UNSUPERVISED DISCRIMINANT EMBEDDING IN CLUSTER SPACES

Eniko Szekely, Eric Bruno and Stephane Marchand-Maillet
University of Geneva, 7 Route de Drize, Geneva, Switzerland

Keywords: Dimension reduction, Clustering, High dimensionality.

Abstract: This paper proposes a new representation space, called the *cluster space*, for data points that originate from high dimensions. Whereas existing dedicated methods concentrate on revealing manifolds from within the data, we consider here the context of clustered data and derive the dimension reduction process from cluster information. Points are represented in the cluster space by means of their a posteriori probability values estimated using Gaussian Mixture Models. The cluster space obtained is the optimal space for discrimination in terms of the Quadratic Discriminant Analysis (QDA). Moreover, it is shown to alleviate the negative impact of the curse of dimensionality on the quality of cluster discrimination and is a useful preprocessing tool for other dimension reduction methods. Various experiments illustrate the effectiveness of the cluster space both on synthetic and real data.

1 INTRODUCTION

Data mining and knowledge discovery are concerned with detecting relevant information or knowledge in data. Structures or clusters constitute such type of information and their detection, performed with the goal of better understanding the data that is being analyzed, represents an active research area of data mining. Therefore two aspects become important here: structure detection - the algorithms - and structure understanding - by humans. The importance of both these aspects has led our work to concentrate on providing a cluster-driven dimension reduction method capable of jointly accounting for both these aspects.

Reducing the dimensionality of the data is a problem that is capturing more and more attention of the data mining and machine learning communities due to the necessity of understanding data in fields like image analysis, information retrieval, bioinformatics, market analysis. Dimension reduction is motivated by: 1) the supposition that data lies in spaces of lower dimensionality than the original spaces; 2) the need of reducing the computational load of high-dimensional data processing and 3) the necessity of visualizing data.

In many datasets, data is naturally organised into clusters. Taking the field of information retrieval and given a query, a document's 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.

1.1 Motivation and Contributions of the Paper

We consider to be given a set of N data points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ that are assumed to come from a multimodal distribution, that is, they are organised into K clusters.

Problem. *We search for a new representation space $\mathcal{S}$ - the cluster space - that can discriminate and emphasize clusters in case they exist.*

Data points are considered to originate from a D-dimensional space $\mathbb{R}^D$ where each point $\mathbf{x}_i$ is represented by the D-dimensional feature vector $\mathbf{x}_i = \{x_i^1, x_i^2, ..., x_i^D\}$ for all $i = 1..N$. Generally, the number of clusters K is (a lot) smaller than the number of original dimensions D.

Existing dimension reduction methods are blind to the structure of the data making identification of clusters in reduced spaces difficult. Nevertheless, the need for structure preserving during the process of reduction is important as, apart from a continuous in-

Szekely E., Bruno E. and Marchand-Maillet S. (2009).
UNSUPERVISED DISCRIMINANT EMBEDDING IN CLUSTER SPACES.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 70-76
DOI: 10.5220/0002306500700076

spection of the reduced space, recovering the structures existing in the original space is of importance in a number of applications: medecine (recovering the groups of diseases), information retrieval (recovering the groups of information (image, text)) etc.

Therefore, we propose an embedding in the cluster space, where point coordinates are calculated by means of their relative distances to each of the clusters. The algorithm starts with a first step of clustering. Once the cluster information is collected in the original space using a Gaussian Mixture Model, the discriminant functions provide the coordinates of points in the *cluster space*. Moreover, considering that the estimation of the GMM parameters is optimal, the cluster space represents the optimal space for discrimination.

The next section revisits related work. Section 3 formally defines the *cluster space*. Experiments on artificial and real data and comparisons with other dimension reduction methods are described in Section 4. The paper ends with discussions and conclusions in Section 5.

