Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 862015, 12 pages
doi:10.1155/2008/862015

Research Article

A Metric Multidimensional Scaling-Based Nonlinear Manifold Learning Approach for Unsupervised Data Reduction

M. Brucher,[1,2] Ch. Heinrich,[1] F. Heitz,[1] and J.-P. Armspach[2]

[1] *Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, LSIIT, UMR 7005, CNRS-Université Louis Pasteur, Strasbourg 1, Boulevard S. Brant, BP 10413, 67412 Illkirch Cedex, France*
[2] *Laboratoire d'Imagerie et de Neurosciences Cognitives, LINC, UMR 7191, CNRS-Université Louis Pasteur, Strasbourg 1, LINC-IPB, 4, rue Kirschleger, 67085 Strasbourg Cedex, France*

Correspondence should be addressed to M. Brucher, brucher@lsiit.u-strasbg.fr

Received 30 September 2007; Revised 21 January 2008; Accepted 7 March 2008

Recommended by Olivier Lezoray

Manifold learning may be seen as a procedure aiming at capturing the degrees of freedom and structure characterizing a set of high-dimensional data, such as images or patterns. The usual goals are data understanding, visualization, classification, and the computation of means. In a linear framework, this problem is typically addressed by principal component analysis (PCA). We propose here a nonlinear extension to PCA. Firstly, the reduced variables are determined in the metric multidimensional scaling framework. Secondly, regression of the original variables with respect to the reduced variables is achieved considering a piecewise linear model. Both steps parameterize the (noisy) manifold holding the original data. Finally, we address the projection of data onto the manifold. The problem is cast in a Bayesian framework. Application of the proposed approach to standard data sets such as the COIL-20 database is presented.

1. INTRODUCTION

Data reduction consists in parameterizing a set of high-dimensional data with a set of reduced coordinates and mapping the original space to the reduced one and vice versa. The (noisy) data are assumed to lie close to a nonlinear manifold whose intrinsic dimension is the dimension of the reduced space. Usual goals of data reduction are visualization (as a scatter plot of the reduced data with labels from the original data, in order to get insight into the structure of the data), data understanding (the degrees of freedom are given a physical interpretation, such as pose angle or lighting intensity), classification (classification is more robust when considering the reduced space, especially in a nonlinear reduction framework), denoising, and the computation of means. Applications are face-recognition, character recognition, and shape analysis, to mention a few examples.

Manifold learning may be addressed in a more formal way. The goal is to determine, for each $\mathbf{y}_i$ on the manifold, a reduced variable $\hat{\mathbf{x}}_i$, an approximation error $\hat{\boldsymbol{\varepsilon}}_i$, and a mapping $\mathbf{f}$ such that $\mathbf{y}_i = \mathbf{f}(\hat{\mathbf{x}}_i) + \hat{\boldsymbol{\varepsilon}}_i$. To tackle the undetermination of this problem ($\mathbf{f}$ and the $\hat{\mathbf{x}}_i$'s are both unknown), the hypothesis that $\mathbf{f}$ preserves distances is introduced. Hence, the data structure in the original space is preserved in the reduced space. Consequently, the mapping $\mathbf{f}$ is determined up to an isometry, which has no practical influence on the goals of manifold learning. Although higher-order models could be considered, we will consider here piecewise linear mappings. The $\hat{\mathbf{x}}_i$'s and $\mathbf{f}$ will be determined sequentially.

Manifold learning is typically addressed by principal component analysis (PCA) with severe limitations, the main one being that only linear manifolds can be comprehended. The main consequence is that the dimension of the reduced space may be significantly overestimated when tackling nonlinear manifolds, thus hampering for example classification, visualization, or the computation of means. The main advantage of PCA over other techniques is that the reduced coordinates and the mapping $\mathbf{f}$ are computed simultaneously and easily.

A straightforward nonlinear extension of PCA is local PCA [1], allowing to fit the manifold locally. The main problem with this approach is that the different sets of reduced variables are unrelated, thus also hampering the final goals of the procedure. Several approaches, such as isometric feature mapping (Isomap, see [2]) and locally linear embedding (LLE, see [3]), address the compression problem globally

instead of blockwise, such as local PCA. The former tries to preserve the approximate geodesic distances on the manifold, considering classical multidimensional scaling (MDS) [4] while the latter aims at preserving the local structure of the manifold considering local barycentric coordinates. Both approaches suffer from noise and errors on the estimated distances in the original space [5]. We propose here a metric MDS approach [6, 7], with a view to reducing this effect: contrary to standard MDS, we consider a robust cost function to estimate the $\mathbf{x}_i$'s. We also propose a piecewise linear regression framework to map the reduced space to the original one and a Bayesian framework to determine the mapping from the original space to the reduced one.

