intrinsic dimensionality of this data set. The results are as follows: in the case of the Euclidean distances, $r = 11260$, $d_{\mathrm{MLE}} = 3$, and in the case of the geodesic distances ($k_{\mathrm{geod}} = 3$), $r = 54919$, $d_{\mathrm{MLE}} = 1$. The results obtained show that, in the case of a rotated duckling, the presence of colours in the pictures does not influence the estimate of the intrinsic dimensionality as compared with the greyscale case. Taking into account that the minimal intrinsic dimensionality of these data is 1, the value of d_{MLE} is false if the Euclidean distances are used. However, the minimal intrinsic dimensionality of these data is evaluated truly by the MLE if the geodesic distances between the data points are evaluated.

The next investigation is performed with the high-dimensional data points ($m = 698$, $n = 4096$), corresponding to photos of the same person's face observed in different poses with a different lighting direction (Fig. 2(g)). We can see the results in Figs. 15 and 16. In the case of the Euclidean distances, we obtained $r = 20.19$, $d_{\mathrm{MLE}} = 4$. When investigating the case of geodesic distances, we used different values of k_{geod}, i.e., the fixed number k_{geod} of the nearest points from each data point X_i in the geodesic distance calculation algorithm. We obtained $r = 72.64$ as $k_{\mathrm{geod}} = 4$, $r = 49.06$ as $k_{\mathrm{geod}} = 10$, and $r = 45.78$ as $k_{\mathrm{geod}} = 13$, but in all these cases $d_{\mathrm{MLE}} = 2$ (see Fig. 15(b)–(d)). Obviously, in Fig. 15, we got different dependences of the intrinsic dimensionality on geodesic distances, using various values of the parameter k_{geod}. In the case of (b), we see two possible values of d_{MLE}: 2 and 1. In the case of (c), there are three possible values of d_{MLE}: 3, 2, and 1. In the case of (d), it is possible to get even four values

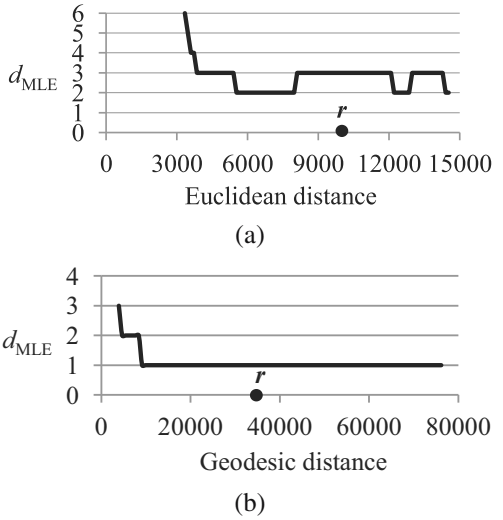

Fig. 13. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic, $k_{\mathrm{geod}} = 3$) (b); data set: pictures of a rotated duckling.

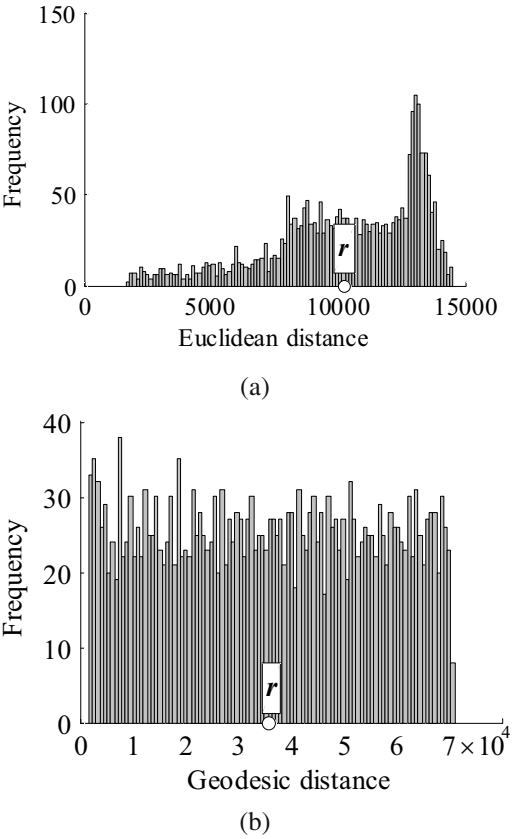

Fig. 14. Histograms of Euclidean (a) and geodesic ($k_{\mathrm{geod}} = 3$) (b) distances between the data points corresponding to the pictures of a rotated duckling.

