UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

DEPT OF COMPUTER SCIENCE

http://viper.unige.ch

VIPER Technical Report 08.01

December 12, 2008

Unsupervised Dimension Reduction of High-Dimensional Data for Cluster Preservation

Enikö Szekely, Eric Bruno
and Stéphane Marchand-Maillet

Contact: Eniko.Szekely@unige.ch

http://viper.unige.ch/

VIPER - Multimedia Information Retrieval
CVMLab, Dept of Computer Science
University of Geneva – Route de Drize, 7
CH - 1227 Carouge SWITZERLAND

Abstract

High-dimensional data is receiving increasing attention in more and more application fields, but the analysis of such data has shown to be difficult due to the "curse of dimensionality". Dimension reduction methods have emerged as successful tools to overcome the problem of high-dimensionality. However, even if they are designed to preserve the most important properties of the data, they are generally blind to the preservation of structures (e.g. multimodal distributions, clusters). In this paper, we propose a class of dimension reduction strategies, called High-Dimensional Multimodal Embedding (HDME), that aim to find low-dimensio-nal representations of high-dimensional data that preserve cluster information. The difficulty of analysing high-dimen-sional data arises from the fact that, in high-dimensional representation spaces, all pairwise distances between points tend to become equal. To overcome the problem of equidistancy, HDME performs a processing of the distances, consisting of a scaling of the distances between similar data points. Similarity may be estimated based on neighbourhood, cluster or class information. We show that the neigh-bourhood-based variant is a competitive alternative to clustering. After the scaling, the points are embedded in a low-dimensional space using a distance-based embedding method. Experiments show that HDME is effective both in terms of retrieval and clustering when compared to known state-of-the-art methods operating in high-dimensional spaces. The code and data are available from `http://viper.unige.ch/doku.php/viper_private:HDME`.

1 Introduction

The need for effective tools of analysis of high-dimensional data gave rise to the field of *dimension reduction*. Dimension reduction methods reduce the dimensionality of the original dataset so as to simplify the data while trying to keep most of its important properties. As a general definition, dimension reduction is a process meant to find meaningful low-dimensional representations of high-dimensional data.

In many datasets, data naturally follows a multimodal distribution and can therefore be organised into clusters. For example, taking the field of information retrieval and given a query, a document relevance to the query can be associated to the document-cluster membership (a document is more relevant to a query if it belongs to the same cluster). In such a context, when reducing the dimensionality of the data, cluster preservation becomes critical for efficient retrieval. However, the preservation of clusters, despite its importance in numerous fields, has still received only little attention. Traditional dimension reduction methods reduce the dimensionality but are generally blind to the preservation of structures. Moreover, clustering in high dimensions has revealed itself to be a difficult problem and methods for high-dimensional data clustering have started to be proposed. However, joint methods that combine dimension reduction with clustering with the purpose of finding low-dimensional spaces that preserve cluster information are still very rare.

In this paper, we propose the High-Dimensional Multimodal Embedding (HDME), a class of dimension reduction methods that aim to preserve clusters. In high dimensions, distances between data points tend to be equal [6]. This behaviour is due to the "curse of dimensionality" and can hide global information, such as clusters. To overcome the equidistancy problem, HDME proposes a processing of the original distances, followed by the low-dimensional embedding. Experiments are performed on real data - the classical MNIST handwritten digits database and the 20 Newsgroups text database. Performance is evaluated in terms of mean average precision and k-means accuracy and show good results of HDME over existing methods.

The rest of the paper is organised as follows. The rest of Section 1 succintly discusses the observations on high-dimensional data, as presented in the literature. Section 2 reviews

related work on dimension reduction and clustering of high-dimensional data. Section 3 formally introduces HDME, along with different approaches for the estimation of similarity between points in Section 4. Experiments and results on real data appear in Section 5. The paper ends with discussions and conclusions in Section 6.

1.1 On the nature of high-dimensional data

The term "high-dimensional data" is employed to designate data points represented as feature vectors of dimensionality D, with D varying from 10s to 1000s or even more dimensions. The analysis of such datasets, whether it is for retrieval, clustering, exploration or visualisation, revealed to be difficult due to the "curse of dimensionality" [5].

The effect of sparsity in high dimensions on pairwise distances was investigated in [3, **?**, **?**, **?**]. In high-dimensional spaces, points tend to be equidistant for a wide range of data distributions [6]. Given a query point $\mathbf{x}_i$ and δ_i^{min} and δ_i^{max} the distances from the $\mathbf{x}_i$ to its closest and farthest neighbours, observations state that, under certain conditions, the relative contrast between the closest and farthest neighbour decreases with the increase in dimensionality, converging to zero:

$$\lim_{D\to\infty} \frac{\delta_i^{max} - \delta_i^{min}}{\delta_i^{min}} \to 0 \tag{1}$$

Equation (1) can be rewritten in terms of the ratio between the distances to the closest and farthest neighbour from the query point:

$$\lim_{D\to\infty} \frac{\delta_i^{max}}{\delta_i^{min}} \to 1 \tag{2}$$

In [12] the authors make an important observation, stating that, despite the equidistancy in ratio (Equation (2)), distances do not necessarily tend to be equal in absolute contrast $\delta_i^{max} - \delta_i^{min}$. The value of the absolute contrast does not necessarily go to zero. It is shown that $\delta_i^{max} - \delta_i^{min}$ behaves differently for different L_n norms, as it grows with $D^{\left(\frac{1}{n}-\frac{1}{2}\right)}$. It results that for the Manhattan metric L_1, the absolute difference increases with the dimensionality and for the Euclidean metric L_2, it converges to a constant, while for the other Minkowski metrics L_n with $n \geq 3$, it decreases with the dimensionality. Thus, L_n metrics, $n \geq 3$, do not seem adapted for high dimensional data, while L_1 and L_2 norms can reveal important properties of the data. The above observations show that it is important to consider the behaviour of distances not only in terms of ratio, but of absolute difference too. As discussed later in this paper, we use the absolute contrast locally as it can capture important properties and globally, we perform a processing of distances in order to avoid the equidistancy in ratio. The processing is necessary for proper preservation of certain global pro-perties, like clusters.

