# Geodesic distances in the maximum likelihood estimator of intrinsic dimensionality

**Rasa Karbauskaitė**[a], **Gintautas Dzemyda**[a], **Edmundas Mazėtis**[b]

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

[b]Lithuanian University of Educational Sciences
Studentų str. 39, 08106 Vilnius, Lithuania
edmundas.mazetis@vpu.lt

**Received:** 3 July 2011 / **Revised:** 9 November 2011 / **Published online:** 7 December 2011

**Abstract.** While analyzing multidimensional data, we often have to reduce their dimensionality so that to preserve as much information on the analyzed data set as possible. To this end, it is reasonable to find out the intrinsic dimensionality of the data. In this paper, two techniques for the intrinsic dimensionality are analyzed and compared, i.e., the maximum likelihood estimator (MLE) and ISOMAP method. We also propose the way how to get good estimates of the intrinsic dimensionality by the MLE method.

**Keywords:** multidimensional data, intrinsic dimensionality, maximum likelihood estimator, Isomap.

## 1 Introduction

In the exploratory data analysis, we often confront with real-life data that are of a very high-dimensionality. However, these data are usually not truly high-dimensional, i.e., they are only embedded in a high-dimensional space, but can be efficiently summarized in a space of much lower dimensionality, such as a nonlinear manifold. It means that these data points locate on some manifold of lower dimensionality or they are close to that manifold. Recently, there has been a surge of interest in manifold learning methods (locally linear embedding (LLE) [1,2], ISOMAP [3], Laplacian eigenmaps (LE) [4], Hessian LLE (HLLE) [5], local tangent space analysis (LTSA) [6], and others [7]), which focus on finding a nonlinear low-dimensional projection of manifold-type high-dimensional data. The manifold learning methods require at least two parameters to be determined: the intrinsic dimensionality $d$ of the high-dimensional data and the number $k$ of the nearest neighbours. Improper values of these parameters greatly influence the results. The ways to select the value of the parameter $k$ are proposed in [8, 9]. The dimensionality of the projection is a key parameter for manifold learning methods. On one hand, a large

value chosen of the intrinsic dimensionality $d$ amplifies noise effects, while a low value leads to overlaps in mapping results (excessively reduced) [10]. It is noted in [11] that if the dimensionality $d$ is too small, important data features are "collapsed" onto the same dimensionality, and if the dimensionality is too large, the projections become noisy and, in some cases, unstable. Therefore, the problem is to disclose the exact dimensionality of that manifold, i.e., the intrinsic dimensionality *d* of the analyzed data.

At first, the term of manifold needs to be defined. 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. A line and a circle are one-dimensional manifolds. 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 plane segment.

The intrinsic dimensionality of a data set is usually defined as the minimal number of parameters or latent variables necessary to describe the data [7]. Latent variables are still often called as degrees of freedom of a data set [3, 7]. Let the dimensionality of the analyzed 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 $d \ll n$.

Principal component analysis (PCA) is the most-known dimensionality reduction method that integrates an estimator of the intrinsic dimensionality. However, the model of PCA is linear, meaning that the estimator works only for manifolds containing linear dependencies (i.e., linear subspaces). For more complex manifols, PCA gives at best an estimate of the global dimensionality of an object [7].

Due to the increased interest in dimensionality reduction and manifold learning, several approaches have been proposed in order to estimate the intrinsic dimensionality of a data set $X$ in the last decade [11–16]. Techniques for intrinsic dimensionality estimation can be subdivided into two main groups: estimators based on the analysis of local properties of the data and estimators based on the analysis of global properties of the data.

Six techniques for intrinsic dimensionality estimation are overlooked in [17]. Local intrinsic dimensionality estimators are based on the observation that the number of data points, covered by a hypersphere around a data point with radius $r$, grows proportional to $r^d$, where $d$ is the intrinsic dimensionality of the data manifold around that data 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 $r$. Three local estimators for intrinsic dimensionality – the correlation dimension estimator, the nearest neighbour dimension estimator, and the maximum likelihood estimator – are described in short in [17]. Whereas local estimators for intrinsic dimensionality compute the average over local estimates of intrinsic dimensionality, global estimators consider the data as a whole when estimating the intrinsic dimensionality. Van der Maaten (2007) overlooks these global intrinsic dimensionality estimators: the eigenvalue-based estimator, the packing number estimator, and the geodesic minimum spanning tree estimator.

In [18], the maximum likelihood estimator of the intrinsic dimensionality is applied

to the real problem, i.e., to the issue of determining the number of pure components in a mixture from Raman spectroscopy data. 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 and etc. Other possible application is given in Section 4 to find the number of degrees of freedom of motion of the object in a set of photographs.

