Automatic Configuration of Spectral Dimensionality Reduction Methods

Michał Lewandowski, Dimitrios Makris and Jean-Christophe Nebel
Digital Imaging Research Centre, Kingston University, KT1 2EE, UK

Abstract

We propose an advanced framework for the automatic configuration of spectral dimensionality reduction methods. This is achieved by introducing, first, the mutual information measure to assess the quality of discovered embedded spaces. Secondly, unsupervised Radial Basis Function network is designated for mapping between spaces where the learning process is derived from graph theory and based on Markov cluster algorithm. Experiments on synthetic and real datasets demonstrate the effectiveness of the proposed methodology.

Keywords: Dimensionality reduction, Locally Linear Embedding, Isomap, Laplacian Eigenmaps, Mutual Information, Radial Basis Function network, Markov Cluster algorithm

1. Introduction

With the exponential increase of data production driven by applications such as the internet, mobile communication, computer vision, medical imaging, speech recognition and genomics, powerful tools are required by scientists to allow the analysis of these data. Since they are usually highly dimensional,

dimensionality reduction has become an essential process in the exploration of large volumes of multivariate data.

Dimensionality reduction can be defined as the transformation of high-dimensional data, $X = \{x_i\}_{(i=1..N)}$ $(x_i \in R^D)$, into a meaningful and compact representation of reduced dimensionality, $Y = \{y_i\}_{(i=1..N)}(y_i \in R^d)$ where $d < D$ (and often $d << D$), to obtain more informative, descriptive and practical data representation for further analysis. This process is achieved by eliminating redundancies present in datasets while ensuring the maximum possible preservation of information.

Since most real datasets are highly nonlinear, many nonlinear dimensionality reduction techniques have been proposed. They can be classified in two main categories: mapping-based and embedding-based. Mapping-based approaches such as GPLVM (Lawrence, 2004) and generative topographic mapping (Bishop et al., 1998) use probabilistic nonlinear functions to map the embedded space to the data space. Their main limitation comes from the computational cost of their learning process which restricts their usage to relatively small datasets. On the other hand, embedded-based approaches, also called spectral methods, estimate the structure of the data underlying manifold by approximating each data point according to their neighbours on the manifold. Although these methods do not provide any explicit mapping between low and high dimensional spaces, they have proved very popular because they can handle very large high dimensional datasets with a reasonable computational cost.

Spectral methods can broadly be divided into three families, i.e. Isometric Feature Mapping (Isomap) (Tenenbaum et al., 2000), Locally Linear Embed-

ding (LLE) (Roweis and Saul, 2000) and Laplacian Eigenmaps (LE) (Belkin and Niyogi, 2001), according to the way data point positions are expressed in function of their neighbours. Since they have been a very active area of research, many extensions and improvements have been suggested (Choi and Choi, 2004; De Ridder et al., 2003; De Silva and Tenenbaum, 2003; Donoho and Grimes, 2003; Goldberg and Ritov, 2008; He and Niyogi, 2004; He et al., 2005; Kokiopoulou and Saad, 2007; Wang and Li, 2009; Yang, 2003; Zhang and Wang, 2007; Zheng et al., 2008). Despite this research effort, these approaches still suffer from the fact they rely on a set of values which are chosen empirically, i.e. neighbourhood size and mapping function parameters.

In this paper, we address this fundamental problem by proposing two extensions of spectral dimensionality reduction methods allowing their automatic configuration. First, optimal values of neighbourhoods are estimated by adopting mutual information measure (MI) (Cover and Thomas, 1991). Secondly, mapping functions are customised to datasets with a novel usage of Radial Basis Function network (RBFN) (Poggio et al., 1990), where network topology is automatically learnt by Markov cluster algorithm (MCL) (Dongen, 2000).

After a detailed description of the main spectral dimensionality reduction approaches and their limitations, we describe the new techniques we propose to allow their automatic configuration. Finally, they are validated on synthetic and real datasets. Since spectral dimensionality reduction methods derives from either Isomap, LLE or LE, our contribution will be applied to these three methods which are used as representatives of all embedded-based approaches.

2. Spectral dimensionality reduction methods and their limitations

Spectral or embedding-based approaches model the structure of data by preserving some geometrical property of the underlying manifold. While the Isomap (Tenenbaum et al., 2000) method attempts to maintain global properties, LE (Belkin and Niyogi, 2001) and LLE (Roweis and Saul, 2000) aim at preserving local geometry which implicitly tends to keep the global layout of the data manifold. After a brief description of these techniques, we list the main limitations we address in this paper.

2.1. Processing pipeline

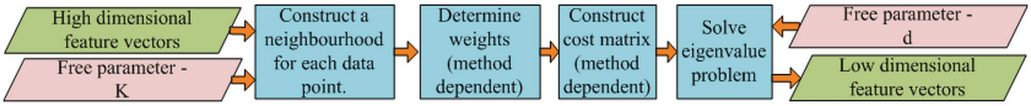


Figure 1: Dimensionality reduction using spectral methods.