2 RELATED WORK

Many different approaches were proposed for the embedding high-dimensional data into low-dimensional spaces. Among the coordinate-based methods, the linear method of Principal Components Analysis is the most commonly used. It tries to linearly capture as much as possible from the variance in the data. Methods based on pairwise distance matrices were designed either: 1) to preserve as faithfully as possible the original Euclidean interpoint distances (Multidimensional Scaling (MDS) (Borg and Groenen, 2005), Sammon Mapping (Sammon, 1969) - which increases the weight given to small distances) or 2) to preserve non-linear transformation of distances (Nonlinear MDS (Borg and Groenen, 2005)) or 3) to unfold data that lies on manifolds (Isomap (Tenenbaum et al., 2000), Curvilinear Component Analysis (CCA) (Demartines and Hérault, 1997), Curvilinear Distance Analysis (CDA) (Lee et al., 2000)).

Manifolds are non-linear structures where two points, even if close with respect to the Euclidean distance, can still be located far away on the manifold. Isomap and CDA use the *geodesic* distance, that is, the distance over the manifold and not through the manifold. Both CCA and CDA weight the distances in the output space and not in the input space like MDS, Isomap or Sammon Mapping do. Different from Isomap, which is a global method, Locally Linear Embedding (Roweis and Saul, 2000) is a local method which tries to preserve the local structure - the linear reconstruction of a point from its neighbours. Similar to LLE, Laplacian Eigenmaps (Belkin and Niyogi, 2002) build a neighborhood graph and embed points with respect to the eigenvectors of the Laplacian matrix. Stochastic Neighbour Embedding (Hinton and Roweis, 2002) rather than preserving distances, preserves probabilities of points of being neighbours of other points. The methods presented are not capable of projecting new testing points in the reduced space, since the embedding has to be recomputed each time a new point is added.

In the introduction we discussed the importance of preserving cluster information in reduced spaces. Clustering is generally approached through hierarchical or partitional methods. Hierarchical clustering generates a tree (a dendrogram) with each node being connected to its parent and with nodes at lower levels being more similar than nodes at higher levels. Partitional methods partition the data into different clusters by doing a hard assignement - each point belongs to exactly one cluster. Soft clustering, on the other side, assigns to each point different degrees of belonging to clusters. The most common example of soft clustering is the probabilistic Gaussian Mixture Model, which assumes that data comes from a mixture of gaussians with different covariance matrices.

The idea of representing points in the space of the clusters was discussed in (Gupta and Ghosh, 2001) and in (Iwata et al., 2007). In (Gupta and Ghosh, 2001) the authors propose a Cluster Space model in order to analyze the similarity between a customer and a cluster in the transactional application area. The solution uses hard clustering on different datasets and then maps the results of the different clustering algorithms into a common space, the cluster space, where analysis is further performed to model the dynamics of the clients. In (Iwata et al., 2007) a Parametric Embedding is proposed that embeds the posterior probabilities of points to belong to clusters in a lower-dimensional space using Kullback-Leibler divergence (here posterior probabilities are considered to be given as input to the algorithm). Our approach differs from the above ones in that it proposes a solution that captures the discriminant information in the embedding space.

3 CLUSTER SPACE

Let us consider that the dataset is grouped into clusters and model it using a full Gaussian Mixture Model (F-GMM). F-GMM makes the general assumption that clusters follow Gaussian distributions and they

have different general covariances $\mathbf{\Sigma}_k$.

GMM models the data as a mixture of Gaussians of the form:

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \tag{1}$$

The *a posteriori* probabilities in the GMM are given as follows:

$$p(k|\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j)} \tag{2}$$

Applying the logarithm to the *a posteriori* probabilities from (2) gives:

$$\begin{aligned} \log p(k|\mathbf{x}_i) &= \log \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j)} \\ &= \log \pi_k - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_k| \\ &\quad - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \\ &\quad - \log \sum_j \pi_j \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \mathbf{\Sigma}_j) \end{aligned} \tag{3}$$