Besides, a class of dimension reduction methods known as eigenmap techniques has recently gained interest. Those methods aim at preserving local (Laplacian eigenmaps [8], Hessian eigenmaps [9]) or global (diffusion maps [10]) structure by minimizing a cost function. In practice, the minimizations are carried out using eigenvector computations. Comparison of the proposed approach with such techniques is provided in this article. Well-identified shortcomings of eigenmap techniques are exhibited by the examples we consider.

Several other attempts have been reported in the literature to reduce the data groupwise [11–13], with an alignment correction in the reduced space, and using a piecewise linear function for the mapping between the original and the reduced spaces. To our opinion, the main drawback of these approaches is the patching of local reduced variables which may introduce distortion in the reduced space considered globally. Nevertheless, these approaches certainly deserve further attention and work.

Finally, a class of dimension reduction methods known as principal curves and principal surfaces [14, 15] attempt to reduce the data nonlinearly. These methods fundamentally only handle reduced spaces of dimension 1 or 2 which is too restrictive for the present case where we want to be able to consider reduced variables of higher dimension. LeBlanc and Tibshirani [16] have proposed an extension of such approaches in higher dimensions. In their setting, the reduced variables are re-estimated during the procedure, which does not correspond to our goal, since we want the reduced variables to reflect the structure of the original data. The model is constructed by adding (and pruning) linear pieces, which is what we also consider in our approach.

This article is organized as follows. Our approach is described in Section 2. Application examples are given in Section 3. Conclusion and prospects are presented in Section 4.

2. THE NONLINEAR MANIFOLD LEARNING PROCEDURE

In this section, we detail the proposed two-stage learning algorithm, consisting in compression (i.e., determination of the $\hat{\mathbf{x}}_i$'s) and in regression (i.e., estimation of the mapping $\mathbf{f}$). The compression step is in fact a feature extraction step, where the features are determined as an implicit nonlinear function of the original data. Feature extraction is opposed to feature selection, where a meaningful subset of the original coordinates has to be selected. From a differential geometry point of view, compression and regression correspond to devising an atlas on a manifold (or to charting a manifold). The presentation of a Bayesian projection procedure, mapping new (i.e., incoming) data of original space to the reduced space, concludes this section. An approach close to the present one was proposed by Elgammal and Lee [17] to infer body posefrom silhouettes.

2.1. Learning stage I: compression

Determination of the set $\hat{\mathbf{X}} \triangleq \{\hat{\mathbf{x}}_i,\ i = 1\ \cdots\ I\}$ of reduced coordinates is based on the hypothesis that the mapping f preserves distances, such that for all $(i,\ j)\ d_1(i,\ j) \simeq d_2(\hat{i},\hat{j})$ where d_1 is the (geodesic) distance in the original space and d_2 the (Euclidean) distance in the reduced space. This hypothesis allows to import the structure of the original data in the reduced space, hence bringing clarity in visualization and robustness in classification. The mapping is determined up to an isometry which, as already stated, has no influence on the final goals of the global procedure. The computation of approximate geodesic distances is detailed in [2]. In practice, $\mathrm{x}_i \in \mathbb{R}^{N_x}$, with typically $N_x \leq 4$, whereas $\mathrm{y}_i \in \mathbb{R}^{N_y}$, where N_y may take values in the tens of thousands. The last coordinate of each x_i will be 1, so that affine regression is handled with simplicity and clarity. This coordinate will not be considered when accounting for the dimension of the reduced space.

The $\hat{\mathbf{x}}_i$'s may be obtained as the solution of the following optimization problem:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \mathcal{S}(\mathbf{Y}, \mathbf{X}), \tag{1}$$

where $\mathbf{Y} \triangleq \{\mathbf{y}_i,\ i = 1\ \cdots\ I\}$ and $\mathcal{S}$ is acost (or stress) function. We will consider several stress functions in this work.