In this paper, we also analyze the maximum likelihood estimator (MLE) and explore, which distances – Euclidean or geodesic – must be evaluated between data points in the MLE algorithm in order to get the true estimate of the intrinsic dimensionality. One of the nonlinear manifold learning methods, i.e., ISOMAP, is analysed as well, because of its ability to find out the intrinsic dimensionality of data. Disadvantages of this method in estimating the intrinsic dimensionality are disclosed.

## 2 The maximum likelihood estimator of intrinsic dimensionality

The maximum likelihood estimator [11] belongs to the class of the local estimators for intrinsic dimensionality. The detailed algorithm of MLE is provided in [11]. In this paper, only the idea is suggested.

Let the analyzed data consist of $m$ $n$-dimensional points $X_i = (x_{i1}, \dots, x_{in})$, $i = 1, \dots, m$ ($X_i \in R^n$). MLE finds the intrinsic dimensionality $d_{\mathrm{MLE}}$ of the data set $X$. According to the authors, in practice, it is more convenient to fix the number of neighbours $k$ rather than the radius $r$ of the hypersphere. Therefore, in this paper, we provide an algorithm that is related with the number of the nearest neighbours.

The MLE algorithm [11] has two control parameters: $k_1$ and $k_2$ ($k_1 < k_2$) – the numbers of the nearest neighbours for each data point. The values of these parameters are chosen. The algorithm has the following steps:

1. The number $k_2$ of the nearest neighbours for each data point $X_i$ is found.
2. The estimate of the intrinsic dimensionality ($d_{\mathrm{MLE}}$) is calculated by the maximum likelihood estimator (MLE) according to the formula:

$$d_{\mathrm{MLE}} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} d_k, \tag{1}$$

where

$$d_k = \frac{1}{m} \sum_{i=1}^{m} d_k(X_i), \tag{2}$$

$$d_k(X_i) = \left[ \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{d(X_i, X_{ik})}{d(X_i, X_{ij})} \right]^{-1}. \tag{3}$$

Here $d(X_{i,} X_{ij})$ is the Euclidean distance from the point $X_i$ to the $j$-th nearest neighbour $X_{ij}$, i.e., it represents the radius of the smallest hypersphere with the centre $X_i$ that covers $j$ neighbouring datapoints. In [11], it is shown that one could divide by $k - 2$ rather than $k - 1$ to make the estimator asymptotically unbiased.

It is clear from equation (3) that the estimate depends on the parameter $k$ as well as on the point $X_i$. Levina and Bickel (2004) assume that all the data points come from the same manifold, and therefore they average the estimated dimensions over all observations ($m$ is the number of data points) (2). According to the authors, the choice of $k$ clearly affects the estimate. In general, for MLE to work well, the hypersphere should be small and simultaneously contain enough points. Levina and Bickel 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 (1). According to experimental investigations, Levina and Bickel recommend the values of $k_1 = 10$ and $k_2 = 20$. However, these estimates are valid for some fixed data sets only.

Since it is not known how to choose the values of the parameters $k_1$ and $k_2$ in general case, in this paper, by analyzing the MLE algorithm, we use only one control parameter $k$, i.e., 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 distances: Euclidean and geodesic. In both cases, the values $d_k$ (2) of MLE are calculated with different values $k$ of the nearest neighbours. In such a way, dependences of the estimate of intrinsic dimensionality of the data on the number $k$ of the nearest neighbours are obtained. We choose such a value $d_k$ of MLE that is stable in a long interval of $k$. Levina et al. (2007) suggest to select the value of $k$ equal to 20 on the basis of a dataset with known number of pure components in a mixture from Raman spectroscopy data.

## 3 The analyzed data sets

The following data sets were used in the experiments:

1000 3-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear 2-dimensional S-shaped manifold (Fig. 1(a)).

1000 3-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear 2-dimensional 8-shaped manifold (Fig. 1(b)). The components $(x, y, z)$ of these data are calculated by the parametric equations below:

$$\begin{cases} x = \cos(v), \\ y = \sin(v)\cos(v), \\ z = u, \end{cases}$$

where $v \in [\frac{2\pi}{m}, 2\pi]$, $u \in (0, 5)$, $m$ is the number of data points.

1801 3-dimensional data points ($m = 1801,\ n = 3$) that lie on a nonlinear 2-dimensional manifold – right helicoid (Fig. 1(c)). The components $(x, y, z)$ of these data are calculated by the parametric equations below:

$$\begin{cases} x = u\cos(v), \\ y = u\sin(v), \end{cases}$$

where $u, v \in [0, 10\pi]$, $z = 0.5v$.

