INVESTIGATION OF THE MAXIMUM LIKELIHOOD ESTIMATOR OF INTRINSIC DIMENSIONALITY

R. Karbauskaitė, G. Dzemyda
Vilnius University, Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: {rasa.karbauskaite, gintautas.dzemyda}@mii.vu.lt

Abstract

Real-life data are often hardly understandable because of their high-dimensionality. Therefore, the knowledge about 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, the maximum likelihood estimator (MLE) for the intrinsic dimensionality is analyzed. We propose the way how to improve the estimates of the intrinsic dimensionality. Directions for further investigations are highlighted.

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. The intrinsic dimensionality of a data set is usually defined as the minimal number of parameters or latent variables necessary to describe the data [5]. Latent variables are still often called as degrees of freedom of a data set [5]. 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 \( d \ll n \).

Due to the increased interest in dimensionality reduction and manifold learning, several approaches have been proposed in order to estimate the intrinsic dimensionality during the last decade [1, 2, 4, 6]. Techniques for intrinsic dimensionality estimation can be divided into two main groups [8]: (1) estimators based on the analysis of local properties of the data (the correlation dimension estimator, the nearest neighbor dimension estimator, and the maximum likelihood estimator (MLE)) and (2) estimators based on the analysis of global properties of the data (the eigenvalue-based estimator, the packing number estimator, and the geodesic minimum spanning tree estimator).

In this paper, we analyze the maximum likelihood estimator (MLE) [6] and compare application of both Euclidean and geodesic distances between data points in the MLE algorithm. The research is deeply related with the papers [3, 6]. Possible further investigations are highlighted.
2 The Maximum Likelihood Estimator of Intrinsic Dimensionality

Let the analyzed data set $X$ consists of $m$ $n$-dimensional points $X_i = (x_{i1}, \ldots, x_{in})$, $i = 1, \ldots, m$ ($X_i \in \mathbb{R}^n$). MLE [6] finds the intrinsic dimensionality $d_{\text{MLE}}$ of the data set $X$.

The MLE algorithm has two control parameters: $k_1$ and $k_2$ ($k_1 < k_2$) – the numbers of the nearest neighbors for each data point. The values of these parameters must be chosen. The algorithm has the following steps:

1) The intrinsic dimensionality of the data point $X_i$ is estimated by the formula:

$$d_k(X_i) = \left[ \frac{1}{k-2} \sum_{j=1}^{k-1} \log \frac{d(X_i, X_{ik})}{d(X_i, X_{ij})} \right]^{-1}. \quad (1)$$

Here $d(X_i, X_{ij})$ is the Euclidean distance from the point $X_i$ to the $j$-th nearest neighbor $X_{ij}$.

2) Estimated dimensions over all $m$ data points are averaged:

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

3) The choice of $k$ affects the estimate. Levina and Bickel [6] 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:

$$d_{\text{MLE}}(k_1, k_2) = \frac{1}{k_2-k_1+1} \sum_{k=k_1}^{k_2} d_k. \quad (3)$$

After the experimental investigations [6], the following values of $k_1$ and $k_2$ are recommended: $k_1 = 10$ and $k_2 = 20$. However, these estimates are valid for some fixed data sets only. Like in [3], here we use only one control parameter $k$, i.e. the number of the nearest neighbors for each data point. 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 neighbors. In such a way, dependences of the estimate $d_k$ of intrinsic dimensionality of the data on the number $k$ of the nearest neighbors are obtained. As the estimate of the intrinsic dimensionality, we choose such a value $d_k$ of MLE that remains stable in a long interval of $k$.

3 Experimental exploration of MLE

The following data sets were used in the experiments: a) 1000 3-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear 2-dimensional S-shaped manifold (Fig. 1a), b) 1000 3-dimensional data points ($m = 1000$, $n = 3$) that lie on a nonlinear 1-dimensional manifold – helix (Fig. 1b), c) a data set of uncolored (greyscale) pictures of a rotated duckling [3, 7] (samples of pictures are shown in (Fig. 1c)).
In this section, the MLE method is examined, while Euclidean or geodesic distances [3] 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 first investigation is performed with the points of the 2-dimensional S-shaped manifold (Fig. 2a). The estimates of the intrinsic dimensionality $d_k$ of the data were calculated by MLE with various values of the control parameter $k$, $k \in [3, 200]$. After applying both variants of MLE, the true results are obtained for all $k$ in the case of 2-dimensional S-shaped manifold. However, after investigating the helix (Fig. 2b), it became clear, that MLEe provides wrong results with many values of the parameter $k$. Meanwhile, in the case of MLEg, 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, i.e. uncolored pictures of a rotated duckling (Fig. 2c). The intrinsic dimensionality of these data, obtained by MLEg, is equal to the number of degrees of freedom of a possible motion of the object observed. Since a duckling was 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.
4 Conclusions

The experimental investigations showed that in order to get true estimates by maximum
likelihood estimator, it is better to evaluate geodesic distances between data points in
this algorithm. If the Euclidean distances are used in MLE, sometimes we can get
false estimates of the intrinsic dimensionality. The MLE algorithm [6] may have one of
the control parameters: $r$, i.e. the radius of the hypersphere that contains the nearest
data points, or the parameter $k$, i.e. the number of nearest neighbors. In [6], it is
stated that a more convenient way to estimate the intrinsic dimensionality of the data
is to fix the number $k$ of neighbors instead of the radius $r$ of the hypersphere. As far
as we know, everyone, who has been investigating the MLE until now, including [3],
used the parameter $k$. For our further investigations we are going to fix the radius $r$
of the hypersphere instead of the number $k$ of neighbors. The aim will be not to draw
dependences of the estimate of intrinsic dimensionality of the data on parameter $r$ in
order to know the value of $r$, but to get the value of this parameter automatically.

References

a fractal-based approach. IEEE Transactions on Pattern Analysis and Machine

Vol. 52(8), pp. 2210-2221.

imum Likelihood Estimator of Intrinsic Dimensionality. Nonlinear Analysis: Mod-


York.


MATLAB. Technical Report MICC 07-07, Maastricht University, Maastricht, The
Netherlands.