An Empirical Study on the Performance of Spectral Manifold Learning Techniques

Peter Mysling, Søren Hauberg, and Kim Steenstrup Pedersen

The eScience Center,
Dept. of Computer Science, University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen Ø, Denmark
{mysling,hauberg,kimstp}@diku.dk

Abstract. In recent years, there has been a surge of interest in spectral manifold learning techniques. Despite the interest, only little work has focused on the empirical behavior of these techniques. We construct synthetic data of variable complexity and observe the performance of the techniques as they are subjected to increasingly difficult problems. We evaluate performance in terms of both a classification and a regression task. Our study includes Isomap, LLE, Laplacian eigenmaps, and diffusion maps. Among others, our results indicate that the techniques are highly dependent on data density, sensitive to scaling, and greatly influenced by intrinsic dimensionality.

Key words: Manifold Learning, Evaluation, Synthetic Data.

1 Introduction

In recent years, the development of techniques for nonlinear dimensionality reduction has generated much interest. Spectral manifold learning, in which the data is assumed to lie near an embedded manifold, has emerged as a particularly prominent approach. These techniques compute a low-dimensional representation based on the structure of the manifold, while also guaranteeing a globally optimal solution. During the last decade, a vast number of manifold learning techniques were proposed [1–7].

Surprisingly, only little work has focused on the empirical behavior and performance of these techniques. To our knowledge, only three such studies exist, namely (1) the work of Yeh et al. [8], in which LLE, Kernel PCA, and Isomap are compared in terms of a clustering task; (2) the work of Niskanen & Silven [9] in which five techniques are evaluated on several low-density data sets; and (3) the technical report of van der Maaten et al. [10] in which twelve techniques are compared on a range of both artificial and natural data sets. In the case of the two latter studies, performance is only evaluated in terms of neighborhood preservation. All previous studies only consider problems of fixed difficulty.

Our study deviates from the previous work in two critical ways. First of all, the techniques are evaluated in terms of both a local and a global measure of structure preservation. Secondly, and more importantly, we construct data sets

in which the complexity can be controlled by a single parameter, allowing us to study the performance as a function of the problem difficulty. By systematically applying this scheme to several types of complexity, we are able to identify scenarios under which the techniques break down. Moreover, we are able to highlight strengths and weaknesses, not only of each technique individually, but also of the methods in general. Furthermore, we believe that, by visualizing the performance as a function of the data complexity, we give an intuitive understanding of characteristic behavior not found in previous studies.

We have designed 5 data set variants, each of which can be scaled in complexity, in terms of a certain data property. The suite contains data sets in which (1) the density can be varied, (2) the amount of noise can be varied, (3) the embedded manifold contains a hole of variable size, (4) the scaling can be varied, (5) the intrinsic dimensionality can be varied. All data sets are modifications of the classical swiss roll [2], which has traditionally been applied in qualitative evaluation of manifold learning techniques. Our data sets are synthetic, because natural data sets would have an unknown or at least poorly estimated manifold structure, which would render our study impossible. We have confined our analysis to four canonical manifold learning techniques, namely Isomap [2], LLE [1], Laplacian eigenmaps [3], and diffusion maps [7]. In order to evaluate the discovered embeddings, we construct quality measures based on two common supervised learning tasks—classification and regression. The quality measure based on regression is sensitive to global deformations in data structure and, to our knowledge, this measure is novel in the analysis of manifold learning techniques.

2 Techniques

In the following, we provide a brief review of the applied manifold learning techniques. Due to space constraints, we refer to the original papers for details.

The manifold learning problem is stated as follows. Let $\{\mathbf{x}_i \in \mathbb{R}^D : i \in 1, \ldots, n\}$ be a collection of data points lying near a possibly nonlinear d-dimensional manifold. The aim is to determine a low-dimensional representation in the form of a mapping $\mathbf{x}_i \in \mathbb{R}^D \mapsto \mathbf{y}_i \in \mathbb{R}^d$ which preserves the structure of the embedded manifold. We let $\mathbf{X}$ and $\mathbf{Y}$ denote corresponding design matrices.