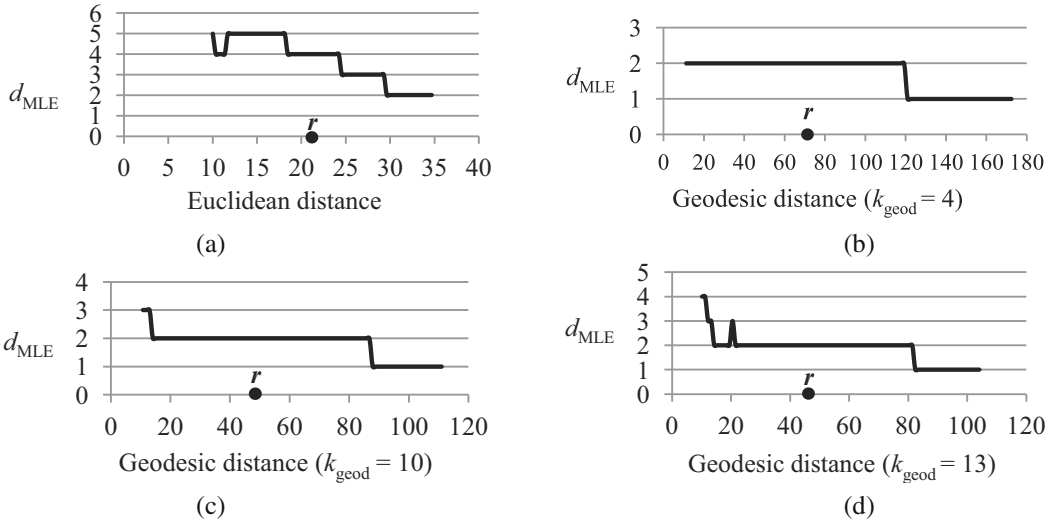

Fig. 15. Estimate of the intrinsic dimensionality depending on distances (Euclidean (a), geodesic (b), (c), (d)); data set: photos of a face.

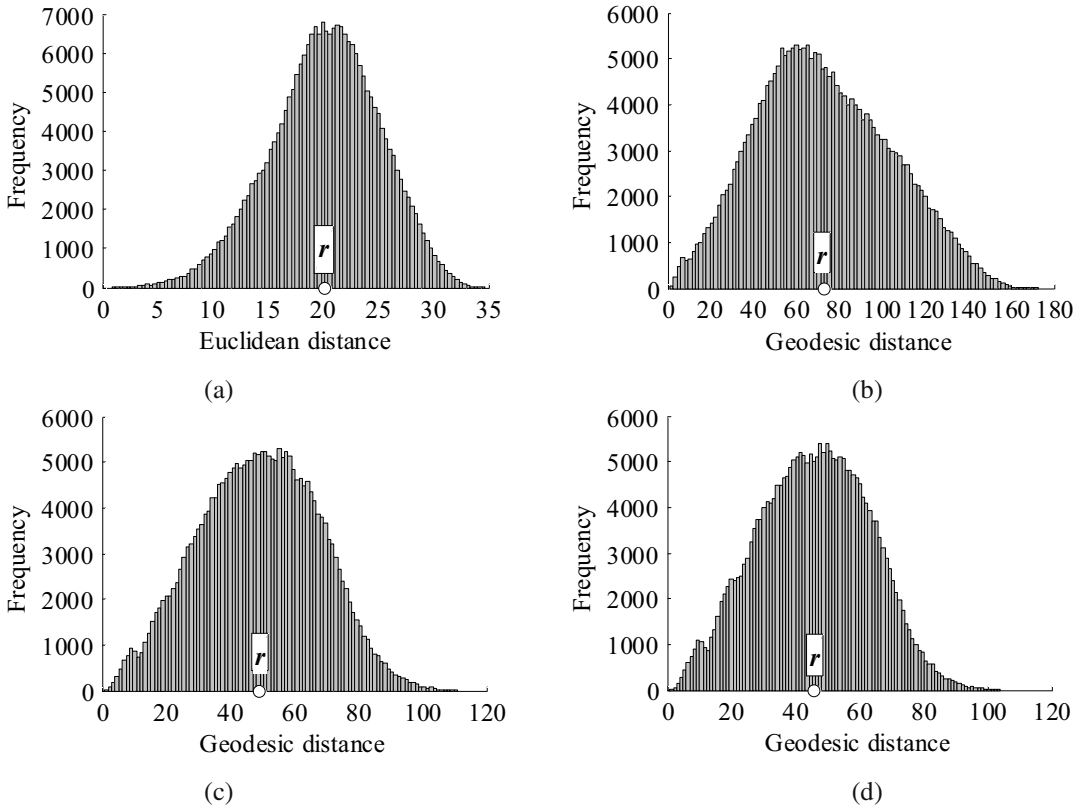

Fig. 16. Histograms of Euclidean (a) and geodesic distances ($k_{\mathrm{geod}} = 4$ (b), $k_{\mathrm{geod}} = 10$ (c), $k_{\mathrm{geod}} = 13$ (d)) between the data points corresponding to the photos of a face.

of d_{MLE} with different distances: 4, 3, 2, and 1. It seems as though the value of the parameter k_{geod} influences the results obtained. However, despite different dependences of d_{MLE} on geodesic distances using different values of k_{geod} (see different cases with the face data set in Fig. 15(b)–(d)), the new way, proposed in Section 3, of choosing the value of the parameter r automatically in the MLE method yields the same value of $d_{\mathrm{MLE}} = 2$ in all these cases, and this value is coincident with the number of degrees of freedom of the face in the photos. So, there

is no dependence of d_{MLE} on k_{geod} in these cases; i.e., in one particular case (face data set), the estimated intrinsic dimensionality d_{MLE} is the same for three different values of k_{geod}. But what about other datasets? The experiments with other data sets from Section 4 are carried out and are concluded in Section 5.3.