2 Related work

Dimension reduction is mostly based on the assumption that the data lies in subspaces of lower dimensionality than the original space. It is hoped that there exists a meaningful intrinsic dimensionality (d) of the data that is smaller than the original dimensionality (D), $d \ll D$. Traditional dimension reduction methods, reviewed next, optimise a formalised objective function, but are generally blind to the presence of structures (e.g. clusters) within the data. PCA is the most employed dimension reduction method in practice. It

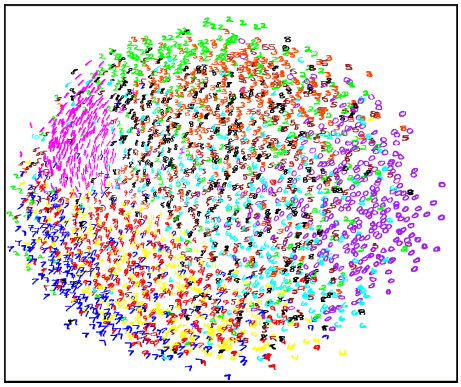


Figure 1: Embedding with Sammon Mapping of a 784-dimensional clustered dataset.

is a linear transform whose objective is to capture as much as possible from the variance in the data, but it is not designed to cope with non-Gaussian distributions and even less with clustered data. Multidimensional Scaling (MDS) [7] is a global approach to pairwise distance preservation, optimising a stress function over all distances. A widely used variant of MDS is Sammon Mapping [21] that increases the importance given to smaller distance values by using a self-normalisation procedure. As in high dimensions distances tend to be equal, global methods that use distances directly (e.g. MDS, Sammon Mapping) fail in preserving any agglomeration of data (e.g. clusters). The resulting embeddings are generally spherical (Figure 1).

The *manifold assumption* has been the key for some of the main algorithms developed recently (e.g. Isomap [23], Locally Linear Embedding (LLE) [20, 22]). The high-dimensional data is assumed to lie on a *manifold* of lower dimensionality than the original space (e.g. Swiss Roll, a 2D-manifold embedded in a 3D-space). A manifold is a space in which local neighbourhoods ressemble a Euclidean space, but the global structure is generally more complex. The local Euclideanity of manifolds only justifies the use of the Euclidean distance for local neighbourhoods. Larger distance values may be estimated using "geodesic" distances computed *along* the manifold and not *through* the manifold. Isomap and Curvilinear Distance Analysis [14] are mappings based on geodesic distances. Unfortunately, whereas the geodesic distance captures well the shape of the manifold, it is not designed to discriminate between clusters.

A global method (e.g. MDS, Isomap) tries to recover the global structure of the data. On the contrary, a local method (e.g. LLE, Laplacian Eigenmaps) aims to recover the local structure of the data and hopes to recover the global structure from local fits. LLE preserves the linear reconstruction of a point from its neighbours. Laplacian Eigenmaps (LE) [4] embed points in the low-dimensional space with respect to the eigenvectors of the Laplacian matrix. SSammon [16] modifies Sammon Mapping so that only the first neighbours of a point contribute to the stress function. DD-HDS [15] uses a sigmoid function, similar to Curvilinear Component Analysis [9], to also allow only the shorter distances to contribute to the error.

Among all the above-cited dimension reduction methods, Laplacian Eigenmaps is the only method tailored for clustered data, since it uses Laplacian graphs, as in spectral clustering, for the embedding.

Clustering high-dimensional data is a difficult task. In high dimensions, clusters are complex structures and rarely convex (e.g. Gaussian). The need for clustering such data gave rise to new directions in cluster analysis. An interesting direction emerged in

the last decade: subspace clustering. Subspace clustering methods (CLIQUE, MAFIA, PROCLUS etc.) search for the specific subspace in which each cluster lives, since in high dimensions clusters rarely live in the same subspace. A complete review on subspace clustering algorithms can be found in [17].

Gaussian Mixture Models (GMMs) in high dimensions have a practical shortcoming: the number of parameters to estimate is too high. In this case, parsimonious models (e.g. spherical instead of full covariances) may be used, that solve the problem, but are more restrictive in shape. High-Dimensional Data Clustering [8] combines subspace clustering and parsimonious models to avoid the problems of high dimensionality and to find the specific subspace of each cluster. Adaptive Dimension Reduction (ADR) [10] iteratively performs clustering in reduced spaces and uses the clustering information to recompute cluster centers in the original space. Once the cluster centers are known, it finds the subspace spanned by the centers. The process is repeated until convergence. The main advantage of ADR is that it avoids performing clustering in the high-dimensional space. LDA-Km [11] is a variant of ADR that uses k-means clustering and Linear Discriminant Analysis (LDA) to select the discriminant subspace. LDA-Km requires the computation of the between- and within-scatter matrices. When the within-scatter matrix computed in high-dimensions can not be inversed, the LDA-Km-B variant [11] can be used as it requires only the between-scatter matrix. A different approach is proposed in Hierarchical Multidimensional Scaling (H-MDS) [19] that combines clustering with dimension reduction. The main motivation is that it is natural to treat between- and within-cluster distances differently. Thus, H-MDS first performs the clustering and then performs an algorithm for positioning all clusters in the same subspace.

Clustering and dimension reduction are powerful tools for the analysis of high-dimensional data. Combining these tools in a unified framework is a challenging task and re-presents the purpose of HDME. As shown later in the paper this does not only simplify the data but also allows for improved retrieval and efficient clustering in low-dimensional spaces.

3 High-Dimensional Multimodal Embedding

3.1 Motivation

HDME relies on the following three observations:

- results from [12] state the meaningfullness of the Euclidean distance even in high dimensions;
- in high dimensions, data often lies on manifolds. As discussed in Section 2, the local Euclideanity of manifolds justifies the use of the Euclidean distance for local neighbourhoods;
- the ratio of distances goes to one [6].