The evaluated techniques represent each data point as a node in a similarity graph G. The graph is constructed in one of three ways: i) by an ϵ-neighborhood approach, in which each point is connected to all points within a ball of radius ϵ; ii) by connecting each point to its k nearest neighbors; iii) by similarity weighting, in which G is a fully connected, weighted graph and weights are assigned according to a Gaussian function of width σ^2.

All techniques compute a low-dimensional representation which retains some measure of the data structure, based on the similarity graph. The optimization amounts to an eigendecomposition of a matrix which is quadratic in the number of data examples.

Isomap [2] estimates the pair-wise geodesic distances by the shortest paths distances in G. The low-dimensional representation is chosen such that the

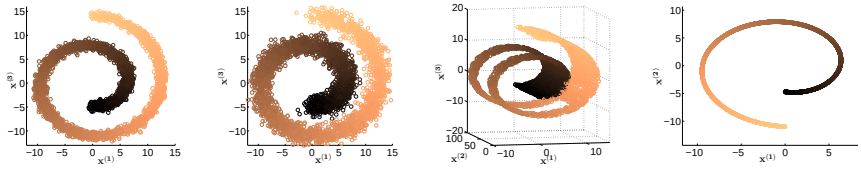
geodesic distances are retained. LLE [1] characterizes the local data structure using linear models and uncovers an embedding which can be described by the same model. Laplacian eigenmaps [3] compute a low-dimensional embedding in which neighboring nodes are proximate, under the weighting of a Gaussian kernel of width σ^2. Diffusion maps [7] apply similarity weighting and treats the distances between data points as transition probabilities in a Markov chain. The similarity between data points is estimated by simulating a Markov random walk between the nodes for t time steps.

3 Synthetic Data Sets

In this study, we construct data sets which vary in complexity as a function of a single argument, which we will refer to as the *data* argument. We evaluate the selected techniques by applying them to data sets of increasing complexity. All constructed data sets are modifications of the traditional swiss roll [2]. The swiss roll data set is a natural basis for several reasons: 1) it is visualizable; 2) it has a simple shape which cannot be modeled by PCA; 3) the chosen techniques are known to perform well on this data set.

A synthetic data set $\mathbf{X}$ is constructed by a mapping $f : \mathbb{R}^d \to \mathbb{R}^D$ of n data points $\{\hat{\mathbf{y}}_i \in \mathbb{R}^d : i = 1, \ldots, n\}$, where $\hat{\mathbf{y}}_i = [\hat{y}_{i,1}, \ldots, \hat{y}_{i,d}]^{\mathrm{T}}$. $\hat{y}_{i,j}$ is sampled from a uniform distribution with finite support $[c_j^{min}, c_j^{max}]$. We refer to these points as the *true* embedding. Letting $\hat{y}_{i,1} \in \left[\frac{3\pi}{2}, \frac{9\pi}{2}\right]$ and $\hat{y}_{i,2} \in [0, 100]$, each data point of the embedded swiss roll is calculated by

$$\mathbf{x}_i = f(\hat{\mathbf{y}}_i) = \left[\hat{y}_{i,1}\cos(\hat{y}_{i,1}),\ \ \hat{y}_{i,2},\ \ \hat{y}_{i,1}\sin(\hat{y}_{i,1})\right]^{\mathrm{T}}.$$



(a) $\sigma^2 = 0.5$ (b) $\sigma^2 = 1$ (c) $h = 0.5$ (d) $d = 1$

Fig. 1. Visualizations of a selection of the applied data sets. (a) 2-dimensional visualization of the noise data set for $\sigma^2 = 0.5$. (b) 2-dimensional visualization of the noise data set for $\sigma^2 = 1$. (c) Visualization of the hole data set for $h = 0.5$ (50%). (d) Visualization of the intrinsic dimensionality data set for $d = 1$.

Below we motivate and describe the five data set variants. We provide visualizations when the structure of the data set is nontrivial.

Density: Machine learning data sets are often of low density and it is unknown how severely this affects the discovered embeddings. We construct data sets which vary in density by varying n, the number of data points.

Noise: Natural data often exhibit irregular structure and contains noisy measurements. We model this by adding Gaussian noise sampled from $\mathcal{N}(0, \sigma^2)$ to each component of the swiss roll data points $\mathbf{x}_i$. Realizations of this data are visualized in Fig. 1(a) and 1(b).