These methods share the same algorithm structure as illustrated in figure 1. First, the neighbourhood for each data point is constructed by choosing K-nearest neighbours based on Euclidean distance. Then, weights, which express the geometrical relationship between each data point and its neighbours, are determined according to the property to be preserved. In LLE, they summarize the neighbours contribution to the reconstruction of a data point (Roweis and Saul, 2000). In LE and Isomap, the weights are related to the distance between a point and its neighbours using respectively heat kernel (Belkin and Niyogi, 2001) and Euclidean distance (Tenenbaum et al., 2000). Then, each method optimises its own cost function subject to constraints that make the problem well-posed. In the case of the LE and Isomap algorithms, the manifold is approximated, first, by an adjacency graph where

nodes correspond to data point and edges represent weights between points. A sparse cost matrix is constructed directly for LLE and LE (Belkin and Niyogi, 2001; Roweis and Saul, 2000), whereas the Isomap dense cost matrix is obtained by calculating geodesic distances between all pairs of points in the graph (Tenenbaum et al., 2000). Finally, spectral embedding is calculated using the eigenvectors of the cost matrix.

2.2. Limitations

The main issue of spectral methods is that the quality of embedded space is very sensitive to the choice of free parameters and they do not provide a mapping function between low and high dimensional spaces.

All approaches have two free parameters: 'd' and 'K'. 'd' is the dimensionality of the embedded space and must be known a priori because it is used in the minimization process. If the number of dimensions is too low, important data features may be collapsed onto the same dimension. 'K' is the neighbourhood size. If it is too small, global feature information is lost since the manifold is split into unconnected pieces. If it is too large, the LE and LLE assumption that a data point and its neighbours are locally linear is not valid. In the case of Isomap, a large K introduces errors in geodesic distances.

Since the effectiveness of a method depends on the choice of these parameters, many techniques have been proposed to estimate automatically their optimal values. The optimal dimensionality of the embedded space is defined as the intrinsic dimension of the high dimensional data. More formally, a dataset $X \in R^D$ is said to have intrinsic dimensionality (ID) equal to d if its elements lie entirely within a d-dimensional subspace of R^d (where $d << D$)

(Fukunaga, 1982). Estimation of 'd' can be achieved using many approaches (see (Camastra, 2003) for a detailed review) including maximum likelihood estimation (Levina and Bickel, 2005), packing numbers (Kegl, 2003), analysis of a geodesic minimum spanning tree (Costa and Hero, 2004), fractal-based methods (Camastra, 2003) and eigenvalue-based estimator (Fukunaga and Olsen, 1971) (EE). However, none of them has achieved consensus as the most accurate method.

The selection of the optimal neighbourhood size 'K' is also an open problem. The main line of research has focused on assessing directly the quality of embedded spaces by a quantitative measure in order to infer the optimal value of 'K'. Although many measures have already been proposed, such as Residual Variance (Kouropteva et al., 2002; Samko et al., 2006), Spearman Rho (Karbauskait et al., 2007; Samko et al., 2006) and Procrustes Analysis (Goldberg and Ritov, 2009), experiments suggest their accuracy depends not only on the choice of intrinsic dimensionality but also on the nature of dataset. Consequently, they are not suitable when dealing with complex non-linear high dimensional data of a nature, which is different from that they have been designated for, e.g. human motion (Lewandowski et al., 2009).

Finally, an inherent limitation of spectral dimensionality reduction approaches is that they do not provide an explicit mapping function between low and high dimensional spaces. Such function is essential for ensuring continuity of low dimensional representation and projecting data between spaces. This issue has been addressed quite satisfactorily by applying Radial Basis Function network (Poggio et al., 1990) to approximate the optimal mapping function (Elgammal and Lee, 2007; He et al., 2004; Lewandowski

et al., 2009). However, the quality of RBFN relies on the careful selection of a few parameters which are chosen empirically.

3. Automatic configuration of spectral dimensionality reduction methods

We contribute to the current state of the art by addressing two essential problems: the selection of the optimal neighbourhood size 'K' and the absence of mapping function between spaces. First, we propose to estimate the optimal neighbourhood size by assessing the quality of discovered embedding spaces using the mutual information measure. Secondly, we overcome the deficiency of mapping function by extending advanced RBFN by exploiting spectral graphs to design the optimal structure of the network in an unsupervised manner. The above schemas are integrated into a general framework for the automatic configuration of spectral dimensionality reduction methods.

3.1. Estimation of optimal neighbourhood size

The optimal neighbourhood size 'K' can be identified directly by assessing embedded space quality. The process is the following. First, data are divided into training and testing sets. Then, for a given value of 'K', dimensionality reduction is applied on the training set and a mapping function is built between the original and embedded spaces. Finally, test data are projected into the low dimensional space and some error metric is calculated. This process is repeated for a range of 'K' values so that the optimal neighbourhood size can be identified.