1801 3-dimensional data points ($m = 1801$, $n = 3$) that lie on a nonlinear 1-dimensional manifold – spiral (Fig. 1(d)). The components $(x, y, z)$ of these data are calculated by the parametric equations below:

$$\begin{cases} x = 100\cos(t), \\ y = 100\sin(t), \\ z = t, \end{cases}$$

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

1000 3-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear 1-dimensional manifold – helix (Fig. 1(e)). The components $(x, y, z)$ of these data are calculated by the parametric equations below:

$$\begin{cases} x = (2 + \cos(8t))\cos(t), \\ y = (2 + \cos(8t))\sin(t), \\ z = \sin(8t), \end{cases}$$

where $t \in [\frac{2\pi}{m}, 2\pi]$, $m$ is the number of data points.

360 2-dimensional data points ($m = 360$, $n = 2$) that lie on a nonlinear 1-dimensional manifold – circle. The components $(x, y)$ of these data are calculated by the parametric equations below:

$$\begin{cases} x = \cos(t), \\ y = \sin(t), \end{cases}$$

where $t \in [0, 2\pi]$.

181 2-dimensional data points ($m = 181,\ n = 2$) that lie on a nonlinear 1-dimensional manifold – semicircle.

*A data set of uncoloured pictures of a rotated duckling* [19] (samples of pictures are shown in Fig. 1(f)). The data are comprised of uncoloured pictures of the same object (a duckling), obtained by gradually rotated a duckling at the $360^\circ$ angle. Each picture is digitized, i.e., a data point is a vector that consists of colour parameters of pixels, and, therefore, it is of a very large dimensionality. The number of pictures (data points) is $m = 72$. The images have $128 \times 128$ greyscale pixels, therefore the dimensionality of points, characterizing each picture in a multidimensional space, is $n = 16384$.

*A data set of coloured pictures of a rotated cup* [19] (samples of pictures are shown in Fig. 1(g)). The data are comprised of coloured pictures of the same object (a cup), obtained by gradually rotated a cup at the $180^\circ$ angle. Each picture is digitized, i.e., a data point is a vector that consists of colour parameters of pixels, and, therefore, it is of a very large dimensionality. The number of pictures (data points) is $m = 35$. The images have $128 \times 128$ colour pixels, therefore the dimensionality of points, characterizing each picture in a multidimensional space, is $n = 49152$.

*A data set of photos of a person's face* [3] (example images are shown in Fig. 1(h)). The data consist of many photos of a person's face observed in different poses (left-and-right pose, up-and-down pose) and lighting conditions, in no particular order. Each picture

is digitized, i.e., a data point is a vector that consists of colour parameters of pixels, and, therefore, it is of a very large dimensionality. The number of *photos* (data points) is $m = 698$. The images have $64 \times 64$ colour pixels, therefore the dimensionality of points that characterize each photo in a multidimensional space is $n = 4096$.