Locally, the first two observations justify the use of the Euclidean distance. Globally, cluster preservation involves first of all cluster discrimination. In terms of distances, discrimination may be expressed through a *scale* difference among distances, roughly distinguishable as within- and between-cluster distances. However, in accordance with the third observation, the equality of all distances in ratio hinders cluster separation as the embeddings get spherical (Figure 1). The natural scale difference observed for low-dimensional

clustered data, disappears in high dimensions. In consequence, we artificially apply, before the actual embedding, a processing of the distances. The processing consists of a non-linear scaling of distances, meant to avoid the "curse of dimensionality". Distances between similar points are scaled down, while leaving the rest intact. That way, we exploit and reinforce the capability of classical dimension reduction techniques (Sammon Mapping, in our case) to preserve clusters and expect to maintain their inner structure.

3.2 Model

Formally, consider a clustered set $\mathcal{X}$ of N data points $\mathbf{x}_i$ in the original high-dimensional space $\mathbb{R}^D$. The problem is to find an accurate projection of the data in a low-dimensional space $\mathbb{R}^d$ ($d \ll D$) that preserves clusters as best as possible. The matrix of distances between the N points is given by $\Delta = (\delta_{ij}) \in \mathbb{R}^{N \times N}$, where δ_{ij} is the Euclidean distance in the original space between point $\mathbf{x}_i$ and point $\mathbf{x}_j$.

Scaling Let $\mathcal{P} = \mathcal{X} \times \mathcal{X}$ be the set of all pairs of points $(\mathbf{x}_i, \mathbf{x}_j)$ for $\mathbf{x}_i \neq \mathbf{x}_j \in \mathcal{X}$.

Given a notion of *similarity* (defined later on), we define $\mathcal{P}_1 \subseteq \mathcal{P}$ as the subset of pairs of similar points and $\mathcal{P}_2 \subseteq \mathcal{P}$ the subset of the rest of the pairs:

$$\mathcal{P}_1 = \{(\mathbf{x}_i, \mathbf{x}_j) |\ \mathbf{x}_i, \mathbf{x}_j \text{ are similar}\} \text{ and } \mathcal{P}_2 = \mathcal{P} \setminus \mathcal{P}_1 \tag{3}$$

Given λ, a scaling factor such that $\lambda \geq 1$, we define $\Delta^* = \left(\delta^*_{ij}\right) \in \mathbb{R}^{N \times N}$ as the scaled distance matrix, where the scaled distances δ^*_{ij} are given by:

$$\delta^*_{ij} = \begin{cases} \dfrac{\delta_{ij}}{\lambda}, & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_1; \\ \delta_{ij}, & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_2. \end{cases} \tag{4}$$

Embedding The second step, after distance scaling, performs the embedding of the points into a lower-dimensional space using a distance-preserving method. HDME seeks to preserve both small and large distance values as faithfully as possible. Hence, the presence of a weighting factor in the embedding function to escape from privileging large distance values is desirable. This is typically the case in the Sammon Mapping stress function:

$$E_{SM} = \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}} \tag{5}$$

where $\mathbf{y}_i,\ i = 1..N$ are the projected points in the embedded space ($\mathbf{y}_i \in \mathbb{R}^d$) and d_{ij} are the distances in the low-dimensional space between projected points $\mathbf{y}_i$ and $\mathbf{y}_j$. Replacing the original distances δ_{ij} from Sammon Mapping with the scaled distances δ^*_{ij}, we obtain the final form of HDME:

$$E_H = \frac{1}{\sum_{i<j} \delta^*_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta^*_{ij})^2}{\delta^*_{ij}} \tag{6}$$

The presence of the non-linear weighting by $\dfrac{1}{\delta^*_{ij}}$ naturally increases the importance of smaller distance values in the embedding, thus reinforcing the scaling performed on distances.

The formal justification of HDME comes by looking at the stress function E_H. Let

$$C = \frac{1}{\sum_{i<j} \delta^*_{ij}}.$$

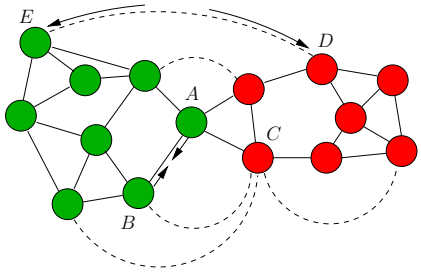


Figure 2: The effect of scaling in HDME on a two-class problem (green and red points). All pairs of similar points are connected with full lines and the rest with dashed arcs.

By substituting the scaled distances in the stress function (Equation (6)), we obtain two terms:

$$\begin{aligned} E_H &= C \sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{P}_1 \\ i<j}} \frac{\left(d_{ij} - \frac{\delta_{ij}}{\lambda}\right)^2}{\frac{\delta_{ij}}{\lambda}} + C \sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{P}_2 \\ i<j}} \frac{(d_{ij}-\delta_{ij})^2}{\delta_{ij}} \\ &= \lambda\, C \sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{P}_1 \\ i<j}} \frac{\left(d_{ij} - \frac{\delta_{ij}}{\lambda}\right)^2}{\delta_{ij}} + C \sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{P}_2 \\ i<j}} \frac{(d_{ij}-\delta_{ij})^2}{\delta_{ij}} \\ &= E_1 + E_2 \end{aligned} \tag{7}$$

The first term E_1 addresses relationships between similar points, $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_1$, and E_2 the remainder, $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_2$. Figure 2 illustrates the effect of scaling (for clarity, the figure indicates only a few pairs of points that are not considered similar). Similar points (e.g. (A, B) or (A, C)) are attracted in the embedding, due to the presence of the scaling factor λ in term E_1, that reduces the scale of the distances. Points considered as not similar (e.g. (B, C) or (D, E)) are pushed far apart in the embedding, due to the preservation in E_2 of the original large distance values.