Hole: Some concern has been expressed regarding the inability of Isomap to model nonconvex manifolds [5]. Motivated by this, we apply the techniques to data where the manifold has a hole. A manifold which contains a square hole, centered in the true embedding and spanning h percent of each true embedding axis, is constructed by rejecting all samples within the hole. Fig. 1(c) shows a realization of this data.

Scaling: Natural data is often a product of a number of measurements; these measurements are frequently not directly comparable and must be rescaled appropriately for analysis. We investigate the sensitivity of the techniques with respect to scaling. Rescalings of the swiss roll data set are constructed by rotating the manifold 45 degrees around each coordinate axis and scaling the first component of the resulting data points by a factor of s.

Intrinsic dimensionality: We investigate the performance of the techniques when subjected to data of variable intrinsic dimensionality. A data set containing one intrinsic dimension is defined to be the 2-dimensional swiss roll. Each additional intrinsic dimension is simply added by including a linear component sampled from $U(0, 100)$. Note that, under this simple scheme, the empirical performance of the techniques degenerate to that of PCA when $d = 3$ and higher. Because of this, we simplify the swiss roll by only sampling $\hat{y}_{i,1}$ from $U\left(\frac{3\pi}{2}, \frac{7\pi}{2}\right)$. A visualization of this is given in Fig. 1(d).

4 Quality Measures

Motivated by the applicability of spectral manifold learning techniques to data analysis, we evaluate the embeddings discovered by these techniques in terms of two common supervised learning tasks—classification and regression. Under this scheme, we associate to each data point $\mathbf{y}_i$ a target value t_i based on its position in the true embedding. In the classification setting, where $t_i^{clas} \in \{0, 1\}$, target values are assigned in a checkerboard pattern. In the regression setting we have $t_i^{reg} \in \mathbb{R}$ and target values are assigned linearly along the first coordinate axis in the true embedding, i.e. $t_i^{reg} = \hat{y}_{i,1}$. Visualizations are given in Fig. 2.

In principle, any classification technique can be applied in the classification setting. In this study, we employ a Nearest Neighbour (NN) classifier for simplicity. Letting p_i^{clas} denote the NN prediction of t_i^{clas} under leave-one-out cross-validation, we define the quality Q_{clas} of $\mathbf{Y}$ as the misclassification rate [10, 9]. The classification measure determines how well the local structure is preserved in the discovered embeddings.

Since the target values were chosen as a linear component in the true embedding, it is reasonable to expect that they can approximately be reconstructed linearly in the embedded coordinate system. Thus, we define the quality Q_{reg} of

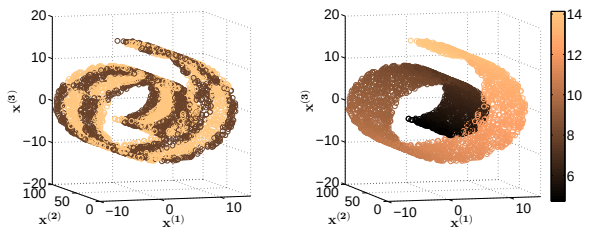


Fig. 2. Target value assignment in the classification and regression settings. Color denotes target value. Left: Classification setting. Right: Regression setting.

the embedding $\mathbf{Y}$ wrt. the data $\mathbf{X}$ as the root mean squared error

$$Q_{reg} = \sqrt{n^{-1} \sum_{i=1}^{n} (p_i^{reg} - t_i^{reg})^2},$$

where p_i^{reg} is the predicted target value under a linear least squares regression model using leave-one-out cross-validation. The regression measure responds to deformations in both global and local data structure. To our knowledge, this measure is novel.

5 Experimental Results

Each technique has a number of parameters which must be fixed. We estimate the optimal parameters in a practical manner, by exhaustively searching a fixed range of viable parameters, and retaining the parameters which maximize quality measures. For Isomap, LLE, and Laplacian eigenmaps, k is varied in $k \in \{4, \ldots, 20\}$. For diffusion maps, the parameters are varied in $t \in \{1, 2, 3, 5, 10, 15, 25\}$ and $\sigma^2 \in \{0.75, 1, 2, 3, 5\}$. Note that we avoid fixing the σ^2 parameter of Laplacian eigenmaps by letting $\sigma^2 \to \infty$, as proposed by Belkin & Niyogi [3].