Equation (3) can be related to the quadratic discriminant function (see (Hastie et al., 2001) for the Quadratic Discriminant Analysis) given by:

$$\delta_k(\mathbf{x}_i) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) + \log \pi_k \tag{4}$$

where π_k, $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ are estimated from the training data, in a supervised context, and new (testing) points are assigned to the cluster for which the value of the discriminant function δ_k from (4) is the highest according to:

$$\arg\max_k \delta_k(\mathbf{x}_i) \tag{5}$$

To capture the discriminant information in the dimension reduction process, we propose the following definition of the *cluster space*:

Definition 1. *The cluster space is a common space $\mathcal{S} = \{c_i^k\}$ with point coordinates c_i^k given by the values of the discriminant functions:*

$$c_i^k = \delta_k(\mathbf{x}_i) \tag{6}$$

To obtain the values of the discriminant functions and therefore of the coordinates in the *cluster space*, in an unsupervised context, the priors π_k, means $\boldsymbol{\mu}_k$ and covariances $\mathbf{\Sigma}_k$ can be estimated with a GMM. The initialization of GMM may be performed with any clustering algorithm such as partitional or hierarchical clustering, graph-based clustering etc. The quality of the embedding is sensitive to a wrong estimation of the mixture parameters, therefore this initialisation step is important. Subspace clustering may be a good choice, especially for high-dimensional data.

The log-scaling of the probabilities that appears in the equation of the discriminant function from (4) is important in the *cluster space* as it corresponds to the Mahalanobis distance value D_M between each point and each cluster center:

$$\delta_k(\mathbf{x}_i) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}\underbrace{(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)}_{D_M^2} + \log \pi_k \tag{7}$$

The Mahalanobis distance from a point to a cluster is the distance of that point to the center of the cluster divided by the width of the ellipsoid along the direction of the point. As the Mahalanobis distance takes into account the shapes of the clusters through the covariance matrices $\mathbf{\Sigma}_k$, it is well suited for the *cluster space* as it allows the capturing of cluster information contained not only in the interdistances between clusters but in their shapes too. Thus, a point close to a cluster in the Euclidean sense may be very far away in the Mahalanobian sense.

The dimensionality of the *cluster space* is given by the number of assumed clusters K. Each point $\mathbf{x}_i$ is thus represented by K coordinates c_i^k (coordinate of point $\mathbf{x}_i$ in dimension k) which correspond to the distances of the point $\mathbf{x}_i$ to the center of cluster k.

The *cluster space* given by equation (6) is the optimal space for discrimination in the framework of QDA given that the parameters of the GMM (π_k, $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$) are optimally estimated.

The *cluster space* can also be used as a gauge for clustering tendency since the more the clusters are separated, the larger the distances to the other clusters will be. Therefore the density of points around boundaries is a good indicator of class separability. A high density indicates a weak separation between the clusters, a low density indicating a high separability. Thus, further algorithms may be designed that use the *cluster space* as a mean for cluster tendency evaluation by analyzing the distribution of points around boundaries.

3.1 The Algorithm

The algorithm for finding the cluster space is presented in Table 1. The algorithm takes as input the dataset $\mathbf{X}$ and the number of clusters K and provides as output the new coordinates in the cluster space $\mathcal{S}$. In Step 1 the priors π_k, means $\boldsymbol{\mu}_k$ and covariances $\mathbf{\Sigma}_k$ are estimated using the Expectation-Maximization (EM) algorithm. The values of the discriminant functions $\delta_k(\mathbf{x}_i)$ in Step 2 are given by Equation (7). Finally, in Step 3, the coordinates of points in the new space $\mathcal{S}$ are given by the values computed in Step 2.

Table 1: The algorithm for the Cluster Space.