We see that the similarity relationship may be or not transitive:

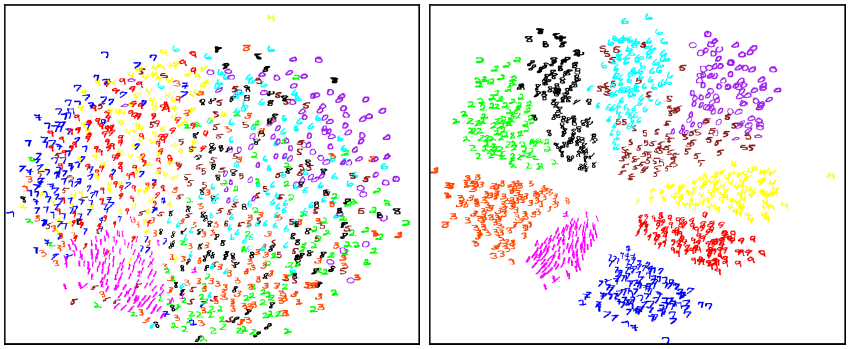
$$\text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_1 \wedge\ (\mathbf{x}_j, \mathbf{x}_k) \in \mathcal{P}_1 \Rightarrow\ (\mathbf{x}_i, \mathbf{x}_k) \in \mathcal{P}_1 \tag{8}$$

For example, points A and B are similar, the same as A and C, but B and C are not. We will discuss later the effect of the transitivity property on the embedding.

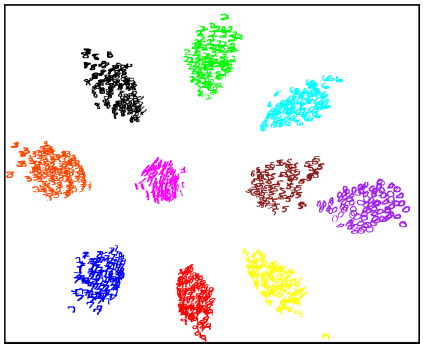
4 Similarity estimation

HDME consists of enforcing proximity relationships during the embedding in a low-dimensional space. The notion of proximity is mapped onto that of s̈imilarityör of b̈elonging to the same cluster.̈ The challenge now is to formalise and estimate these notions.

To first demonstrate the capability of HDME to reach the cluster preservation objective, we first illustrate it in a fully supervised case where class labels over the data are known and are used to define the similarity measure. We then look at unsupervised strategies to estimate similarity, first based on clustering, and then relax it, based on neighbourhoods. For illustrations we use the MNIST handwritten digit database (presented in section 5).



(A) $\lambda = 1$ (B) $\lambda = 2$



(C) $\lambda = 5$

Figure 3: Influence of the scaling factor λ on the embedding with fully supervised HDME on 1000 MNIST digits (0,1,2,3,4,5,6,7,8,9). Higher the value of the scaling factor, higher/clearer the separation between classes.

4.1 Class-based similarity

In a fully supervised scenario, two points are similar if they belong to the same class, i.e. if they have the same label. Formally, using labelled data, the subset $\mathcal{P}_1$, corresponding to similar points, becomes:

$$\mathcal{P}_1 = \{(\mathbf{x}_i, \mathbf{x}_j) | \text{ label}(\mathbf{x}_i) = \text{label}(\mathbf{x}_j)\} \tag{9}$$

Therefore, the scaling is applied on distances between points that belong to the same class. An example is given in Figure 3. The embedding with HDME is performed in a two-dimensional space. In the baseline case, when $\lambda = 1$, Figure 3(a), $\Delta = \Delta^*$ as no scaling is performed and the embedding is equivalent to the original Sammon Mapping. No class is clearly discriminated in the 2D-space. For $\lambda = 2$, Figure 3(b), class discrimination improves, but is still not sufficient to separate well all classes, whereas for $\lambda = 5$, Figure 3(c), the discrimination is clear. The clear separation between classes is due to the transitivity property of the class similarity since the distances that are scaled are all the within-class distances.

The results illustrate well the importance of the scaling for cluster preservation of high-dimensional data. The idea of supervised embedding was previously presented in [13]. A supervised PCA is proposed that helps discriminate among classes, however the linearity

of the method merges some of them. Supervised embeddings are particularly useful for the exploration of a collection of labelled data.

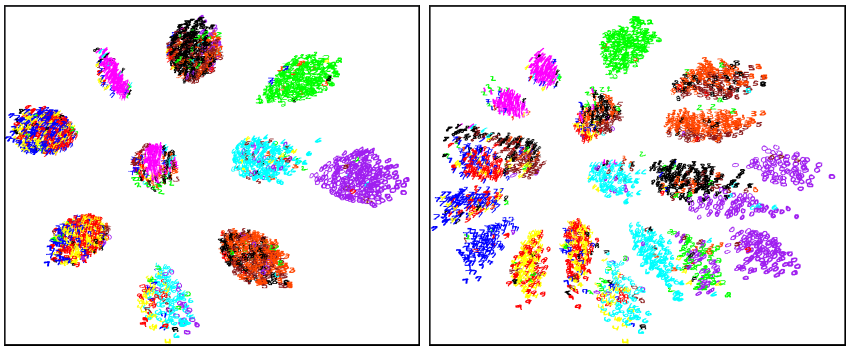
In the following we consider the challenge of providing efficient prior similarity information without supervision.

4.2 Cluster-based similarity

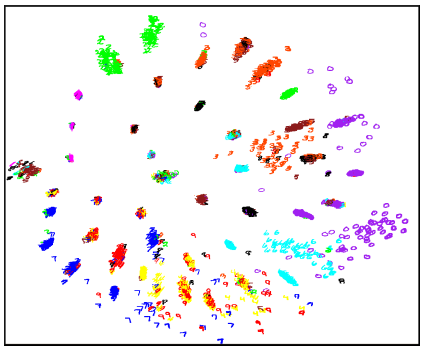
In the absence of class label information, the similarity between points may be estimated based on cluster membership. Let c be the estimated number of clusters and $\mathcal{C} = \{\mathcal{C}_1 \cdots \mathcal{C}_c\}$ be the set of clusters. Any method of clustering may be applied to estimate the clusters (Section 2). Two points are then declared as similar if they belong to the same cluster. The subset $\mathcal{P}_1$ now becomes:

$$\mathcal{P}_1 = \{(\mathbf{x}_i, \mathbf{x}_j) | \ \exists l, l = 1..c \text{ s.t. } \mathbf{x}_i \in \mathcal{C}_l \text{ and } \mathbf{x}_j \in \mathcal{C}_l\} \tag{10}$$