Since this process requires calculating computationally expensive mapping functions for all possible values of 'K', quantitative metrics have been

proposed to evaluate the quality of an embedded space without mapping. The standard procedure of optimal neighbourhood size estimation using a quantitative metric is summarized in pseudo-code 1. There are three metrics commonly used. Residual variance (RV) (Kouropteva et al., 2002; Samko et al., 2006) expresses how well the distance information is preserved between two sets of variables, i.e. it reflects the degree of linear relationship between these variables. Spearman's rho (SR) (Karbauskait et al., 2007; Samko et al., 2006) measures the accuracy of the low-dimensional manifold in retaining the order of pair wise distances of data points of the high-dimensional. Finally, procrustes analysis measure (PA) (Goldberg and Ritov, 2009) reflects the matching of two sets of variables in terms of distances. PA determines how well linear transformations of the points in one space conforms to the points in the second space. Since experiments have suggested that these measures depend on the specific nature of datasets (Lewandowski et al., 2009), they are not suitable for the automatic selection of the free parameter 'K' in an untested domain.

Algorithm 1 Estimation of optimal neighbourhood size

Input: high dimension dataset, maximum K ($maxK$), ID estimate
Output: optimal K

Find minimum K ($minK$) which produces a fully connected graph
for each K in range $< minK, maxK >$ **do**
  Reduce dimensionality of the dataset using a spectral method
  Use metric to assess the quality of the embedded space
end for
Select optimal K according to metric

In this work, we tackle this fundamental issue by adopting a metric which

can deal with variables without any linear relationship. We propose to use the mutual information measure (Cover and Thomas, 1991) which has proved to be able to discover even marginal dependency between two spaces of variables, since, in contrast to linear correlation coefficients, it is also sensitive to dependencies which do not manifest themselves in the covariance. MI is null if and only if the two random variables are strictly independent. The first idea would be to design a cost function directly in the spectral dimensionality reduction framework using MI, however since MI expresses relationship between two sets of variables rather than individual points, it is not an appropriate metric for this purpose. As the consequence, we propose to employ it in post processing step to evaluate the quality of spaces.

The most straightforward and widespread approach for estimating MI is to partition the data and approximate MI by the following finite sum:

$$I(X,Y) = \sum_{i}^{N} \sum_{j}^{N} p(i,j) log \frac{p(i,j)}{p_x(i) p_y(j)} \tag{1}$$

where *p(i,j)* is the joint probability distribution function, and $p_x(i)$ and $p_y(j)$ are the marginal probability distribution functions of X and Y respectively. This formulation can be equivalently expressed as (Cover and Thomas, 1991):

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

where $H(X)$ and $H(Y)$ are the marginal entropies and *H(X,Y)* is the joint entropy of X and Y.

However, this standard approach can only be applied for $D = d = 1$,

because the estimation of entropy is based on data binning. Since, in our framework, we need to estimate MI measure for higher dimensional variables ($D > 1, d \geq 1$), we calculate the entropy using K-nearest neighbour statistics as proposed in (Kraskov et al., 2004). Assuming that some metric is defined on the spaces spanned by X and Y, all neighbours of a given data point are ranked according to their distance to that point. Then the entropy $H(Z)$, where $Z \in \{X, Y\}$, is estimated by the average distance to the K-nearest neighbours, averaged over all z ($z \in \{x, y\}$). This leads to the following equation (Kraskov et al., 2004):

$$H(Z) = N^{-1} \sum_{i=1}^{N} (\gamma(n_z(i) + 1)) - \gamma(N) - \log c_{d_z} - \frac{d_z}{N} \sum_{i=1}^{N} (\log \epsilon(i)) \qquad (3)$$

Here, $n_z(i)$ denotes the number of points which fulfil the condition: $\|z(i) - z(j)\| < \epsilon(i)$ and $\gamma(\cdot)$ is the digamma function (Kraskov et al., 2004). d denotes the dimension of z and c_{d_z} is the volume of the d-dimensional unit ball. Similarly, the joint entropy of X and Y for a given K (Kraskov et al., 2004) is expressed by:

$$H(X, Y) = \gamma(K) - \gamma(N) - \log(c_{d_x} c_{d_y}) - \frac{d_x + d_y}{N} \sum_{i=1}^{N} (\log \epsilon(i)) \qquad (4)$$

Combining equations 2, 3 and 4 results in the expression of multi dimensional MI:

$$I(X, Y) = \gamma(K) + \gamma(N) - N^{-1} \sum_{i=1}^{N} (\gamma(n_x(i) + 1) + \gamma(n_y(i) + 1)) \qquad (5)$$

Although mutual information has never been used in this context, the use of the multidimensional extension allows MI becoming an intuitive measure for analysing the mutual correlation between high and low dimensional spaces.

3.2. Unsupervised mapping

All spectral approaches suffer from the deficiency of not providing a mapping function. A solution has been to use RBFN based mapping (Elgammal and Lee, 2007; He et al., 2004). However, this process relies on manual adjustment of its structure according to data. In previous work, we have addressed this by introducing unsupervised RBFN (Lewandowski et al., 2009). Since that approach has some limitations (that we discuss later), we propose a novel method for designing the structure of the network which originates from graph clustering theory.