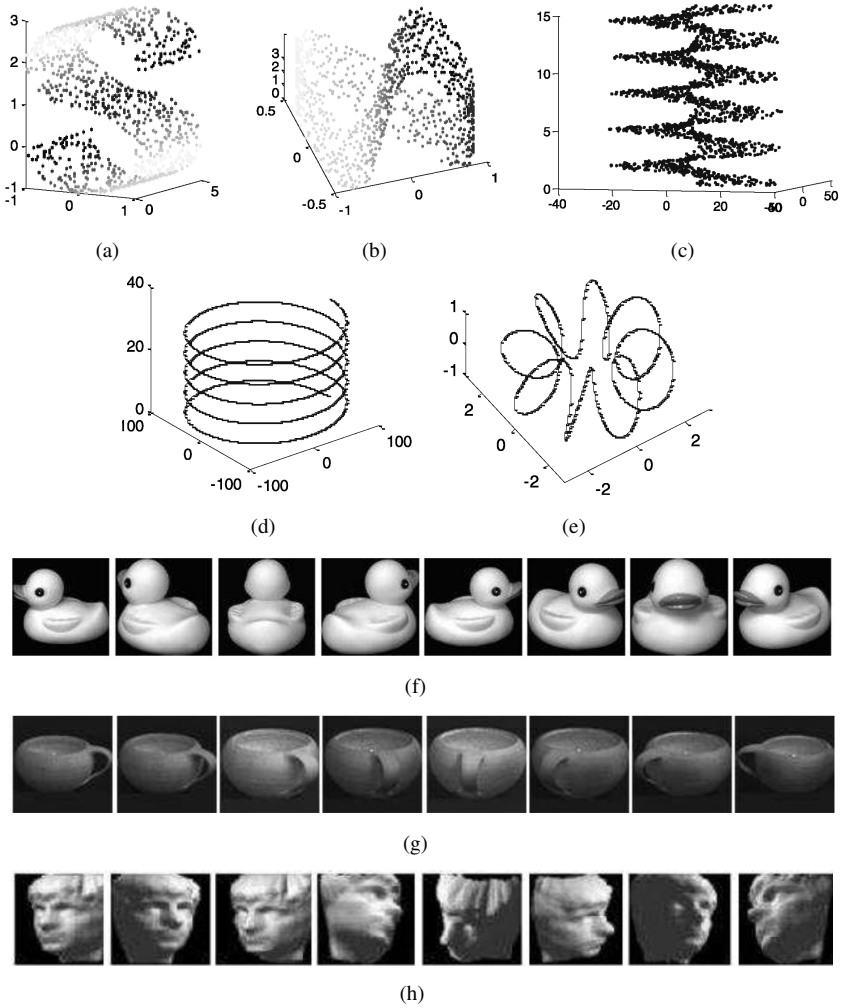

Fig. 1. Manifold datasets.

## 4 Experimental exploration of MLE

In this section, the MLE method is explored experimentally, while Euclidean or geodesic distances are evaluated between data points. For brevity, we denote the MLE method as

MLEe, if Euclidean distances are used, and MLEg, if geodesic distances are used. The estimates of the intrinsic dimensionality, obtained by MLEe and MLEg, are denoted as $d^*_{\mathrm{MLEe}}$ and $d^*_{\mathrm{MLEg}}$, respectively.

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 of the neighbours of each point $X_i$ can be organized in two ways: (1) by the fixed number $k_{\mathrm{geod}}$ 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 derived, 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.

The first investigation is performed with the points of the 2-dimensional S-shaped and 8-shaped manifolds and with the points of the 1-dimensional manifolds: a circle and a semicircle. The estimates of the intrinsic dimensionality $d$ of the data were calculated by MLE with various values of the control parameter $k$, $k \in [3,\ 100]$. After applying both variants of MLE, the true results are obtained, i.e., $d^*_{\mathrm{MLEe}} = d^*_{\mathrm{MLEg}} = 2$ for all $k$ in the case of 2-dimensional manifolds ($k_{\mathrm{geod}} = 5$, $k_{\mathrm{geod}}$ is the number of the nearest neighbours chosen when geodesic distances are calculated), and $d^*_{\mathrm{MLEe}} = d^*_{\mathrm{MLEg}} = 1$ for all $k$ in the case of 1-dimensional manifolds ($k_{\mathrm{geod}} = 2$).

However, after investigating such manifolds as a helicoid (Fig. 2), a helix (Fig. 3) and a spiral (Fig. 4), it became clear, that MLEe provides wrong results with many values of the parameter $k$. Meanwhile, in the case of MLEg ($k_{\mathrm{geod}} = 5$ in the case of the helicoid, and $k_{\mathrm{geod}} = 2$ in the case of the helix and the spiral), the true results are obtained with $k \in [5, 200]$.

An advantage of MLEg over MLEe became also evident while investigating the high-dimensional data, obtained after digitizing real pictures (uncoloured pictures of a rotated duckling, coloured pictures of a rotated cup, and photos of a person's face observed in different poses) (Figs. 5–7). The intrinsic dimensionality of these data, obtained by MLEg ($k_{\mathrm{geod}} = 2$ in the case of pictures of a rotated duckling, $k_{\mathrm{geod}} = 3$ in the case of pictures of a rotated cup, $k_{\mathrm{geod}} = 5$ in the case of photos of a person's face), is equal to the number of degrees of freedom of a possible motion of the object observed. Since a duckling or a cup were gradually rotated at a certain angle in the same plane, i.e., without turning the object itself, these data have only one degree of freedom, i.e., the intrinsic dimensionality of these data is equal to 1. A person's face analyzed in [3] 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 intrinsic dimensionality of these data is equal to 2. However, the intrinsic dimensionality of these data, obtained by MLEe, is not equal to the number of degrees of freedom of a possible motion of the observed object. Thus, MLEe fails in indentifying the true intrinsic dimensionality.

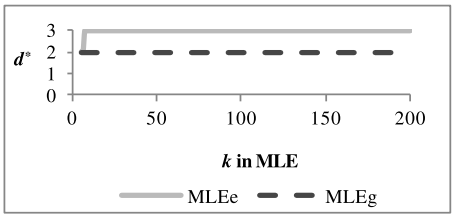


Fig. 2. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the points of the helicoid by MLE.

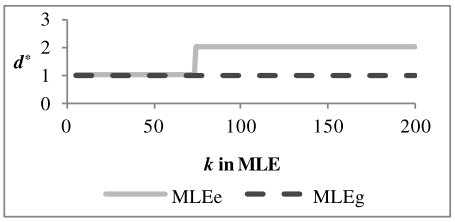


Fig. 3. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the points of the helix by MLE.