The scaling is applied on distances between points that belong to the same cluster. As an example of clustering, we apply the k-means algorithm for different values of the number of clusters c and different values of the scaling factor λ (Figure 4). Due to the transi-



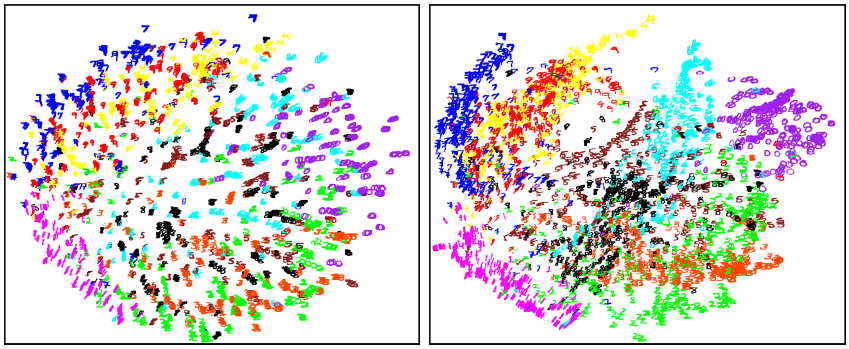
(A) $c = 10$, $\lambda = 5$ (B) $c = 20$, $\lambda = 5$



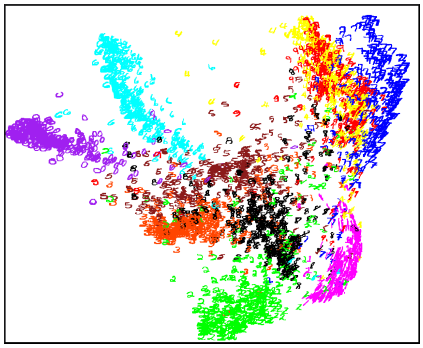
(C) $c = 50$, $\lambda = 100$

Figure 4: Unsupervised HDME on 3000 MNIST digits (0,1,2,3,4,5,6,7,8,9) with k-means for different values of the number of clusters c and of the scaling factor λ.

tivity property of the cluster relationship and given a sufficient scaling, clusters are well separated. However, the embedding shows that a hard clustering decision, like k-means,



(A) $k = 1$ (B) $k = 2$



(C) $k = 10$

Figure 5: Influence of the number of nearest neighbours k on the embedding with unsupervised HDME of 3000 MNIST digits (0,1,2,3,4,5,6,7,8,9) ($\lambda = 5000$).

does not reflect the real global structure of the data. Points from different classes are merged and points from the same class are taken apart. Thus, in the cluster-based case, the quality of the embedding will mainly depend on the quality of the clustering.

4.3 Neighbourhood-based similarity

In the absence of good clustering results, the estimation of similarity can rely on local neighbourhoods. The motivation for using local information for cluster preservation is the assumption that a point, together with its neighbours, belong to the same class or cluster. Futhermore, it is generally assumed that, in high dimensions, the data lies on a manifold. We approximate the sought manifold by a k-nearest neighbour graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_k)$ built using the Euclidean distance. Then, two points are declared similar if they are neighbours, i.e. they are connected in $\mathcal{G}$. The subset $\mathcal{P}_1$ becomes:

$$\mathcal{P}_1 = \{(\mathbf{x}_i, \mathbf{x}_j) | \; \exists \; e_{ij} \in \mathcal{E}_k\} \tag{11}$$

Different values for the number of nearest neighbours k generate different embeddings (Figure 5). Low values of k create many small components, while high values connect components together such that, when $k = N$, the graph $\mathcal{G}$ is totally connected and the embedding is equivalent to the original space (Figure 1).

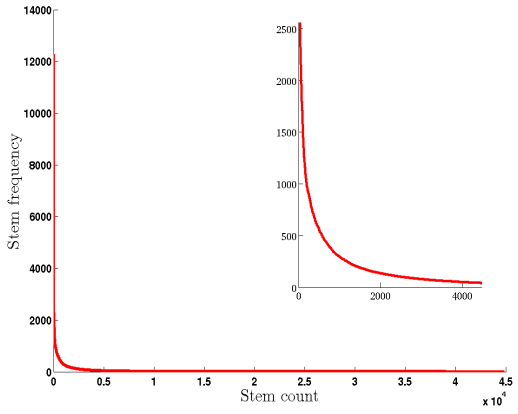


Figure 6: The stem frequency for the 11,269 Newsgroups documents.

The connectivity of the graph allows the propagation of similarity information, therefore of the scaling operation. Moreover, as the neighbourhood relationship is not transitive in $\mathcal{G}$, the embedding is constrained by the degree of connectivity of the graph (Figure 5). Thus, neighbourhood-based embeddings reflect more faithfully - than class- and cluster-based cases - the global structure of real data, where often clusters overlap. We have observed that the neighbour-hood-based HDME requires higher values of the scaling factor, than the cluster- and class-based HDME, on the same data, to obtain a good embedding.

5 Experimental results

5.1 Datasets

MNIST The MNIST handwritten digits dataset [2] is widely used in dimension reduction. All digits are normalized to fit in a $20 * 20$ pixel box and centered in a $28 * 28$ image by computing the center of the mass of the points. The digits are represented in the original space as 784-dimensional vectors (each pixel represents a dimension). The distance matrix is computed using the Euclidean distance in the 784-dimensional space.

20 Newsgroups The 20 Newsgroups dataset [1] contains approximately 20,000 articles from 20 different newsgroups divided into training (60%) and testing (40%). For the experiments, we use the data from the training set of 11,269 documents. The dataset is processed as follows: 1) stopwords are eliminated and stemming [18] is performed; 2) stems that appear too few times ($\leq$ 100) or too many times ($\geq$ 2000) are deleted as they are not relevant in a cluster preservation context. The stem frequency is plotted in Figure 6. The dimensionality of the data after processing is of 2436 dimensions. In this new feature space we eliminate duplicate documents and documents with Euclidean norm equal to zero. The final collection contains 11,213 documents from 20 categories in the 2436-dimensional space.