RBFN from high to low dimensional space is expressed by the following over-constrained nonlinear system of equations:

$$y = f(x) = B * \psi(x) \tag{6}$$

where B is a $D \times L$ matrix of network weights and vector $\psi(x)$ is given by:

$$\psi(x) = [\phi(\|x - c_1\|), \phi(\|x - c_2\|), \ldots, \phi(\|x - c_L\|)]^T \tag{7}$$

where L is the number of hidden layers in the network, which correspond to the coordinates of centres c_j and $\phi(.)$ is a real-valued basis function. We exploit Gaussian basis function $\phi(\|x_i - c_j\|) = e^{\frac{\|x_i - c_j\|^2}{2\sigma^2}}$, where σ denotes the

average distance between all centres, because it has excellent approximation properties (Poggio et al., 1990). The solution for B can be found by applying the Moore-Penrose pseudo-inverse on matrix $\psi(X)$ in equation 6 and solving the obtained linear system of equations.

The RBFN structure is formed by centres c_j which summarize training data points in order to provide generalization properties of the network. However, the performance of RBFN critically depends upon the chosen centres (Chen et al., 1991). K-means clustering (Kanungo et al., 2002) (KMC) and rival penalized competitive learning (Xu et al., 1993) (RPCL) are currently the most popular and well studied methods which address this task. A key drawback of the KMC algorithm is that it requires prior knowledge about the correct number of centres. This can be addressed using the RPCL algorithm which is capable of finding the optimal localisation of centres as well as their correct number L in an automatic way. First, L' centres are randomly initialised ($L'>>L$). Subsequently, in each iteration, the algorithm randomly selects a sample s from the training set and moves the closest centre (the so called competition winner) towards the considered point s by a weighted distance *w1*. In the same step the second closest centre (or rival) is pushed away from the sample s by a weighted distance *w2* (where *w1*>>*w2*). Learning rates, i.e. *w1*,*w2* are monotonically decreased after each iteration. The entire procedure is repeated until its converges or reaches a given threshold. This mechanism allows automatic determination of the centres' positions by locating them at the core of data point clusters and gradually driving unrequired centres away from those clusters.

In earlier work (Lewandowski et al., 2009), we automated the mapping

process by applying RPCL for training of RBFN. However, RPCL, as KMC, depends on the initial random localization of centres and relies on the Euclidean distance, which is not the most appropriate metric to model high dimensional relationships (Aggarwal et al., 2001). In order to improve accuracy, we extend our idea of unsupervised mapping learning and propose to use the Markov cluster algorithm (MCL) (Dongen, 2000) to identify the suitable number and localization of centres automatically by exploiting the adjacency graph constructed during spectral reduction of dimensionality. As it will be demonstrated in the results section, the computational cost of a mapping learning process is greatly reduced and the obtained mapping exhibits better accuracy in comparison to standard approaches such as KMC and RPCL.

At the heart of the MCL algorithm (Dongen, 2000) lies the idea to simulate flow within a graph: flows are promoted where current is strong and demoted where current is weak. Flow simulation is achieved by transforming a graph into a Markov graph using the standard definition of a random walk on a graph. Then a flow is defined by two simple algebraic operations, i.e. expansion and inflation, which are applied connectively, so that the flow becomes thicker in regions of higher current and thinner in regions of lower current.

According to this paradigm, if natural groups are present in the spectral graph obtained in the first step of dimensionality reduction, then, current across borders between different groups will wither away. As the result, a fully connected graph is divided into few subgraphs (figure 2), thus revealing the optimal number L as well as coordinates of clusters c_j. Application of

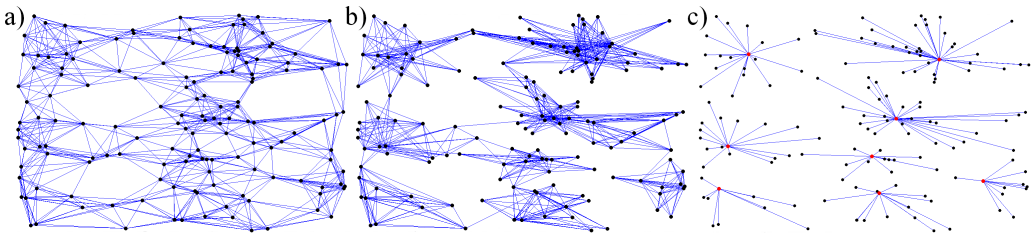


Figure 2: 2D representation of successive stages of flow simulation using the MCL process for discovery of the localisation and the number of centres in RBFN.

this procedure enables the discovery of more representative clusters of high dimensional data and subsequently customise RBFN structure to dataset in an automatic and efficient manner.

4. Experimental results and discussion

4.1. Datasets

The proposed framework was validated with both artificial and real datasets. Standard datasets were selected to extensively evaluate the performance and robustness of the proposed methodology in different scenarios. Figure 3 illustrates the datasets used in this work. Since the intrinsic dimensionalities of the digits and face datasets are unknown, we used both low and high values of their estimates in order to perform our experiments.