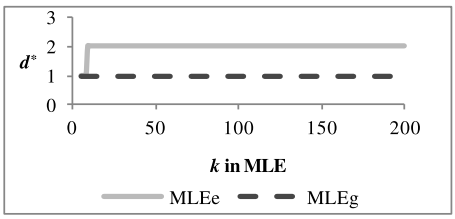


Fig. 4. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the points of the spiral by MLE.

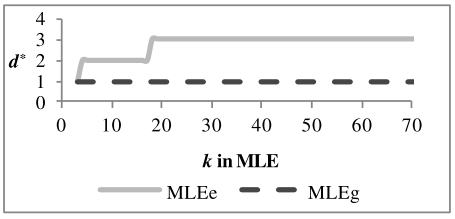


Fig. 5. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the data points, corresponding to uncoloured pictures of a rotated duckling, by MLE ($k \in [3, 71]$).

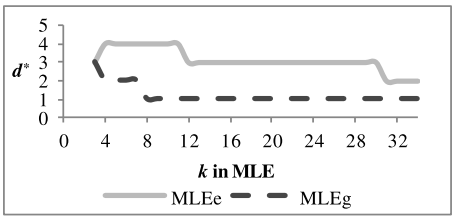


Fig. 6. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the data points, corresponding to the coloured pictures of a rotated cup, by MLE ($k \in [3, 34]$).

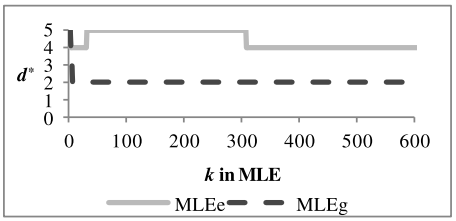


Fig. 7. Dependences of the estimate $d^*$ of the intrinsic dimensionality on $k$, obtained after analyzing the data points, corresponding to the photos of a person's face, by MLE ($k \in [3, 600]$) ($k_{\mathrm{geod}} = 5$).

## 5 Isometric feature mapping (ISOMAP)

ISOMAP can be assigned to the group of multidimensional scaling methods. This method is designed for dimensionality reduction as well as for visualization of multidimensional data [3]. Using ISOMAP, an assumption that the points of the initial space are located on a lower-dimensional manifold is made. Geodesic distances are used as a measure of proximity between the points analyzed in the ISOMAP. 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. In the further experiments, we use the fixed number $k_{\mathrm{geod}}$ of the nearest points from $X_i$.

The ISOMAP algorithm can be generalized as follows:

1. The neighbours of each point are derived in the input multidimensional space.
2. The geodesic distance between the pairs of all the points are computed; a dissimilarity matrix is formed.
3. The projection of multidimensional points to a lower-dimensional space is obtained by multidimensional scaling.

Since ISOMAP is designed to analyse manifold-type high-dimensional data, it was selected to investigate the intrinsic dimensionality of the data as one of the MDS-type methods.

In [3], it is provided the possibility to determine the dimensionality of the manifold, i.e., the intrinsic dimensionality $d$ of the initial data by using ISOMAP. ISOMAP provides error curve that can be "eyeballed" to estimate dimensionality [11]. The interval of the projection space $d^*$ is chosen for data projection, for example, $d^* = 1, \ldots, 10$, and the residual variance [3] is computed for each $d^*$.

This quantitative measure illustrates how well the distance information is preserved. It is defined as $1 - \rho^2_{D_x D_y}$, where $\rho_{D_x D_y}$ is the standard linear correlation coefficient, taken over all the entries of $D_x$ and $D_y$, where $D_y$ is the matrix of Euclidean distances between the pairs of points in the low-dimensional space, and $D_x$ is the matrix of geodesic distances between the pairs of points (the graph distance matrix) in the high-dimensional space, respectively. The lower the residual variance, the better the high-dimensional data are represented in the embedded space.

However, the ISOMAP method has some drawbacks [7]. When the manifold to be embedded is not developable, ISOMAP yields disappointing results. In this case, the guarantee of determining a global optimum does not really matter, since actually the model and its associated error function are not appropriate anymore. Another problem encountered when running ISOMAP is the practical computation of the geodesic distances. The approximations given by the graph distances may be very rough, and their quality depends on both the data (number of points, noise) and the method parameter $k_{\mathrm{geod}}$. Badly chosen value of this parameter may totally jeopardize the quality of the dimensionality reduction (data projection as well as the intrinsic dimensionality).