5.2 Evaluation

Mean average precision. Average precision (AP) is an evaluation method widely used in information retrieval, defined as follows:

$$AP = \frac{\sum_{i=1}^{R}(P(i) \times rel(i))}{number\ of\ relevant\ elements} \tag{12}$$

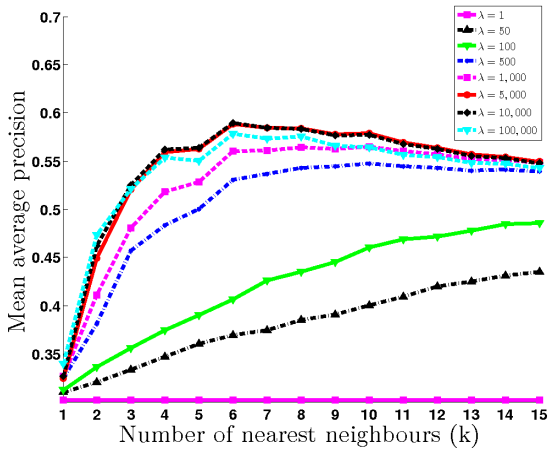


Figure 7: Influence of the parameters λ and k on the quality of the retrieval for 3000 MNIST digits (0,1,2,3,4,5,6,7,8,9).

where R - number of retrieved elements, $P(i)$ - the precision of a given cut-off rank i and $rel(i)$ - a binary function on the relevance of element i. Mean average precision (MAP) is the mean of average precision over all documents and emphasises retrieving relevant documents earlier. A document is relevant to a query if it has the same label. As HDME aims to preserve clusters, it can relate relevancy to the label information. Thus, MAP becomes a good indicator of the quality of the embedding in the low-dimensional space.

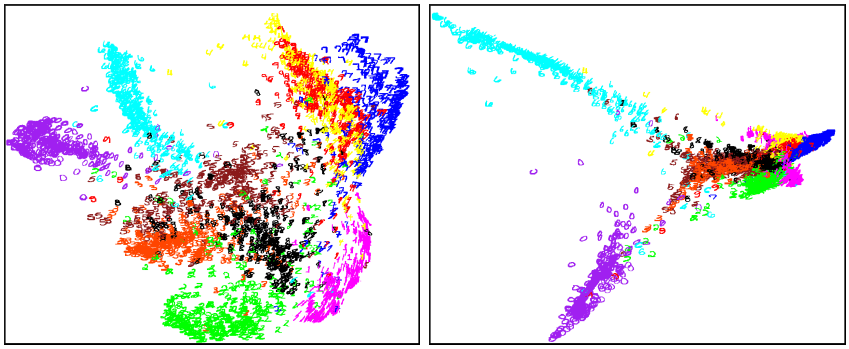
k-means clustering. The second evaluation measure consists in estimating the accuracy of the k-means clustering in the low-dimensional space. We choose k-means for its wide utilisation, however we mention here that cluster shapes may not always be well approximated by the spherical Gaussianity of k-means. The accuracy is estimated over three runs of k-means indicating the mean value. We provide always the correct number of clusters, 10 for the MNIST and respectively, 20 for the Newsgroups.

Visualisation. Two-dimensional spaces allow the inspection of data collections and provide a good understanding of the data.

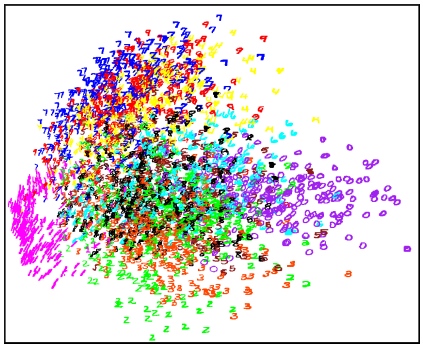
5.3 Experiments

The rest of the experiments are performed for the neigh-bourhood-based HDME, due to its high proven performance and space limitations. The first experiment (Figure 7) shows the evolution of the two parameters of the model, the scaling factor λ and the number of nearest neighbours k, on the quality of the retrieval. The retrieval improves for λ varying from 1 (Sammon Mapping) to $10,000$, whereas for a value of $100,000$ it shows decreased performance. The choice of λ becomes thus important, but not critical as it allows for a good retrieval for a wide range of values $(500 - 100,000)$. Cross-validation may help in choosing the appropriate values. Concerning the choice of k, we already discussed it intuitively in Section 4.3. MAP values confirm the idea that low values of k should be avoided as they generate many small components and do not reflect the real data global structure.

The next experiment consists in the visualisation of the MNIST digits. Figure 8 depicts the results obtained with HDME (a), LE (b) and PCA (c) for 3000 digits projected into a 2D-space. The scaling factor for the rest of the experiments involving the digits is $\lambda = 5000$. We chose to visualise the best embedding in terms of mean average precision



(A) HDME ($k = 6$) (B) LE ($k = 3$)



(C) PCA

Figure 8: Embedding for 3000 MNIST digits (0,1,2,3,4,5,6,7,8,9) with HDME, LE and PCA.

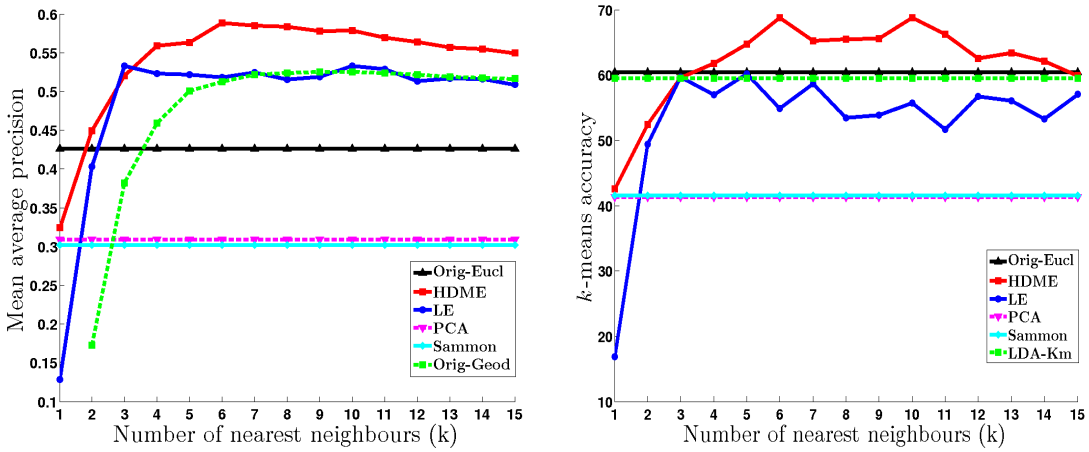
for each method. Thus, the HDME nearest neighbour graph is built with $k = 6$ and the LE nearest neighbour graph with $k = 3$. For Laplacian Eigenmaps the heat kernel parameter is chosen to be $t = \infty$ (that was shown to work well in practice in the original article [4]). HDME and LE show improved visualisation in a 2D space when compared with PCA. Moreover, we observe that the arrangement of digit classes generated by HDME distinguishes nearly all classes of digits giving them roughly equal importance in the embedding, as opposed to LE that increases the importance given to zeros and sixs (two classes with high intra-class variability).