The 'swissroll' dataset is a synthetic and nonlinear example of a two dimensional flat submanifold which lies in a three-dimensional space. This dataset exhibits significant disagreement between geodesic and Euclidean distances (figure 3a). 2000 points were randomly sampled from the manifold and used in all our experiments. In addition, we generated a second smaller dataset consisting of 1000 points (denoted by a star in our experiments) in

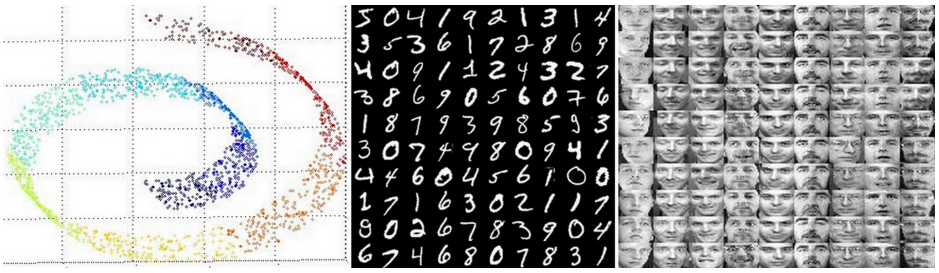


Figure 3: Datasets used in the experiments: from left to right, 'swissroll' manifold, handwritten digits and face images.

order to compare Isomap results with those of the original Isomap paper (Tenenbaum et al., 2000).

The MNIST dataset (LeCun, 2000) consists of handwritten characters images containing digits from 0 to 9 (figure 3b). The size of each image is 28 x 28 pixels, with 256 gray levels per pixel. Thus, each image is represented by a 784-dimensional vector. Due to computational and memory constraints, in our experiments we used a subset of the MNIST database consisting of 6000 images. According to (Camastra and Vinciarelli, 2001), the optimum ID of handwritten digits is 7, whereas the upper bound of the ID as determined by EE equals 10.

The ORL (formerly Olivetti) face database contains 400 images of 40 distinct subjects (Samaria and Harter, 1994) (figure 3c). All images were captured against a dark homogeneous background with the subjects in an upright, frontal position, with tolerance for some side movements. There are variations in facial expression (open/closed eyes, smiling/nonsmiling), and facial details (glasses/no glasses, different skin colours). The images are greyscale with a resolution of 64x64 pixels which gives a 4096 dimension feature vector. The analysis of relation between recognition rates and dimensionality

of embedded space in (Yin et al., 2008) suggests a value of 10 as the optimal ID for this dataset. The upper bound of the ID as determined by EE equals 40.

4.2. Experiments

All experiments were performed with Isomap, LLE and LE using K values in the range $\langle 4, 30 \rangle$. In multidimensional spaces, geodesic distances are used, whereas on the plane we employ Euclidean distances as suggested in (Samko et al., 2006).

First, we evaluate qualitatively the novel MI estimator against current approaches, i.e. Residual Variance, Spearman Rho and Procrustes Analysis measures. This was performed using the synthetic dataset for which the underlying structure is known so the quality of embedded space can be judged visually.

Then, two classical pattern classification problems, face and handwritten digit recognition, are considered in order to analyze the quantitative performance of the MI metric. We do not perform any preprocessing or normalization of the data in order to prevent any information lost. It is important to note that, in this work, we did not focus on designing a state of art classification system, but to compare existing metrics with the one we propose using on a standard classification framework based on a real application.

Finally, in the last experiment we show superiority of graph based RBFN in comparison with standard RBFN. This is achieved by repeating the classification experiments with digits and faces recognition using the new mapping function whose structure is inferred automatically from the spectral graphs.

4.2.1. Dimension reduction of 'swissroll' dataset

Table 1 presents the low dimensional spaces of 'swissroll' dataset produced by Isomap, LE and LLE using the estimated neighbourhood sizes calculated by RV, SR, PA and MI.

In all cases, the MI measure was able to identify very good low dimensional representation of 'swissroll' dataset, i.e. embedded space which manages to unroll manifold and preserves local structure. Moreover, estimated values of K using MI are in agreement with parameters which were recommended in the original papers (Belkin and Niyogi, 2001; Roweis and Saul, 2000; Tenenbaum et al., 2000). Although, other measures usually select reasonable low dimensional representations, their quality is not consistent. For instance, the local structure is distorted in most experiments involving RV/SR. Although PA seems to behave similarly to MI, in the case of LLE the very different neighbourhood size returned by PA leads to the production of an embedded space of inferior quality.