The intrinsic dimensionality of the face data set (Tenenbaum *et al*., 2000) is analysed in several papers. It is shown by Tenenbaum *et al*. (2000) that the intrinsic dimensionality of this data set is 3. Levina and Bickel (2005) state that the estimated dimensionality of about 4 is very reasonable. In the work of Karbauskaitė *et al*. (2011), 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 the MLE. A question arises as to which estimated dimensionality can be taken as the true intrinsic dimensionality.

In order to answer this question, let us analyse the face database (Tenenbaum *et al*., 2000) in detail. At first, the 4096-dimensional data points are projected on the 5-dimensional space by the ISOMAP method (Tenenbaum *et al*., 2000). ISOMAP is used in the investigation because currently it is one of the most popular manifold learning methods. Thus, we get a matrix of dimensions 698×5. 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 $y_1, y_2, \ldots, y_{n^*}$, $n^* = 5$, which characterize the objects. Then the covariance matrix C of the features is obtained:

$$C = \begin{pmatrix} 1538.8 & 0 & 0 & 0 & 0 \\ 0 & 419.3 & 0 & 0 & 0 \\ 0 & 0 & 276.3 & 0 & 0 \\ 0 & 0 & 0 & 86.8 & 0 \\ 0 & 0 & 0 & 0 & 79.1 \end{pmatrix}. \quad (9)$$

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 feature y_k. We see from (9) and Fig. 17 that the variances of the first three features are much larger than others. The variances of the fourth and fifth features are more than three times smaller than the variance of the third one. It means that there are three main features, but they are not the only ones. Therefore, the estimated dimension of about 4 or 5 is very reasonable. A question arises as to 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 the first three features pairwise on the plane (see Figs. 18–20). Figures 18–20 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 first two features, i.e., both poses (left-and-right and up-and-down), are more essential than the third feature that corresponds to the lighting direction.

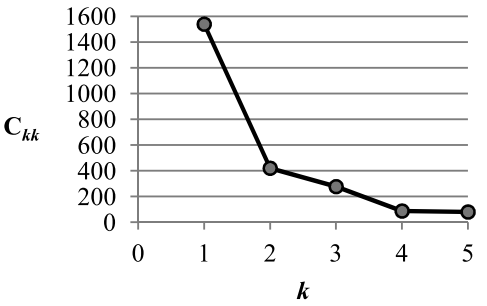

Fig. 17. Variances of features.

Since the face database consists of images of an artificial face under three changing conditions: vertical and horizontal orientation as well as illumination (lighting direction), it is possible to assume that the intrinsic dimensionality of this data set should be 3. The person's face has two directions of motion (two poses): the left-and-right pose and the up-and-down pose. So, the minimal intrinsic dimensionality of these data can be assigned to 2, which is the number of degrees of freedom of a possible motion of the object in the image. Of course, the true intrinsic dimensionality is larger. However, the most essential dimensions correspond to the directions of motions. Thus, after such a discussion, we dare say that, in this investigation, the minimal intrinsic dimensionality of these data is evaluated well if the geodesic distances between the data points are calculated.

The next investigation is based on the sets of coloured pictures of rotated objects (Nene *et al*., 1996). The real-valued estimates $\widehat{d}_{\mathrm{MLE}}$ of the intrinsic dimensionality of various data sets were calculated in both cases where (a) Euclidean distances and (b) geodesic ($k_{\mathrm{geod}} = 3$) distances between data points are used in the MLE algorithm (Fig. 21). The average of the intrinsic dimensionality estimates is 4.23 in the case (a) and 1.26 in the case (b). Figure 21(a) shows that the estimates $\widehat{d}_{\mathrm{MLE}}$ of the intrinsic dimensionality of all the data sets analysed are larger than 2. It is easy to notice in Fig. 21(b) that, in the case of the geodesic distances, only several samples of the analysed data sets obtain the estimate $\widehat{d}_{\mathrm{MLE}}$ of the intrinsic dimensionality that is larger than 1.5, i.e., $d_{\mathrm{MLE}} = 2$. Most data sets have the estimate lower than 1.5, i.e., $d_{\mathrm{MLE}} = 1$. As the particular data set analysed consists of a set of pictures of a rotated object with the degree of freedom equal to 1, the obtained d_{MLE} value is equal to that degree of freedom in most cases. This fact cannot be stated when the Euclidean distances are used.

5.3. Recommendations for selecting k_{geod}. Our realization of the MLE method needs calculation of geodesic distances between the points of the analysed data set. To this end, we need to set the number k_{geod} of the nearest neighbours which are used to construct a weighted graph while looking for geodesic distances. In Section 5.2, we did not observe the dependence of d_{MLE} on k_{geod} in the case of the face data set when k_{geod} has the values 4, 10 and 13. But what are general recommendations for selecting the values of k_{geod}?