Next (Figures 9, 10), we compute the mean average precision and k-means accuracy for 3000 MNIST digits (0,1,2,3,4, 5,6,7,8,9) and 11,213 documents from Newsgroups. MAP is estimated for the original space with the euclidean and the geodesic distance, for HDME, LE, PCA and Sammon Mapping. k-means accuracy is estimated in the original space, for HDME, LE, PCA, Sammon Mapping and LDA-Km. PCA is chosen as it is the most widely employed dimension reduction method in practice. Sammon Mapping is the baseline for HDME and LE gives the best results for clustered data, to our knowledge. The geodesic distance is chosen as it captures better than the euclidean distance the complex structure of high-dimensional data. And LDA-Km is chosen as it is a clustering method especially designed for high-dimensional data. Given c is the number of clusters, LDA-Km projects the data in a space of dimensionality $d = c - 1$ ($d = 9$ for MNIST and $d = 19$ for

Newsgroups). Due to the difference in size of the two datasets, we vary $k = 1..15$ (MNIST) and $k = 1..150$ for multiples of 10 (20 Newsgroups).

As a first observation, local-based methods (HDME, LE and geodesic) generally perform better. Concerning MAP, the 2D-space of HDME gives better performances than both the original space and LE for both datasets. The geodesic performs fairly the same for the Newsgroups data.

In terms of clustering accuracy, for the MNIST, HDME performs better than the original space, but not for the Newsgroups data. One possible explanation is the fact that the 2D space might be too low to separate all the 20 different newsgroups. For LDA-Km, the clustering accuracy is estimated in the $d = c - 1$ dimensions of the most discriminant space.



(A) MAP

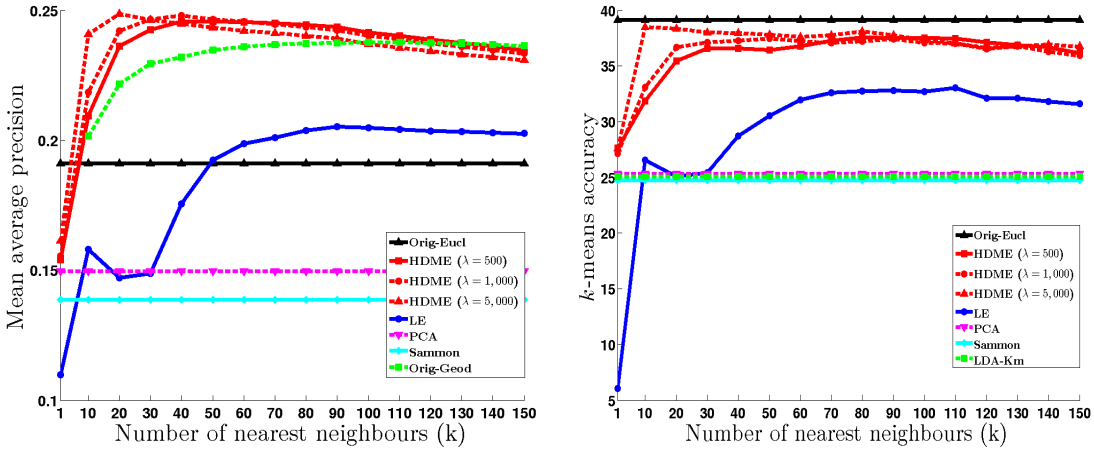
(B) k-means

Figure 9: Mean average precision and k-means accuracy for MNIST for different values of k ($\lambda = 5,000$).

The last experiment gives two examples of the Newsgroups data: 1) first 200 documents from each of the 4 categories: N8: rec.autos, N9: rec.motorcycles, N10: rec.sport.baseball, N11: rec.sport.hockey and 2) first 200 documents from each of the 4 categories: N1: alt.atheism, N5: comp.sys.mac.hard-ware, N10: rec.sport.baseball, N15: sci.space. We apply the same processing as for the whole collection and eliminate duplicates. Mean average precision, k-means accuracy and visualisation results are displayed in Figures 11 and 12.

6 Conclusions

Information mining involves high-dimensional representation spaces. In these spaces, dimension reduction techniques were proposed to avoid the curse of dimensionality. However, due to difficult behaviours of distance measurements in high dimensions, structures in the data (e.g. clusters) are difficult to be preserved. In this paper, we proposed the High-Dimensional Multimodal Embedding (HDME) as a mapping from high- to low-dimensional representation space that is able to preserve structures in the data organization. This contrasts to many classical dimension reduction strategies that are blind to structures in data.



(A) MAP (B) k-means

Figure 10: Mean average precision and k-means accuracy for 20 Newsgroups for different values of k.

We have demonstrated the relevance of HDME in the supervised case where class information is available. We have then further evaluated it in an unsupervised context over two classical datasets, naturally embedded in high-dimensional spaces and compared its performance with classical strategies such as PCA and Laplacian Eigenmaps. We have shown that HDME helps further retrieval operations by mapping the data in a space where proximity relationships are better measurable.