Input:	$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}, i = 1..N$, $\mathbf{x}_i \in \mathbb{R}^D$, K-number of clusters
Output:	$\mathcal{S} = \{c_i^k\}$, $i = 1..N$, $k = 1..K$
Step1:	Estimate the priors π_k, means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$.
Step2:	Compute the discriminant functions $\delta_k(\mathbf{x}_i)$
Step3:	$c_i^k = \delta_k(\mathbf{x}_i)$

4 EXPERIMENTS

4.1 Artificial Data

Experiment 1. We generate artificial data from 3 Gaussians in 3 dimensions as shown in Fig.1. The Gaussians are given by $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ of 200 points, $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ of 200 points, $\mathcal{N}_3(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ of 200 points with $\boldsymbol{\mu}_1 = [0\ 0\ 0]$, $\boldsymbol{\mu}_2 = [1.5\ 0\ 0]$, $\boldsymbol{\mu}_3 = [1.5\ 0\ 1]$ and the covariances:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.01 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.01 \end{pmatrix}, \ \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.01 \end{pmatrix} \tag{8}$$

Results. We see in Figure 1 that algorithms like PCA (d) and MDS (f) are not capable of separating the 3 clusters that are well separated in (a). In the *cluster space* (b) the clusters are well separated. A further dimension reduction in this space using Isomap with a Manhattan distance shows in (c) that the clusters are separated. Isomap (e) also gives good results but is dependent on the number of neighbours given to build the fully connected graph (in such cases - of well separated clusters - the number of neighbours should be quite high).

Experiment 2. The choice of K (the number of clusters, and implicitly the dimension of the cluster space) plays an important role on the quality of the embedding in the cluster space. Figure 2 presents two cases when the number of chosen K is different from the number of real clusters in the data. The goal is to confirm that the choice of K does not force unclustered data to be clustered. K is kept fixed ($K = 3$) and the number of clusters varies. We chose $K = 3$ to be able to visualize the results in a 3D space.

Results. In the first example of Figure 2, (a) and (c), K is higher than the number of clusters and we observe that a higher K does not force clusters to break. This is an important aspect since the embedding, even if based on an initial clustering, should not artificially create structures that do not exist inside the data itself. Using a soft clustering like GMM avoids forcing clusters to break, like it would happen in a hard clustering approach (k-means). In the second example, (b) and (d), K is lower that the number of clusters and we observe that two of the clusters are embedded in the same plane but they are however kept separated. In conclusion, the choice of K is important but a number of situations work well even with different values. However, as observed during experimentation, a lower K influences more drastically the quality than a higher K, thus using higher estimates for K is preferred.