From Table 2 with other data sets we see that the value of d_{MLE} obtained by the algorithm proposed in Section 3 does not depend on the chosen value of k_{geod}, if it is not large. However, we should note that the estimate d_{MLE} cannot be calculated with very small values of k_{geod}. For one-dimensional manifolds, k_{geod} cannot be set equal to one, and, in the case of two-dimensional manifolds, the value of k_{geod} cannot be set as 1, 2, 3 or even 4 (for the helicoid). The reason is that the neighbourhood graph over all the points of the analysed data set appears not to be connected. Let us show that by the example of the one-dimensional manifold (helix). A neighbourhood graph over the points of the helix is constructed: each point is connected with its k_{geod} nearest neighbours. The neighbourhood graphs with various values of k_{geod} are drawn in Fig. 22.

As shown in Table 2, the estimate d_{MLE} is not calculated if $k_{\mathrm{geod}} = 1$. The reason is that the graph is not a connected graph and it is impossible to find any path from some points to other particular points in the graph in this case (see Fig. 22(a)). A graph is a connected one if there is a path from any point to any other point in the graph. As a result, the geodesic distances between a part of data points cannot be calculated and the estimate d_{MLE} is not obtained. However, the connected graphs were obtained in the remaining cases (Fig. 22 (b)–(d)). So, the values of d_{MLE} can be calculated in all the three cases. It is worth noticing that the neighbourhood graph in the case (d) is different from the graphs in the cases (b) and (c). If the value of k_{geod} is rather large, the neighbours of each data point may be found wrongly, i.e., false neighbours may be set to some data points. The reason is as follows: since the Euclidean distances are calculated between a point and its neighbours, the nearest neighbours of the point may be points that are distant on the manifold; i.e., the structure of a nonlinear manifold is ignored if k_{geod} is too large. In Fig. 22(d), we see the graph with a false neighbourhood. The wrong neighbourhood graph and obtained geodesic distances between data points according to this graph may influence the false estimates d_{MLE} of the intrinsic dimensionality. It is obvious in Fig. 23. If $k_{\mathrm{geod}} = 2, \ldots, 30$, the estimate $d_{\mathrm{MLE}} = 1$ that is the true intrinsic dimensionality of the helix but the false value of the intrinsic dimensionality is obtained starting from $k_{\mathrm{geod}} = 31$.

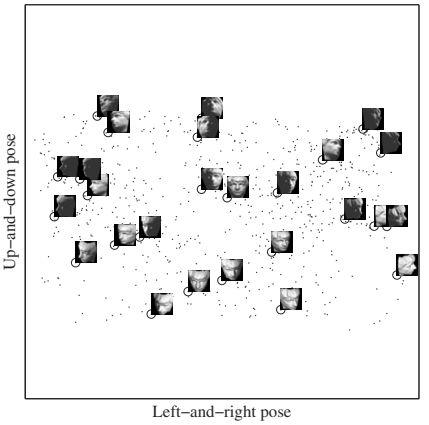

Fig. 18. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: left-and-right pose, up-and-down pose.

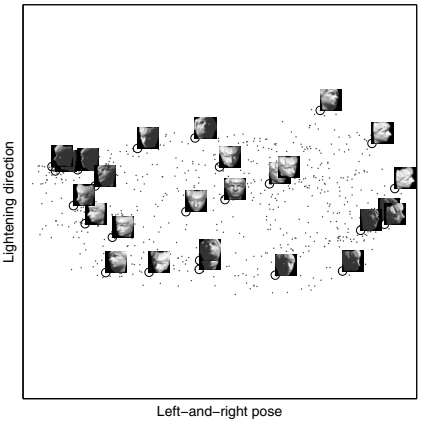

Fig. 19. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: left-and-right pose, lightening direction.

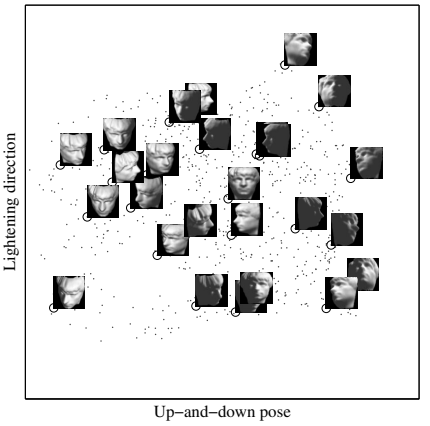

Fig. 20. Projections of the high-dimensional data points corresponding to the photos of a face on a plane: up-and-down pose, lightening direction.