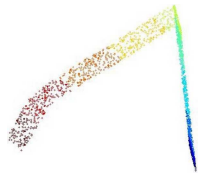
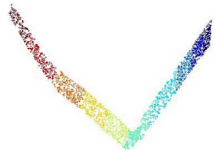
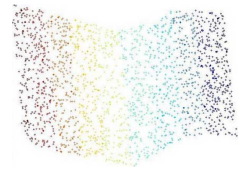
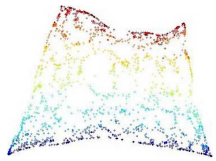
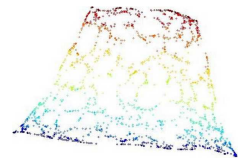
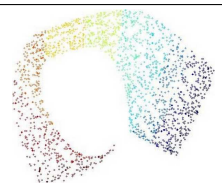
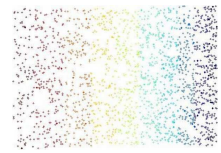
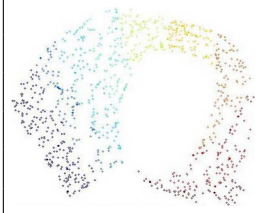
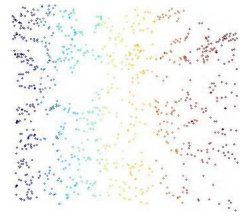
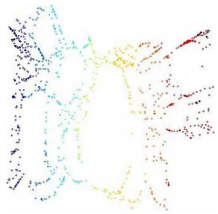
Method (Recommended K)	Coefficient (Estimated K)	Visualization	Coefficient (Estimated K)	Visualization
LLE (20) (Roweis and Saul, 2000)	RV (11)		SR (22)	
	PA (8)		MI (20)	
LE (5-15) (Belkin and Niyogi, 2001)	RV/SR (8)		PA/MI (5)	
Isomap (-)	RV (21)		SR/PA/MI (18)	
Isomap* (7) (Tenenbaum et al., 2000)	RV (9)		PA/MI (7)	
	SR (4)		* denote the 'swissroll' dataset with 1000 points instead of 2000 points	

Table 1: The low dimensional spaces of 'swissroll' with estimated and recommended neighbourhood sizes for Isomap, LE and LLE according to coefficients RV, SR, PA and MI.

4.2.2. Classification evaluation

The recognition of either digits or faces is performed according to the 10-fold cross validation strategy, where we divide a dataset into ten distinct partitions. For each partition, we reduce dimensionality of remaining dataset and train RBFN with the standard RPCL algorithm. Then, each partition is projected into the low dimensional space and classification is performed using a first nearest neighbour classifier (Ho, 1998). Finally, classification accuracy is calculated by averaging over the ten partitions. For each dataset, estimation of optimal neighbourhood size for dimentionality reduction is calculated using RV, SR, PA and MI. Moreover, the actual optimal K, 'Opt', is calculated experimentally by an exhaustive evaluation of classification accuracy for all values of K within the range $\langle 4, 30 \rangle$. In addition, using that value, we evaluate the classification accuracy of the scheme, 'Opt*', which includes graph-based RBFN (G-RBFN). Tables 2 and 3 show the results of these experiments which were conducted with two sets of IDs as defined in section 4.1. Note that the huge computational cost of applying PA on the very high dimensional faces dataset (dimensionality of 4096) did not allow us to obtain the results for this measure using our processing capabilities (16-node cluster).

	ID	RV	SR	PA	MI	Opt	Opt*
Iso	10	88	88	88	88	89	90
LLE		62	63	59	78	78	82
LE		79	79	80	80	80	84
Iso	7	85	82	84	85	85	87
LLE		56	63	53	74	74	77
LE		75	75	74	76	77	80

Table 2: Percentage accuracy of hand-written digits recognition.

	ID	RV	SR	PA	MI	Opt	Opt*
Iso	40	76	73	-	77	77	77
LLE		78	78	-	80	80	80
LE		67	67	-	67	68	73
Iso	10	65	57	-	76	76	76
LLE		55	55	-	61	62	62
LE		62	50	-	63	63	63

Table 3: Percentage accuracy of faces recognition.

In agreement with our previous experiments, neighbourhood sizes estimated by the MI measure produce consistently better accuracy than those suggested by other metrics regardless of the chosen ID. Moreover, it allows classification performances which are either optimal or near-optimal for a given dimensionality reduction method. Results also reveal that unlike LLE and Isomap, LE is not very sensitive to neighbourhood size selection. As expected, decrease of ID results in a decline of accuracy since more discriminative information is discarded during dimensionality reduction. Two dimensional visualization of the best low dimensional space obtained with Isomap for the digit dataset is presented in figure 4.

Regarding the efficiency of graph-based RBFN, tables 2 and 3 show that this new scheme improves significantly the quality of the mapping produced by standard RPCL RBFN. Further comparison between those two mapping methods is provided in figure 5, where classification accuracy and processing time are measured for various sizes of the digits dataset. Here, LE is used for dimentionality reduction as a representative of spectral methods.