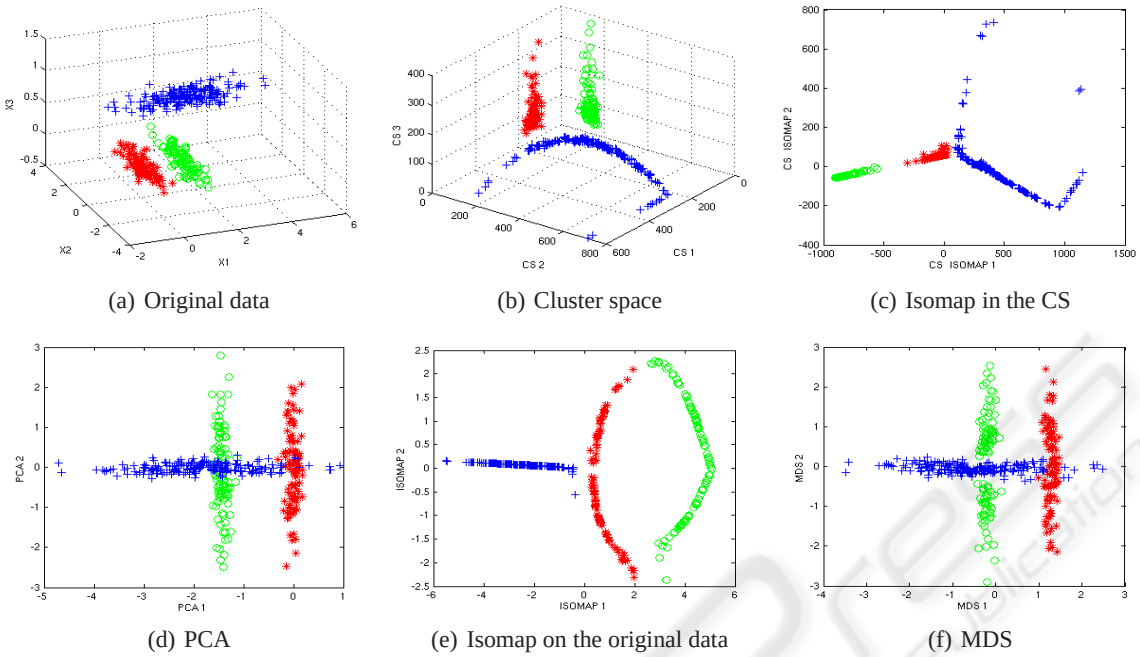
4.2 Real Data

Experiment 1. We give a first example using the Wine dataset from the UCI Machine Learning Repository. The dataset contains 3 clusters with 178 data points in a 13-dimensional space. The embedding of the dataset in a 3-dimensional space is showed in Figure 3. For evaluation we estimated the Mean Average Precision (MAP), the purity of the clustering obtained with k-means and the error of the k-Nearest Neighbour with $k = 5$. Results appear in Table 2.

Results. We observe that the cluster space captures all clusters well as opposed to other dimension reduction method like PCA or Sammon. The new representation space also allows for a clear visualization in a 3-dimensional space.

Experiment 2. One of the main application of the cluster space can be seen as a preprocessing step for further data analysis. The cluster space is useful as a preprocessing step especially when a lower-dimensional space of dimension 2 or 3 is desired for example for visualization. In this case, a dimension reduction in the cluster space can be performed using a metric that preserves the geometry of the *cluster space* especially cluster separability. To illustrate this we use a high-dimensional dataset (MNIST digit dataset originally embedded in a 784-dimensional space). Features are extracted from the data with PCA and the dimension reduction methods that we imply in the following apply in this space. The examples presented in Figure 4 show that the preprocessing in the cluster space helps Isomap to separate clusters in the 2-dimensional space.

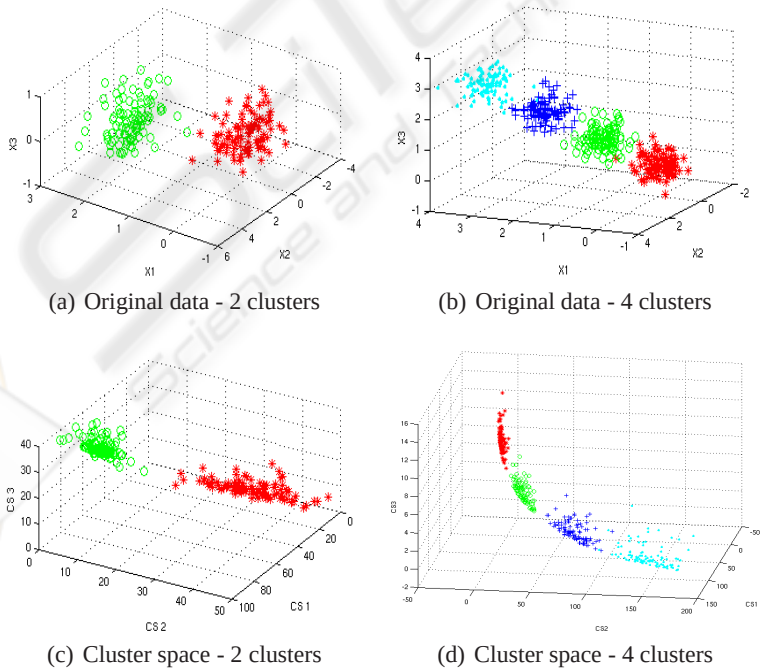
In high-dimensional spaces, estimating all the parameters necessary for a full covariance model is difficult due to the sparsity of the data. Multiple solutions are possible. One is provided by *parsimonious models*. Multiple parsimonious models have been proposed with varying complexities according



(a) Original data (b) Cluster space (c) Isomap in the CS

(d) PCA (e) Isomap on the original data (f) MDS

Figure 1: Artificial data from 3 gaussians in 3 dimensions reduced using dimension reduction methods: a) Original data projected in the 3 dimensions; b) Data projected in the *cluster space* using an EM with full covariances, $K = 3$ and the Euclidean distance; c) Data from b) reduced using Isomap with the Manhattan distance and 30 neighbours to build the graph; d) PCA in the original space; e) Isomap in the original space with the Euclidean distance and 30 neighbours; f) MDS in the original space with the Euclidean distance.