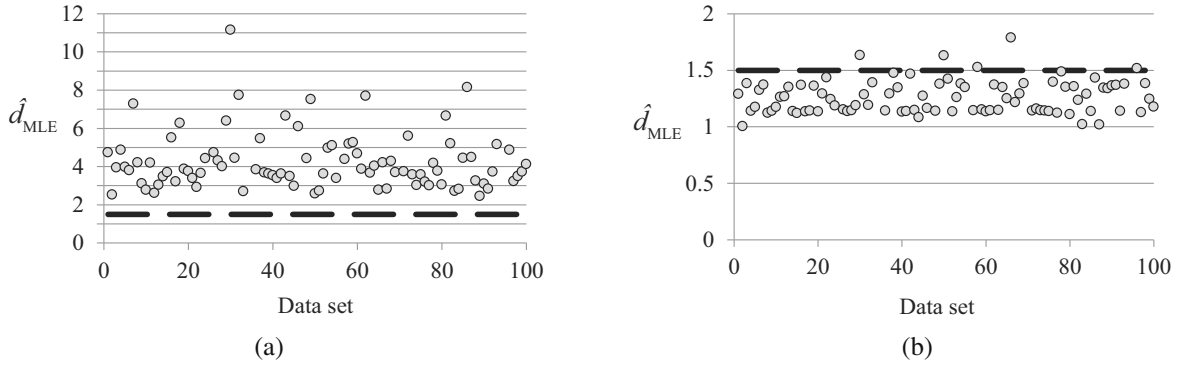

Fig. 21. Estimate $\widehat{d}_{\mathrm{MLE}}$ of the intrinsic dimensionality of data sets of coloured pictures of a rotated object: Euclidean distances (a) and geodesic distances ($k_{\mathrm{geod}} = 3$) (b).

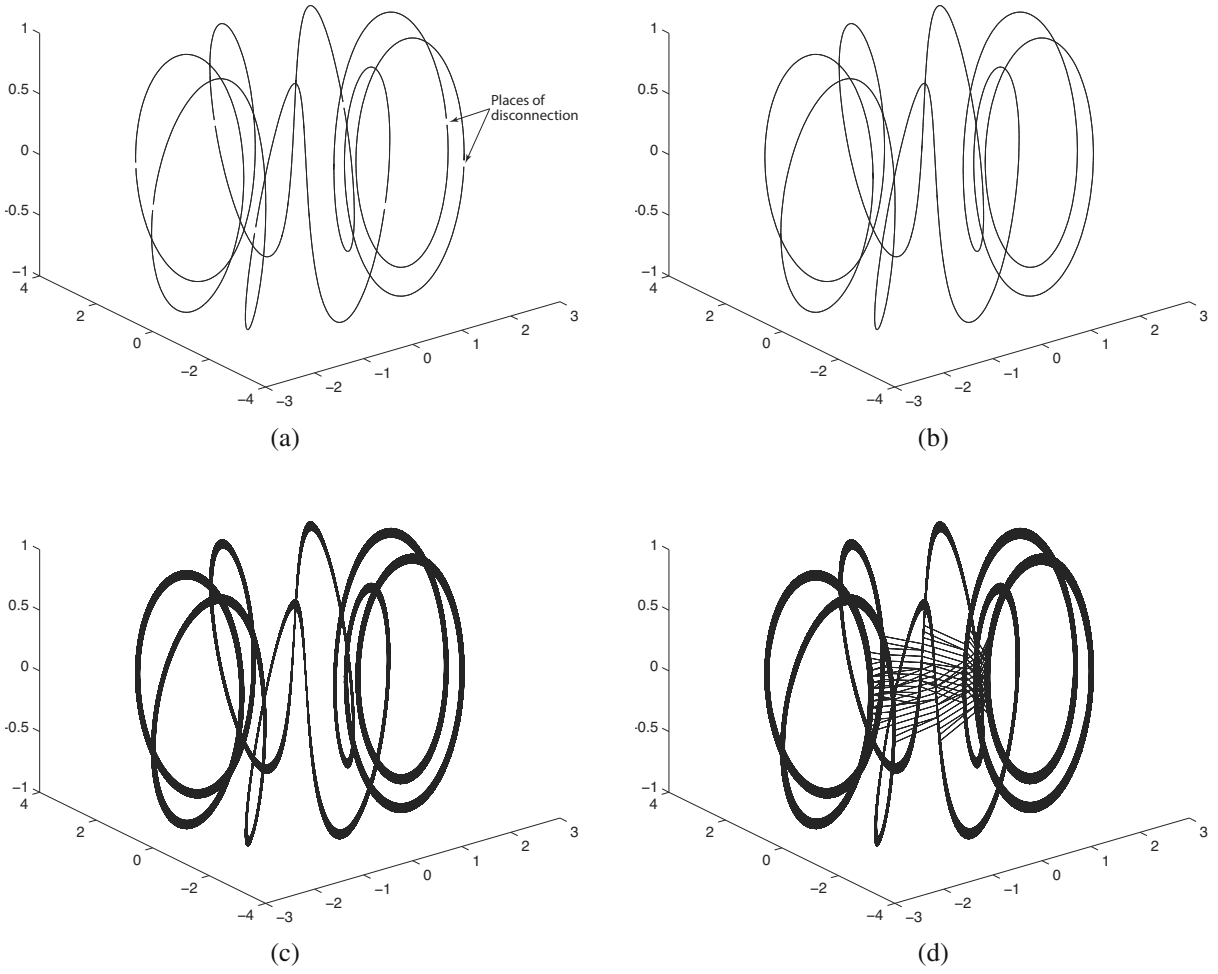

Fig. 22. Neighbourhood graphs with the various numbers k_{geod} of the nearest neighbours of each data point: $k_{\mathrm{geod}} = 1$ (a), $k_{\mathrm{geod}} = 2$ (b), $k_{\mathrm{geod}} = 30$ (c), $k_{\mathrm{geod}} = 31$ (d); data set: the helix.