HDME, as a class of dimension reduction techniques tailored for structured dataset is valuable as a preprocessing for mining, retrieval, and visualisation. HDME shows good performance in quality, but it has a high complexity in time. Parameter estimation is important but not too sensible, as good results were obtained for a wide range of values for both datasets. We plan to further test the ability of the method to detect outliers and to perform visualisation in spaces of dimensionality higher than two, hoping to allow an even better embedding than the 2D space.

7 Acknowledgements

This work has been partly funded by SNF fund No. 200020-121842 in parallel with the Swiss NCCR(IM)2.

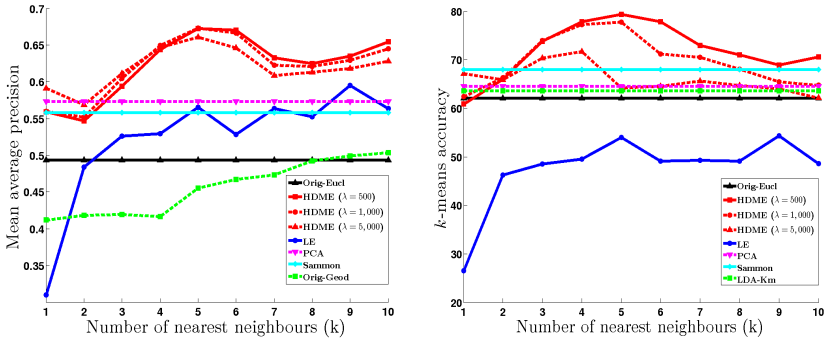
References

[1] 20 newsgroups data. `http://people.csail.mit.edu/jrennie/20Newsgroups/`.

[2] The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`.

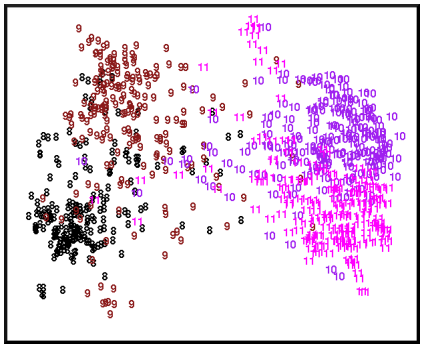
[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the International Conference on Database Theory*, pages 420–434, 2001.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.



(A) MAP

(B) k-means



(C) Visualisation ($\lambda = 1000$, $k = 5$)

Figure 11: Results for the newgroups N8, N9, N10, N11.

[5] R. Bellman. Adaptive control processes: A guided tour. Princeton University Press, 1961.

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, U., and Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, volume 1540, pages 217–235. Springer, 1999.

[7] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. 2005.

[8] C. Bouveyron, S. Girard, and C. Schmid. High dimensional data clustering. In *17th International Conference on Computational Statistics*, pages 812–820, June 2006.

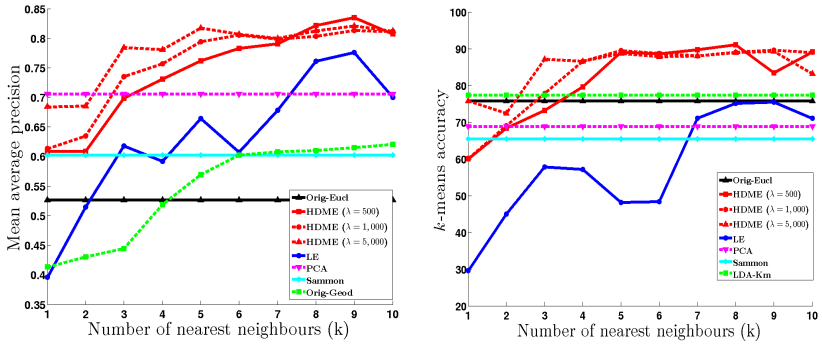
[9] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizig neural network for nonlinear mapping of data sets. In *IEEE Transactions on Neural Network*, volume 8. 1997.

[10] C. Ding, X. He, H. Zha, and H. Simon. Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the 2nd IEEE International Conference on Data Mining*, pages 147–154, 2002.

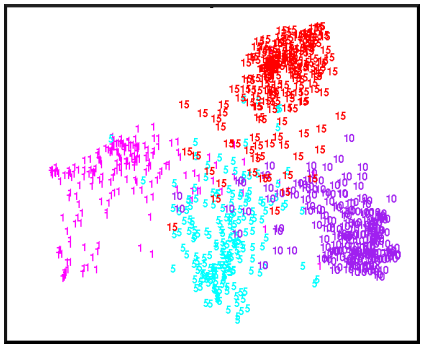
[11] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th International Conference on Machine Learning*, pages 521–528. ACM, 2007.

[12] A. Hinneburg, C. Aggarwal, and D. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th VLDB Conference*, 2000.

[13] Y. Koren and L. Carmel. Robust linear dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 2004.



(A) MAP (B) k-means



(C) Visualisation ($\lambda = 500$, $k = 9$)

Figure 12: Results for the newgroups N1, N5, N10, N15.

[14] J. A. Lee, A. Lendasse, and M. Verleysen. A robust nonlinear projection method. In *Proceddings of ESANN'2000, Belgium*, pages 13–20, 2000.

[15] S. Lespinats, M. Verleysen, A. Giron, and B. Fertil. Dd-hds: a method for visualization and explorationof high-dimensional data. *IEEE Transactions on Neural Networks*, 18, 2007.

[16] M. Martin-Merino and A. Blanco. A local semi-supervised sammon algorithm for textual data visualization. *Journal of Intelligent Systems*, 2008.

[17] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 6(1):90–105, 2004.

[18] M. Porter. The porter stemming algorithm. `http://tartarus.org/~martin/PorterStemmer/`.

[19] M. Quist and G.Yona. Distributional scaling: An algorithm for structure-preserving embedding of metric and nonmetric spaces. *Journal of Machine Learning Research*, 5:399–420, 2004.

[20] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[21] J. W. Sammon. A nonlinear mapping for data structure analysis. In *IEEE Transactions on Computers*, volume C-18. 1969.

[22] L. K. Saul, S. T. Roweis, and Y. Singer. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[23] J. B. Tennenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.