Having fixed the parameters for each data set, we estimate the mean performance over a series of 10 trials; each trial uses a new realization of the data set. The results are plotted along with the standard error. We report the performance of PCA as a baseline measure. Except for the density experiment, each data set is constructed with a density of 3500 data points. We remind the reader that, for both quality measures, a lower score is indicative of better performance.

The experimental results are given in Fig. 3–7. For clarity, the markers have been slighty displaced. Before inspecting each experiment in turn, we make two general observations. First, we note that the two quality measures are highly correlated; when the measures disagree, it is an indication that a global deformation of the embedded manifold has occurred. Secondly, we observe that LLE tends to perform less stable than the remaining techniques, especially in the regression setting. We do, however, not believe that this is an effect of attempting

to uncover global manifold structure from models of local geometry; our results show that Laplacian eigenmaps is capable of this with considerable stability. Rather, we speculate that this is a weakness of modeling the local geometry by reconstruction weights.

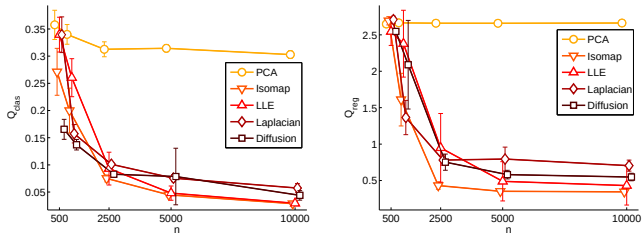


Fig. 3. Results of the density experiment. Left: Classification measure. Right: Regression measure.

Density (Fig. 3): We make two key observations. First, the performance of the techniques does not converge until $n \geq 2500$; note that this is a fairly densely sampled manifold. Additionally, we observe that diffusion maps, in the low-density cases, outperform the remaining techniques with significant stability, according to the classification measure. Since this is not the case in the regression setting, we conclude that only the local manifold structure is preserved.

Noise (Fig. 4): We note that the performance of the techniques deteriorates as the noise is increased beyond $\sigma^2 = 0.5$. Surprisingly, the sensitivity of Isomap with respect to short-circuiting does not result in more rapid deterioration than the remaining techniques. Diffusion maps and Laplacian eigenmaps tend to be especially robust when subjected to low noise data. We also observe that diffusion maps are capable of preserving the local structure, even noise levels increase.

Hole (Fig. 5): We observe that holes on the manifold, regardless of the size, does not significantly affect the performance of the applied techniques. Note that this is not necessarily an indication that the applied techniques accurately determine the structure of the true embedding, but rather that the discovered embeddings are satisfactory in terms of the classification and regression tasks.

Scaling (Fig. 6): We observe that, generally, the techniques more easily reconstruct an embedding which is satisfactory in terms of the classification measure than the regression measure. Again, this gives an indication that the local structure is more easily retained than the global structure. Additionally, we observe that the techniques struggle to recover the global manifold structure when the data is scaled beyond $s \in [0.5; 2]$.This effect is most pronounced for LLE, Laplacian eigenmaps, and diffusion maps.

Intrinsic Dimensionality (Fig. 7): We observe that the performance of the techniques begins to deteriorate when $d > 3$ and that the techniques do not have a significant advantage over PCA when $d > 4$.

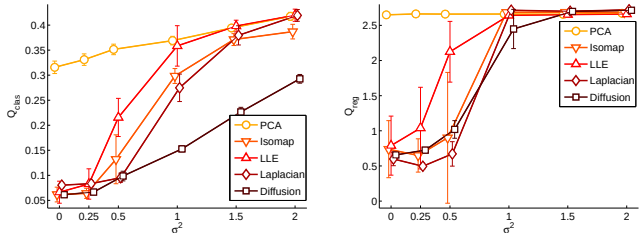


Fig. 4. Results of the noise experiment. Left: Classification measure. Right: Regression measure.