## 6 Experimental exploration of ISOMAP to estimate the intrinsic dimensionality

The first investigation is performed with the points of the 2-dimensional S-shaped manifold. The ISOMAP method has been run for 10 times by selecting a different dimensionality of spaces for data projection, i.e., $d^* = 1, \ldots, 10$. Each time, after obtaining data projections in a space of lower dimensionality, the residual variance was calculated. The dependence of the residual variance on the projection space $d^*$ has been obtained. The lower the value of the residual variance, the more precise projections of multidimensional data, i.e., geodesic distances between multidimensional data points and their projections, have been preserved. We can see in Fig. 8, that in the case of the S-shaped manifold, the value of the residual variance decreases a great deal, if $d^* = 2$. Although small values of the residual variance (almost zero) are obtained, if $d^* \geq 2$, but only the first value of the interval $d^* \in [2, 10]$ is taken as the intrinsic dimensionality in [3]. Thus, the most precise data projections are obtained if data are transferred to a 2-dimensional space, i.e., the intrinsic dimensionality of these data is equal to 2, which is the truth.

The analogical investigation is performed with the points of the 2-dimensional manifolds: 8-shaped manifold (Fig. 9) and helicoid (Fig. 10). In both cases, the values of the residual variance considerably decrease, if $d^* = 2$. However, the relative value of the residual variance with $d^* = 2$ and $d^* = 3$ decreases up to 65,8% in the case of the helicoid and by 47,5% in the case of the 8-shaped manifold. Thus a question arises, which

projection space and the intrinsic dimensionality of the data thereby, is more suitable? It cannot be specified strictly by ISOMAP.

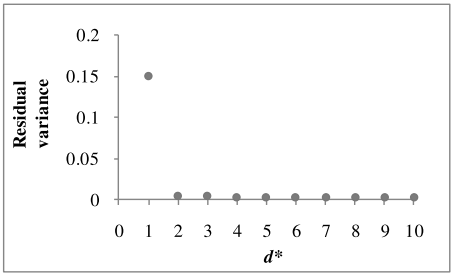


Fig. 8. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the S-shaped manifold by ISOMAP $(k_{\mathrm{geod}} = 5)$.

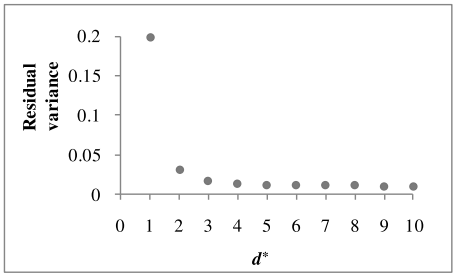


Fig. 9. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the 8-shaped manifold by ISOMAP $(k_{\mathrm{geod}} = 5)$.

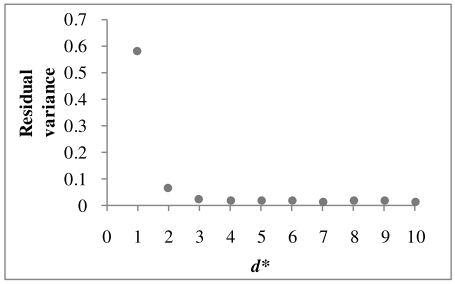


Fig. 10. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the helicoid by ISOMAP $(k_{\mathrm{geod}} = 5)$.

After investigating a spiral, it has been noticed, that the ISOMAP method ($k_{\mathrm{geod}} = 2$) does not define the intrinsic dimensionality $d$ of these data (the best projection space) at all, because the value of the residual variance is equal to 0 with all $d^* = 1, \ldots, 10$. It means that any integer number from the interval $[1, 10]$ can be the value of the parameter $d^*$. Unfortunately, it is not the truth, because the intrinsic dimensionality $d = 1$ in the case of a spiral.

Afterwards we investigated the closed 1-dimensional manifolds (a helix, a circle) that have neither the beginning nor the end. It became clear, that, in this case, the true intrinsic dimensionality $d = 1$ of these data is increased by 1 by ISOMAP, i.e., $d^* = 2$ (Fig. 11, Fig. 12). However, if a curve is unclosed, for example, a semicircle, then the intrinsic dimensionality of the data is set true by this method, i.e., $d = d^* = 1$ (Fig. 13). As the values of the residual variance in the graph are very small (equal almost zero), maybe it is risky to draw some conclusions. However, the first lowest value is obtained with $d^* = 1$.