Thus, our recommendation for selecting k_{geod} is as follows: one should pick out the value of this parameter rather small, i.e., such that the neighbourhood graph would be connected and the value of d_{MLE} calculated, because too large a value of k_{geod} may lead to an inaccurate estimate of the intrinsic dimensionality.

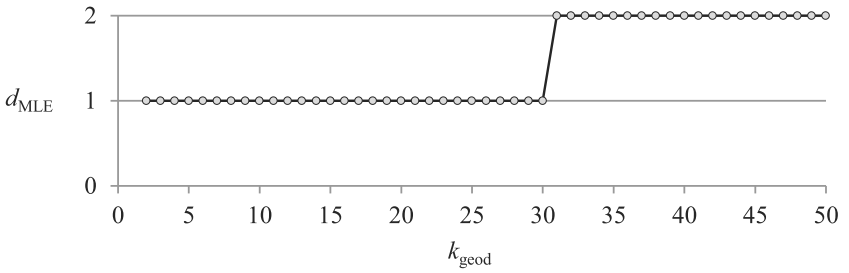

Fig. 23. Estimate d_{MLE} of the intrinsic dimensionality depending on the number k_{geod} of the nearest neighbours in the calculation algorithm of geodesic distances; data set: the helix.

Table 1. Estimates of the average distance r and the intrinsic dimensionality d_{MLE}.

Data sets	Intrinsic dimensionality d	Euclidean distances		Geodesic distances	
		r	d_{MLE}	r	d_{MLE}
S-shaped manifold	2	2.81	2	4.42	2
8-shaped manifold	2	2.07	2	2.65	2
Helicoid	2	24.25	2	31.96	2
Spiral	1	128.51	1	1048.70	1
Helix	1	2.92	2	13.01	1
Greyscale pictures of a rotated duckling	1	10243	3	35798	1
Coloured pictures of a rotated duckling	1	11260	3	54919	1
Photos of a person's face	3	20.19	4	72.64	2

6. Conclusions

Image analysis—face pose detection, face recognition, the analysis of facial expressions, human motion data interpretation, gait analysis, medical data analysis – is a very challenging topic in exploratory data analysis. The data points obtained from the images are of very high dimensionality, because the picture is digitized, i.e., a data point consists of colour parameters of pixels. However, such high-dimensional data can be efficiently summarized in a space of much lower dimensionality, that is, on a nonlinear manifold, because high-dimensional data sets can have meaningful low-dimensional structures hidden in the observation space (i.e., the data are of low intrinsic dimensionality). The knowledge of the intrinsic dimensionality of a data set is very useful information in exploratory data analysis.

In this paper, one of the local estimators of the intrinsic dimensionality—the maximum likelihood estimator (MLE)—has been analysed and developed. As far as we know, everyone who has been investigating the MLE until now has used the formula (2) or (3), i.e., they selected the number k (or an interval of k) of the nearest neighbours in the MLE. In this paper, we suggest to use the formula (1), i.e., we suggest to fix the radius r of the hypersphere that covers neighbours of the analysed points instead of the number k of the nearest neighbours. A new way of choosing the value of the parameter r in the MLE method has been proposed. It enables us to find the true value of this parameter by the formula (7).

An advantage of this approach as compared with that described by Karbauskaitė *et al*. (2011) is that there is no need to draw dependences of the estimate of the intrinsic dimensionality on the distances and to make some human decisions, because we get the value of the parameter r automatically. In our investigations, we have discovered that the number of latent variables is highly related to that of degrees of freedom of a possible motion of the object. Therefore, 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. However, the true intrinsic dimensionality may be larger than the number of degrees of freedom of a possible motion of the object. With a view to discover the influence of illumination and colours on the estimates of the intrinsic dimensionality, we have analysed both greyscale and coloured image data sets. The results have shown that the presence of colours in the pictures does not influence the estimate of the intrinsic dimensionality, as compared with the greyscale case. We have also explored which distances—Euclidean or geodesic—should be evaluated between the data points in the MLE algorithm. The obtained results are generalised in Table 1. The experiments with different data sets,

Table 2. Estimates d_{MLE} of the intrinsic dimensionality obtained using the algorithm proposed in Section 3.

Data sets	k_{geod}									
	1	2	3	4	5	6	7	8	9	10
S-shaped manifold	–	–	–	2	2	2	2	2	2	2
8-shaped manifold	–	–	–	–	2	2	2	2	2	2
Helicoid	–	–	–	2	2	2	2	2	2	2
Spiral	–	1	1	1	1	1	1	1	1	1
Helix	–	1	1	1	1	1	1	1	1	1
Greyscale pictures of a rotated duckling	–	1	1	1	1	1	1	1	1	1

especially real data sets (images), have shown that the MLE provides the right estimate of the intrinsic dimensionality if the geodesic distances are used and the value of the parameter r is equal to the expected value of the distances.