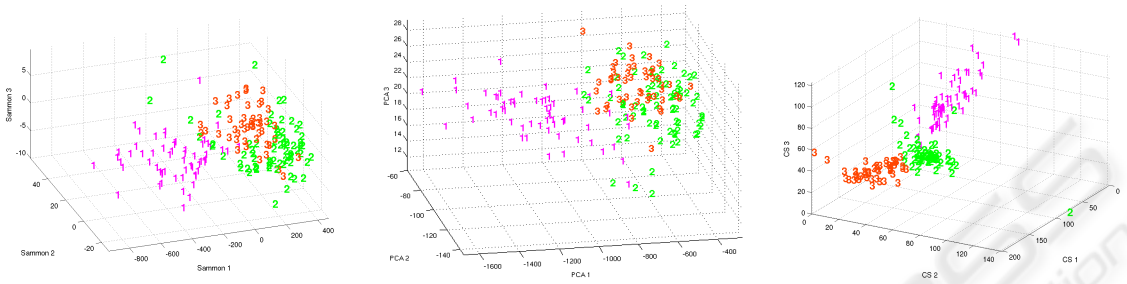
(a) Original data - 2 clusters (b) Original data - 4 clusters

(c) Cluster space - 2 clusters (d) Cluster space - 4 clusters

Figure 2: Examples on the quality of the embedding in the *cluster space* for cases when the assumed number of clusters K (here $K = 3$) is different from the real number of clusters.

Table 2: Evaluation of the Wine dataset.

Wine	Orig	PCA	Sammon	Isomap	CS
Purity	67.42	70.22	69.29	69.28	84.83
MAP	0.6433	0.6422	0.6429	0.6424	0.8499
*k*NN	69.66	69.66	69.66	68.54	94.94



(a) Wine-Sammon (b) Wine-PCA (c) Wine-Cluster Space

Figure 3: Dimension reduction for Wine with different methods.

to the specific models (intracluster and intercluster) of the covariance matrices chosen: full different covariances, full common covariance, spherical covariance (see (Fraley and Raftery, 2002) for a review). A second solution is given by truncated Singular Value Decomposition (T-SVD). The covariance matrix is ill-conditioned in high-dimensional spaces. The estimation of the inverse matrix can be resolved by using T-SVD with the first t eigenvectors $\mathbf{\Sigma}_t^{-1} = U_t^T D_t^{-1} U_t$. A third solution, that we applied, is possible by first denoising the high-dimensional data with a method such as PCA, and further start the analysis in this reduced space.