The previous fact is validated by the following investigation with high-dimensional data points, corresponding to real pictures of a rotated duckling. If the object (a duckling) has been gradually rotated by the $360^\circ$ angle, the data points are located on a circle. However, the estimate of the intrinsic dimensionality of these data obtained by ISOMAP is equal to 2 (Fig. 14). But if a duckling is rotated at the $180^\circ$ angle little by little, then

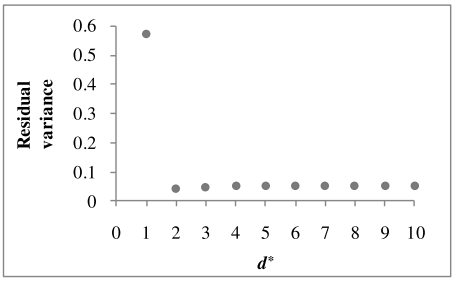


Fig. 11. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the helix by ISOMAP ($k_{\mathrm{geod}} = 2$).

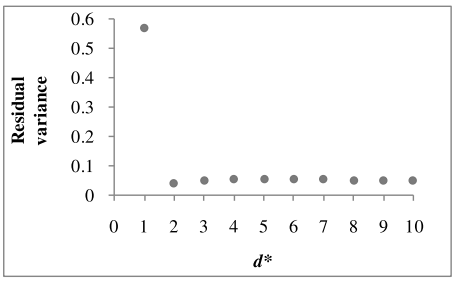


Fig. 12. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the circle by ISOMAP ($k_{\mathrm{geod}} = 2$).

the data points locate themselves on a semicircle. In this case, the intrinsic dimensionality of the data, obtained by ISOMAP, is equal to 1 (Fig. 15). In both cases, the high-dimensional data points, corresponding to real pictures of a rotated duckling, lie on a 1-dimensional manifold (a curve). Therefore, the true intrinsic dimensionality of the data is 1.

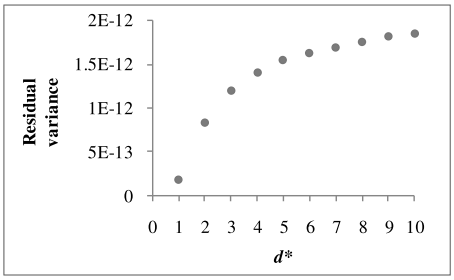


Fig. 13. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points of the semicircle by ISOMAP ($k_{\mathrm{geod}} = 2$).

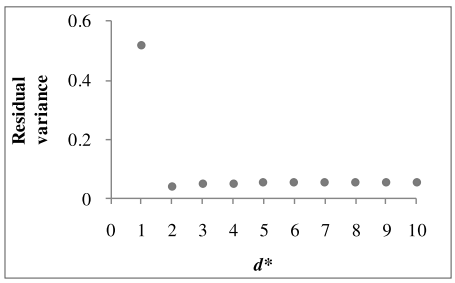


Fig. 14. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points, corresponding to the pictures of a rotated duckling at $360^\circ$, by ISOMAP ($k_{\mathrm{geod}} = 2$).

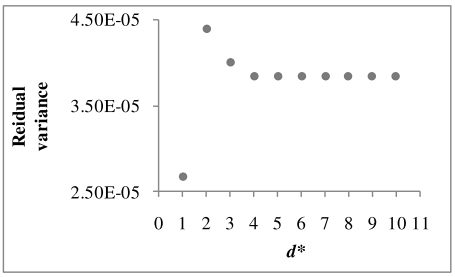


Fig. 15. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points, corresponding to the pictures of a rotated duckling at $180^\circ$, by ISOMAP ($k_{\mathrm{geod}} = 2$).

The last investigation is performed with the high-dimensional data points that correspond to coloured pictures. It has been noticed that the ISOMAP method increases the intrinsic dimensionality by 1, if colours and lighting dominate in the pictures. The intrinsic dimensionality is 1 of the high-dimensional data, obtained by digitizing the coloured pictures of a rotated cup, because these data points are located on a 1-dimensional manifold (a semicircle) as well. Besides, they have one degree of freedom of a motion. However, we can see from Fig. 16 that the dimensionality obtained by ISOMAP is equal to 2. An analogous situation is obtained by analyzing the high-dimensional data that correspond to photos of a person's face observed in different poses (left-and-right, up-and-down). Due to the lighting in the photos, the dimensionality obtained by ISOMAP is 3, but not 2 (Fig. 17).

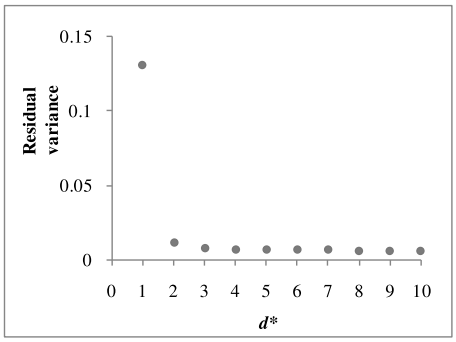


Fig. 16. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points, corresponding to coloured pictures of a rotated cup, by ISOMAP ($k_{\mathrm{geod}} = 3$).