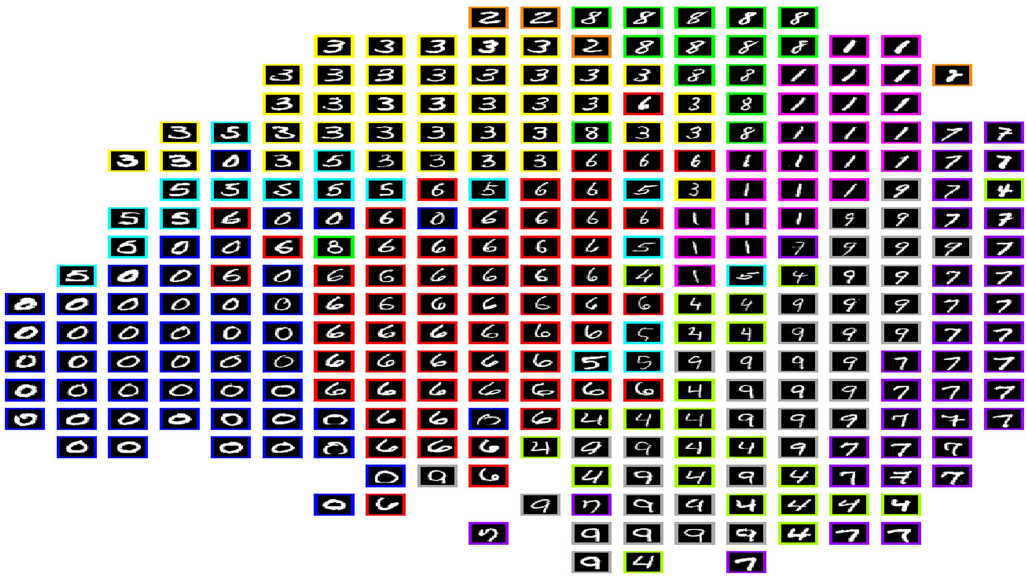


Figure 4: Two dimensional visualization of the best low dimensional space obtained with Isomap for MNIST data subset.

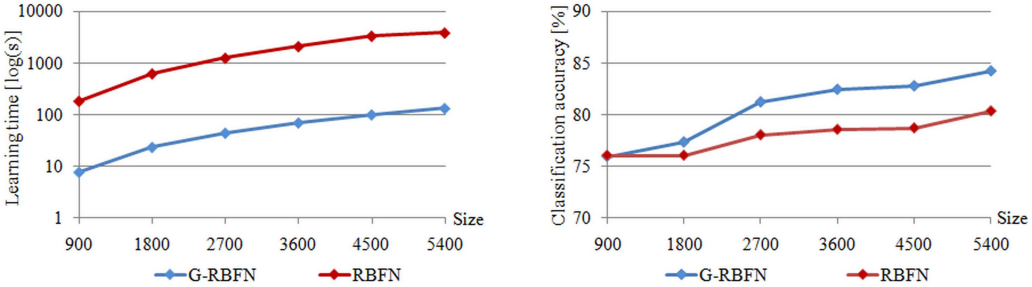


Figure 5: Classification processing time (left) and accuracy (right) comparisons between graph-based RBFN and standard RPCL RBFN according to digits dataset size (ID=10).

First, whatever the size of the training set, classification accuracy using graph-based RBFN is higher than for standard RBFN. Secondly, graph-based RBFN is computationally very efficient since the learning process time tends to increase linearly with the size of the database, while it grows quadratically when it is performed using the RPCL procedure.

4.3. Discussion

All experiments demonstrate that MI is a better metric to estimate neighbourhood size than currently used measures. Not only are embedded spaces produced by MI visually convincing, but our quantitative study, i.e. classification experiments, confirm its superiority since it consistently provides better accuracy regardless of the estimated ID. Moreover, unlike PA, MI proved able to handle very high dimensional datasets. Our quantitative experiments also validate our proposal of using graph-based RBFN to provide mapping between embedded and data spaces. This scheme outperforms significantly standard RBFN mapping in both accuracy and computational efficiency when combined with spectral dimensionality reduction methods.

Although we used classification experiments to validate quantitatively the value of our contribution to spectral dimensionality reduction methods, our aim was not to produce a state of the art classifier, but to demonstrate that our innovations could be applied successfully to representatives of the three main spectral families, i.e. Isomap, LLE and LE. We would suggest readers with a special interest in classification to apply our advanced techniques to spectral methods which were developed especially to handle that task. They include discriminant Isomap (Yang, 2003), supervised LLE (De Ridder et al., 2003) and semi-supervised LE (Zheng et al., 2008).

5. Conclusions

In this paper, a framework has been proposed to configure automatically spectral dimensionality reduction methods. This is achieved by, first, estimating the optimal neighbourhood size. We introduce the MI metric as a powerful alternative to existing techniques. Secondly, RBFN based mapping between spaces has to be learnt in a unsupervised manner. Although the RPCL algorithm is the standard approach, we enhance significantly the learning process by using the efficient graph based MCL algorithm. Our contributions have been validated qualitatively and quantitatively using various datasets. Results demonstrate that neighbourhoods selected by the MI metric allow spectral dimensionality reduction methods to produce better quality embedded spaces. Moreover, our new mapping functions improve both mapping accuracy and computational efficiency.

Aggarwal, C., Hinneburg, A., Keim, D., 2001. On the surprising behavior of distance metrics in high dimensional spaces. Lecture Notes in Computer Science 1973, 420–434.

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems 14 14, 585–591.

Bishop, C., Svensén, M., Williams, C., 1998. GTM: The generative topographic mapping. Neural computation 10 (1), 215–234.

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

Camastra, F., Vinciarelli, A., 2001. Intrinsic dimension estimation of data: An approach based on grassberger-procacciaś algorithm. Neural Processing Letters 14 (1), 27–34.