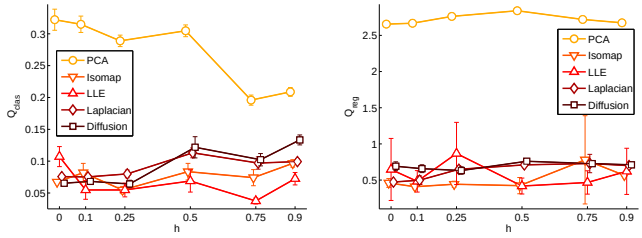


Fig. 5. Results of the hole experiment. Left: Classification measure. Right: Regression measure.

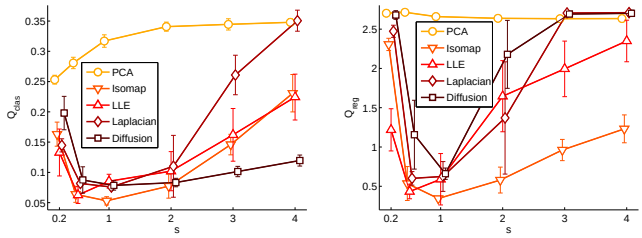


Fig. 6. Results of the scaling experiment. Left: Classification measure. Right: Regression measure.

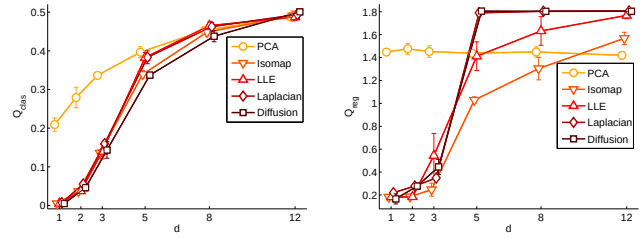


Fig. 7. Results of the intrinsic dimensionality experiment. Left: Classification measure. Right: Regression measure.

6 Discussion

In summary, our experiments indicate that the evaluated techniques are 1) highly dependent on data density, 2) invariant to holes on the manifold with respect to the classification and regression tasks, 3) sensitive to scaling, and 4) highly dependent on intrinsic dimensionality. Clearly, 1) and 4) are tightly related.

Although it is expected that high intrinsic dimensionality and low data density have a negative impact on the discovered embeddings, the severity of these effects is nevertheless surprising. As limited amounts of data is the rule rather than the exception, we consider this a severe problem. Note that these techniques require quadratic memory in the amount of data examples, making problems of more than 10.000 examples virtually infeasible on modern computers.

The experiments showed that the methods were sensitive to scaling of the original data. This is a problem of practical concern as it questions the use of e.g. whitening as a pre-processing step. Such pre-processing does not take the manifold structure into account, which is why we see performance drops when data is scaled.

We believe that our study provides four important contributions to the community: 1) we have presented a novel quality measure which is sensitive to global deformations in data structure; 2) we exemplify how to view performance as a function of complexity; 3) we facilitate an intuitive understanding of manifold learning performance; 4) our study can help practitioners evaluate whether spectral manifold learning is applicable for a certain data set.

References

1. Roweis, S., Saul, L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 290, 2323–2326 (2000)
2. Tenenbaum, J., de Silva, V., Langford, J.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science. 290, 2319–2323 (2000)
3. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15, 1373–1396 (2003)
4. Brand, M.: Charting a Manifold. In: NIPS 15, pp. 977–984. IEEE Press (2003)
5. Donoho, D.L., Grimes, C.: Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data. In: PNAS 100, pp. 5591–5596. National Academy Sciences, Washington (2003)
6. Zhang, Z., Zha, H.: Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. SIAM J. Sci. Comput 26, 313–338 (2004)
7. Coifman, R.R., Lafon, S.: Diffusion Maps. Applied and Computational Harmonics Analysis 21, 5–30 (2006)
8. Yeh, M.C., Lee, I.H., Wu, G., Wu, Y., Chang, E.Y.: Manifold Learning, a Promised Land or Work in Progress. In: Proc. of IEEE Intl. Conf. on Multimedia and Expo, pp. 1154–1157. IEEE Press, New York (2005)
9. Niskanen, M., Silven, O.: Comparison of Dimensionality Reduction Methods for Wood Surface Inspection. In: Proc. of the 6th Intl. Conference on Quality Control by Artificial Vision, pp. 179–188 (2003)
10. van der Maaten, L., Postma, E.O., van den Herik, H.J.: Dimensionality Reduction: A Comparative Review. Technical report, Tilburg Uni. (2009)