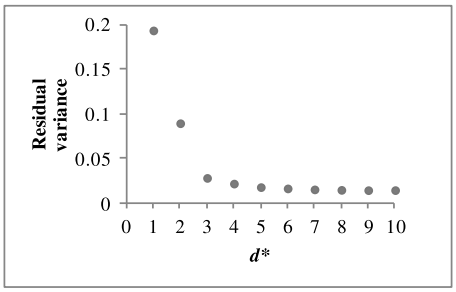


Fig. 17. Dependence of the residual variance on the dimensionality of the projection space $d^*$ obtained after analysing the points, corresponding to the photos of a person's face, by ISOMAP ($k_{\mathrm{geod}} = 5$).

Table 1. The values of the intrinsic dimensionality, obtained by different method.

| Data sets | $d$ | $d^*_{\mathrm{MLEe}}$ | $d^*_{\mathrm{MLEg}}$ | $d^*_{\mathrm{ISOMAP}}$ |
|---|---|---|---|---|
| S-shaped manifold | 2 | 2 | 2 | 2 |
| 8-shaped manifold | 2 | 2 | 2 | 2 or 3 |
| Right helicoid | 2 | 3 | 2 | 2 or 3 |
| Spiral | 1 | 2 | 1 | 1–10 |
| Helix | 1 | 1 or 2 | 1 | 2 |
| Circle | 1 | 1 | 1 | 2 |
| Semicircle | 1 | 1 | 1 | 1 |
| Uncoloured pictures of a rotated duckling at the $360^\circ$ angle | 1 | 2 or 3 | 1 | 2 |
| Coloured pictures of a rotated cup | 1 | 3 or 4 | 1 | 2 |
| Photos of a person's face | 2 | 4 or 5 | 2 | 3 |

## 7 Conclusions

Real-life data are often hardly understandable because of their high-dimensionality. Therefore, the ability to find the intrinsic dimensionality of a data set is very useful. Several methods for estimating the intrinsic dimensionality are proposed in the literature.

In this paper, we have analysed two methods for the intrinsic dimensionality: the maximum likelihood estimator (MLE) and the ISOMAP method. The obtained results are generalized in Table 1. We have shown that, in order to get true estimates by MLE, it is necessary to evaluate geodesic distances between data points in this algorithm ($d^*_{\mathrm{MLEg}} = d$ in all the cases). If the Euclidean distances are used in MLE, sometimes we can get false estimates of the intrinsic dimensionality.

ISOMAP is a nonlinear manifold learning method, but it has the ability to define the intrinsic dimensionality of data, too. Disadvantages of this method became evident in the estimation of the intrinsic dimensionality while analyzing closed manifolds and coloured pictures. In these cases, ISOMAP has failed to estimate the intrinsic dimensionality of the data, because it increased the true intrinsic dimensionality of these data by 1, as compared with the maximum likelihood estimator.

## References

1. S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, **290**, pp. 2323–2326, 2000.

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

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

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

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

6. Z. Zhang, H. Zha, Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment, *SIAM J. Sci. Comput.*, **26**(1), pp. 313–338, 2004.

7. J.A. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction*, Springer, New York, 2007.

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

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), pp. 359–364, 2007.

10. J. Yin, D. Hu, Z. Zhou, Growing locally linear embedding for manifold learning, *Journal of Pattern Recognition Research*, **2**(1), pp. 1-16, 2007.

11. E. Levina, P.J. Bickel, Maximum likelihood estimation of intrinsic dimension, *Advances in Neural Information Processing Systems*, **17**, pp. 777–784, 2005.

12. M. Brand, Charting a manifold, *Advances in Neural Information Processing Systems*, **15**, pp. 961–968, 2003.

13. F. Camastra, A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(10), pp. 1404–1407, 2002.

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

15. B. Kegl, Intrinsic dimension estimation using packing numbers, *Advances in Neural Information Processing Systems*, **15**, pp. 681–688, 2005.

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

17. L.J.P. van der Maaten, An introduction to dimensionality reduction using MATLAB, Technical Report MICC 07-07, Maastricht University, Maastricht, The Netherlands, 2007.

18. E. Levina, A.S. Wagaman, A.F. Callender, G.S. Mandair, M.D. Morris, Estimating the number of pure chemical components in a mixture by maximum likelihood, *J. Chemometr.*, **21**, pp. 24–34, 2007.

19. S.A. Nene, S.K. Nayar, H. Murase, Columbia object image library (COIL-20), Technical Report CUCS-005-96, 1996, `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`.