Chen, S., Cowan, C., Grant, P., 1991. Orthogonal least squares learning algorithm for radial basisfunction networks. IEEE Transactions on neural networks 2 (2), 302–309.

Choi, H., Choi, S., 2004. Kernel isomap. IET Electronics letters 40 (25), 1612–1613.

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

Cover, T. M., Thomas, J., 1991. Elements of information theory. Wiley.

De Ridder, D., Kouropteva, O., Okun, O., Pietikainen, M., Duin, R., 2003. Supervised locally linear embedding. Lecture Notes in Computer Science 2714, 333–341.

De Silva, V., Tenenbaum, J., 2003. Global Versus Local Methods in Nonlinear Dimensionality Reduction. Advances in neural information processing systems 15, 705–712.

Dongen, S., 2000. Graph clustering by flow simulation. Ph.D. thesis, PhD Thesis, University of Utrecht, The Netherlands.

Donoho, D., Grimes, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences 100 (10), 5591–5596.

Elgammal, A., Lee, C., 2007. Nonlinear manifold learning for dynamic shape and dynamic appearance. Computer Vision and Image Understanding 106 (1), 31–46.

Fukunaga, K., 1982. Intrinsic dimensionality extraction. Classification, Pattern Recognition and Reduction of Dimensionality, 347–362.

Fukunaga, K., Olsen, D. R., 1971. An algorithm for
nding intrinsic dimensionality of data. IEEE Transactions on Computers C-20 (2), 176–183.

Goldberg, Y., Ritov, Y., 2008. Ldr-lle: Lle with low-dimensional neighborhood representation. ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II, 43–54.

Goldberg, Y., Ritov, Y., 2009. Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. Machine Learning 77 (1), 1–25.

He, X., Cai, D., Yan, S., Zhang, H., 2005. Neighborhood preserving embed-

ding. Proceedings of IEEE International Conference on Computer Vision 2, 1208–1213.

He, X., Ma, W., Zhang, H., 2004. Learning an image manifold for retrieval. Proceedings of ACM Multimedia, 17–23.

He, X., Niyogi, P., 2004. Locality preserving projections. Advances in Neural Information Processing Systems 16, 153–160.

Ho, T., 1998. Nearest neighbors in random subspaces. Lecture Notes in Computer Science 1451, 640–648.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A., 2002. A local search approximation algorithm for k-means clustering. Proceedings of the eighteenth annual symposium on Computational geometry, 10–18.

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

Kegl, B., 2003. Intrinsic dimension estimation using packing numbers. Advances in Neural Information Processing Systems 15, 681–688.

Kokiopoulou, E., Saad, Y., 2007. Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (12), 2143–2156.

Kouropteva, O., Okun, O., Pietikainen, M., 2002. Selection of the optimal parameter value for the locally linear embedding algorithm. Proc. of the 1 st Int. Conf. on Fuzzy Systems and Knowledge Discovery, Singapore, 359–363.

Kraskov, A., Stoegbauer, H., Grassberger, P., 2004. Estimating mutual information. Physical review. E, Statistical, nonlinear, and soft matter physics 69 (6), 66138.

Lawrence, N., 2004. Gaussian process latent variable models for visualisation of high dimensional data. Advances in Neural Information Processing Systems 16.

LeCun, Y., 2000. MNIST handwritten digits dataset (http://yann.lecun.com/exdb/mnist/index.html) [last accessed on 30/10/2009].

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

Lewandowski, M., Makris, D., Nebel, J.-C., 2009. Automatic configuration of spectral dimensionality reduction methods for 3D human pose estimation. Visual Surveillance.

Poggio, T., Girosi, F., MIT, C., 1990. Networks for approximation and learning. Proceedings of the IEEE 78 (9), 1481–1497.

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

Samaria, F., Harter, A., 1994. Parameterisation of a Stochastic Model for Human Face Identification (http://www.cs.toronto.edu/˜roweis/data.html) [last accessed on 30/10/2009]. Workshop on Applications of Computer Vision.

Samko, O., Marshall, A., Rosin, P., 2006. Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognition Letters 27 (9), 968–979.

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

Wang, Q., Li, J., 2009. Combining local and global information for nonlinear dimensionality reduction. Neurocomputing 72 (10-12), 2235–2241.

Xu, L., Krzyzak, A., Oja, E., 1993. Rival penalized competitive learning for clustering analysis, RBFnet, and curve detection. IEEE Transactions on Neural networks 4 (4), 636–649.

Yang, M., 2003. Discriminant isometric mapping for face recognition. Lecture Notes in Computer Science 2626, 470–480.

Yin, J., Hu, D., Zhou, Z., 2008. Growing locally linear embedding for manifold learning. Journal of Pattern Recognition Research 2 (1), 1–16.

Zhang, Z., Wang, J., 2007. Mlle: Modified locally linear embedding using multiple weights. Advances in Neural Information Processing Systems 19, 1593–1600.

Zheng, F., Chen, N., Li, L., 2008. Semi-supervised Laplacian eigenmaps for dimensionality reduction. Wavelet Analysis and Pattern Recognition 2, 843–849.