The knowledge of the intrinsic dimensionality may be very useful in the visualization of high-dimensional data. Further research should be focused on the accuracy of dimensionality reduction using the estimates of the intrinsic dimensionality.

Acknowledgment

The authors are very grateful to the referees for careful reading of the paper, valuable remarks and suggestions that improved the quality of this work.

References

Álvarez-Meza, A.M., Valencia-Aguirre, J., Daza-Santacoloma, G. and Castellanos-Domínguez, G. (2011). Global and local choice of the number of nearest neighbors in locally linear embedding, *Pattern Recognition Letters* **32**(16): 2171–2177.

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

Brand, M. (2003). Charting a manifold, *in* S. Becker, S. Thrun and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA, pp. 961–968.

Camastra, F. (2003). Data dimensionality estimation methods: A survey, *Pattern Recognition* **36**(12): 2945–2954.

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

Chang, Y., Hu, C. and Turk, M. (2004). Probabilistic expression analysis on manifolds, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR(2), Washington, DC, USA*, pp. 520–527.

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

Costa, J.A. and Hero, A.O. (2005). Estimating local intrinsic dimension with k-nearest neighbor graphs, *IEEE Transactions on Statistical Signal Processing* **30**(23): 1432–1436.

Donoho, D.L. and Grimes, C. (2005). Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Sciences* **102**(21): 7426–7431.

Dzemyda, G., Kurasova, O. and Žilinskas, J. (2013). *Multidimensional Data Visualization: Methods and Applications*, Optimization and Its Applications, Vol. 75, Springer-Verlag, New York, NY.

Einbeck, J. and Kalantan, Z. (2013). Intrinsic dimensionality estimation for high-dimensional data sets: New approaches for the computation of correlation dimension, *Journal of Emerging Technologies in Web Intelligence* **5**(2): 91–97.

Elgammal, A. and su Lee, C. (2004a). Inferring 3d body pose from silhouettes using activity manifold learning, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR(2), Washington, DC, USA*, pp. 681–688.

Elgammal, A. and su Lee, C. (2004b). Separating style and content on a nonlinear manifold, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR(1), Washington, DC, USA*, pp. 478–485.

Fan, M., Zhang, X., Chen, S., Bao, H. and Maybank, S.J. (2013). Dimension estimation of image manifolds by minimal cover approximation, *Neurocomputing* **105**: 19–29.

Fukunaga, K. (1982). Intrinsic dimensionality extraction, *in* P. Krishnaiah and L. Kanal (Eds.), *Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics*, Vol. 2, North-Holland, Amsterdam, pp. 347–362.

Fukunaga, K. and Olsen, D. (1971). An algorithm for finding intrinsic dimensionality of data, *IEEE Transactions on Computers* **20**(2): 176–183.

Gong, S., Cristani, M., Yan, S. and Loy, C.C. (Eds.) (2014). *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, Vol. XVIII, Springer, London.

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

Hadid, A., Kouropteva, O. and Pietikäinen, M. (2002). Unsupervised learning using locally linear embedding: experiments with face pose analysis, *16th International*

Conference on Pattern Recognition, ICPR'02(1), Quebec City, Quebec, Canada, pp. 111–114.

He, J., Ding, L., Jiang, L., Li, Z. and Hu, Q. (2014). Intrinsic dimensionality estimation based on manifold assumption, *Journal of Visual Communication and Image Representation* **25**(5): 740–747.

Hein, M. and Audibert, J. (2005). Intrinsic dimensionality estimation of submanifolds in r^d, *Machine Learning: Proceedings of the 22nd International Conference (ICML 2005), Bonn, Germany*, pp. 289–296.

Jenkins, O.C. and Mataric, M.J. (2004). A spatio-temporal extension to isomap nonlinear dimension reduction, *21st International Conference on Machine Learning, ICML(69), Banff, Alberta, Canada,* pp. 441–448.

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

Karbauskaitė, R. and Dzemyda, G. (2014). Geodesic distances in the intrinsic dimensionality estimation using packing numbers, *Nonlinear Analysis: Modelling and Control* **19**(4): 578–591.

Karbauskaitė, R., Dzemyda, G. and Marcinkevičius, V. (2008). Selecting a regularization parameter in the locally linear embedding algorithm, *20th International EURO Mini Conference on Continuous Optimization and Knowledge-based Technologies (EurOPT2008), Neringa, Lithuania*, pp. 59–64.

Karbauskaitė, R., Dzemyda, G. and Marcinkevičius, V. (2010). Dependence of locally linear embedding on the regularization parameter, *An Official Journal of the Spanish Society of Statistics and Operations Research* **18**(2): 354–376.