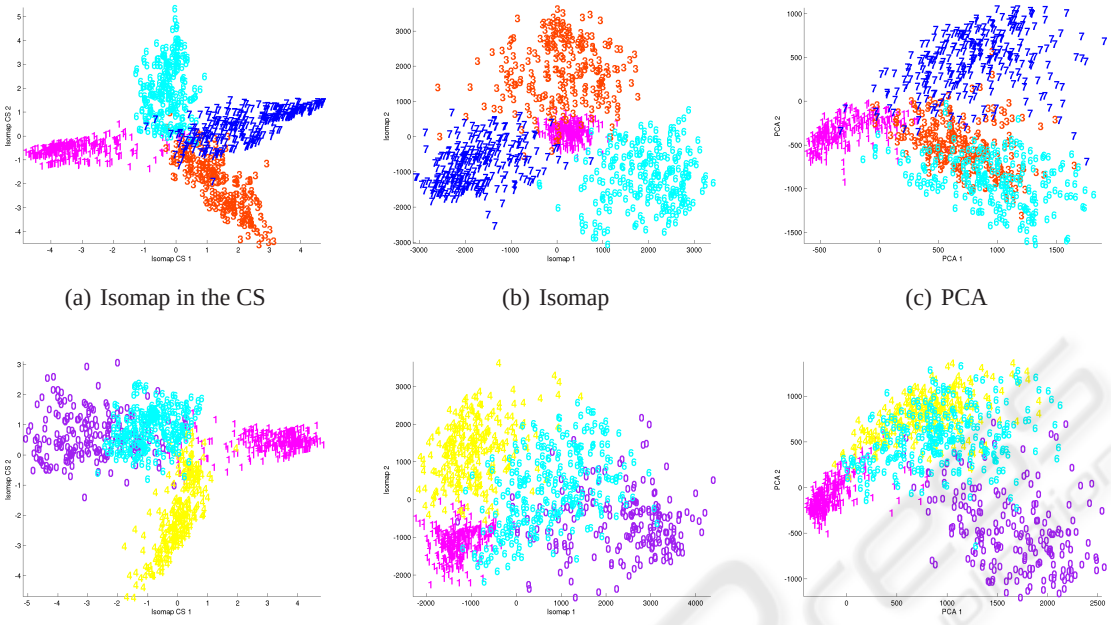
5 DISCUSSION AND CONCLUSIONS

The current construction of the *cluster space* leads to the representation of the data in a lower-dimensional space that emphasizes clusters. However, at least two issues still need to be addressed to make this construction generic:

- the presence of clusters is mapped onto the choice of an initial parameter K, directing both the modeling of the data and the dimensionality of the resulting cluster space. We have shown with different experiments (Figure 2) that our process is not drastically sensitive to a wrong estimation of this parameter (higher values are to be preferred).
- our model is based on clustering, and therefore the initialization of cluster centers is very important. However, due to the sparsity of high-dimensional spaces, a correct unsupervised initialization remains an open issue. We wish to further investigate methods for subspace clustering whose performances overcome those of traditional clustering algorithms as our results in the present reside on outputs of k-means in the initialization of the EM.

One advantage of the model is that new points can be projected in the *cluster space* (as long as they do not represent new clusters), their embedding being computed from the distances to all the clusters. The model can be further developed to estimate the parameters of the GMM in a supervised manner.

In this paper, we proposed a new representation space for embedding clustered data. Typically, the data is mapped onto the space of dimensionality K where K is given by the number of clusters and the coordinates are given by the values of the discriminant functions estimated in an unsupervised manner. We call this reduced space the *cluster space*. This space is optimal for discrimination in terms of QDA when the parameters of the GMM are optimally estimated. The cluster space is a good preprocessing step before applying other dimension reduction methods. In conclusion, the model that we propose is designed with the goal of embedding data into a low-dimensional space - the *cluster space* - where structure is to be preserved (e.g. cluster emphasis and separability).



(a) Isomap in the CS (b) Isomap (c) PCA

(d) Isomap in the CS (e) Isomap (f) PCA

Figure 4: Dimension reduction for 1000 MNIST digits (1, 3, 6, 7) and 1000 MNIST digits (0, 1, 4, 6).

ACKNOWLEDGEMENTS

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

REFERENCES

Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14.

Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Demartines, P. and Hérault, J. (1997). Curvilinear component analysis: A self-organizig neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Network*.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, pages 611–631.

Gupta, G. and Ghosh, J. (2001). Detecting seasonal trends and cluster motion visualization for very high-dimensional transactional data. In *Proceedings of the First International SIAM Conference on Data Mining*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer-Verlag.

Hinton, G. and Roweis, S. (2002). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*.

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T., and Tenenbaum, J. (2007). Parametric embedding for class visualization. *Neural Computation*.

Lee, J., Lendasse, A., and Verleysen, M. (2000). A robust nonlinear projection method. In *Proceedings of ESANN'2000, Belgium*, pages 13–20.

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

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18.

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