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

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

Kégl, B. (2003). Intrinsic dimension estimation using packing numbers, *Advances in Neural Information Processing Systems, NIPS(15), Cambridge, MA, USA*, pp. 697–704.

Kouropteva, O., Okun, O. and Pietikäinen, M. (2002). Selection of the optimal parameter value for the locally linear embedding algorithm, *1st International Conference on Fuzzy Systems and Knowledge Discovery, FSKD(1), Singapore*, pp. 359–363.

Kulczycki, P. and Łukasik, S. (2014). An algorithm for reducing the dimension and size of a sample for data exploration procedures, *International Journal of Applied Mathematics and Computer Science* **24**(1): 133–149, DOI: 10.2478/amcs-2014-0011.

Lee, J.A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*, Springer, New York, NY.

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

Levina, E., Wagaman, A.S., Callender, A.F., Mandair, G.S. and Morris, M.D. (2007). Estimating the number of pure chemical components in a mixture by maximum likelihood, *Journal of Chemometrics* **21**(1–2): 24–34.

Li, S. Z., Xiao, R., Li, Z. and Zhang, H. (2001). Nonlinear mapping from multi-view face patterns to a Gaussian distribution in a low dimensional space, *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS), Vancouver, BC, Canada*, pp. 47–54.

Mo, D. and Huang, S.H. (2012). Fractal-based intrinsic dimension estimation and its application in dimensionality reduction, *IEEE Transactions on Knowledge and Data Engineering* **24**(1): 59–71.

Nene, S.A., Nayar, S.K. and Murase, H. (1996). Columbia object image library (COIL-20), *Technical Report CUCS-005-96*, Columbia University, New York, NY.

Niskanen, M. and Silven, O. (2003). Comparison of dimensionality reduction methods for wood surface inspection, *6th International Conference on Quality Control by Artificial Vision, QCAV(5132), Gatlinburg, TN, USA*, pp. 178–188.

Qiao, M.F.H. and Zhang, B. (2009). Intrinsic dimension estimation of manifolds by incising balls, *Pattern Recognition* **42**(5): 780–787.

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

Saul, L.K. and Roweis, S.T. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds, *Journal of Machine Learning Research* **4**: 119–155.

Shin, Y.J. and Park, C.H. (2011). Analysis of correlation based dimension reduction methods, *International Journal of Applied Mathematics and Computer Science* **21**(3): 549–558, DOI: 10.2478/v10006-011-0043-9.

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

van der Maaten, L.J.P. (2007). An introduction to dimensionality reduction using MATLAB, *Technical Report MICC 07-07*, Maastricht University, Maastricht.

Varini, C., Nattkemper, T. W., Degenhard, A. and Wismuller, A. (2004). Breast MRI data analysis by LLE, *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, Montreal, Canada*, Vol. 3, pp. 2449–2454.

Verveer, P. and Duin, R. (1995). An evaluation of intrinsic dimensionality estimators, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(1): 81–86.

Weinberger, K.Q. and Saul, L.K. (2006). Unsupervised learning of image manifolds by semidefinite programming, *International Journal of Computer Vision* **70**(1): 77–90.

Yang, M.-H. (2002). Face recognition using extended isomap, *IEEE International Conference on Image Processing, ICIP(2), Rochester, NY, USA*, pp. 117–120.

Yata, K. and Aoshima, M. (2010). Intrinsic dimensionality estimation of high-dimension, low sample size data with d-asymptotics, *Communications in Statistics—Theory and Methods* **39**(8–9): 1511–1521.

Zhang, J., Li, S.Z. and Wang, J. (2004). Nearest manifold approach for face recognition, *6th IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea*, pp. 223–228.

Zhang, Z. and Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment, *SIAM Journal of Scientific Computing* **26**(1): 313–338.

Rasa Karbauskaitė is a researcher in the System Analysis Department at the Institute of Mathematics and Informatics of Vilnius University. She received a Bachelor's degree in mathematics and informatics (2003) and a Master's degree in informatics (2005) from Vilnius Pedagogical University, as well as a Ph.D. in informatics from Vytautas Magnus University and the Institute of Mathematics and Informatics (2010). Her research interests include multidimensional data visualization, estimation of the visualization quality, dimensionality reduction, estimation of the intrinsic dimensionality of high-dimensional data, and data clustering.

Gintautas Dzemyda graduated from the Kaunas University of Technology, Lithuania, in 1980, and in 1984 he received there a doctoral degree in technical sciences after postgraduate studies at the Institute of Mathematics and Informatics, Vilnius, Lithuania. In 1997 he obtained the degree of a *doctor habilitatus* from the Kaunas University of Technology. The title of a professor was conferred upon him in 1998 at the Kaunas University of Technology. He is the director of the Vilnius University Institute of Mathematics and Informatics and the head of the System Analysis Department of that institute. His areas of research are the theory, development and application of optimization, and the interaction of optimization and data analysis. His interests include visualization of multidimensional data, optimization theory and applications, data mining in databases, multiple criteria decision support, neural networks, and parallel optimization.

Received: 11 September 2014
Revised: 27 February 2015
Re-revised: 14 April 2015