(i) $\mathcal{S}_{\mathrm{SAM}}$, proposed by Sammon [6] and advocated by Duda, et al. [18, chapter 10, (109)]:

$$\mathcal{S}_{\mathrm{SAM}}(\mathbf{Y}, \mathbf{X}) = \frac{1}{\sum_{i,j;\, i<j} d_1(\mathbf{y}_i, \mathbf{y}_j)} \sum_{i,j;\, i<j} \frac{\left[d_1(\mathbf{y}_i, \mathbf{y}_j) - d_2(\mathbf{x}_i, \mathbf{x}_j)\right]^2}{d_1(\mathbf{y}_i, \mathbf{y}_j)}, \tag{2}$$

where d_1 is geodesic (in [6], d_1 is supposed to be Euclidean though it is mentioned that any distance could be used) and d_2 Euclidean.

(ii) $\mathcal{S}_{\mathrm{ISO}}$, corresponding to the Isomap [2] algorithm

$$\mathcal{S}_{\mathrm{ISO}}(\mathbf{Y}, \mathbf{X}) = \left|\left|\mathbf{J}\left(\mathbf{D}_{\mathbf{Y}}^2 - \mathbf{D}_{\mathbf{X}}^2\right)\mathbf{J}\right|\right|_F^2, \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{J} = \mathbf{Id} - (1/I)\mathbb{1}\mathbb{1}^t$ is the centering matrix, with $\mathbb{1} = [1\ \cdots\ 1]^t$. Matrix $\mathbf{D}_{\mathbf{X}}^2$ (resp., $\mathbf{D}_{\mathbf{Y}}^2$) encompasses the squared d_2 (resp., d_1) distances between the $\mathbf{x}_i, \mathbf{x}_j$ (resp., $\mathbf{y}_i, \mathbf{y}_j$). Distances d_1 and d_2 are, respectively, geodesic and Euclidean.

In the standard MDS approach, $d_1(\mathbf{y}_i, \mathbf{y}_j)$ is replaced by a monotonic transformation of p_{ij}, where p_{ij} is the proximity (to be user-defined according to the type of data considered) between $\mathbf{y}_i$ and $\mathbf{y}_j$ (see [4]),

(iii) $\mathcal{S}_{\text{CCA}}$, corresponding to the Curvilinear Component Analysis approach proposed by Demartines et al. [19]:

$$\mathcal{S}_{\text{CCA}}(\mathbf{Y}, \mathbf{X}) = \sum_{i,j;\, i<j} \left[d_1(\mathbf{y}_i, \mathbf{y}_j) - d_2(\mathbf{x}_i, \mathbf{x}_j)\right]^2 F(d_2(\mathbf{x}_i, \mathbf{x}_j)), \tag{4}$$

where $F(x) = \mathbb{1}_{(|x|\leq\lambda)}$, $\mathbb{1}_{(\cdot)}$ being the indicator function taking value 1 (resp., 0) if the condition between parentheses is true (resp., false).

(iv) $\mathcal{S}_{\text{P}}$, which is the stress function we propose:

$$\mathcal{S}_{\text{P}}(\mathbf{Y}, \mathbf{X}) = \sum_{i,j;\, i<j} \sqrt{\gamma + \left[d_1(\mathbf{y}_i, \mathbf{y}_j) - d_2(\mathbf{x}_i, \mathbf{x}_j)\right]^2} \times \sqrt{\frac{\tau^2 + \left[d_1(\mathbf{y}_i, \mathbf{y}_j) - d_2(\mathbf{x}_i, \mathbf{x}_j)\right]^2}{\tau^2}} \frac{d_1(\mathbf{y}_i, \mathbf{y}_j)}{\sigma + d_1(\mathbf{y}_i, \mathbf{y}_j)}, \tag{5}$$

where d_1 is geodesic and d_2 Euclidean.

PCA can be interpretedas a particular case of this approach (Isomap with Euclidean distances in both spaces, see [20]). Standard MDS, which uses squared differences between distances, is sensitive to outliers, which can be present in real data. This advocates the metric MDS approach proposed here (involving a robust, nonquadratic cost function, see the first factor in the definition of $\mathcal{S}_{\text{P}}$). Besides, small distances can be dominated by noise. Small distances thus should have reduced influence on the stress. This is achieved by the third factor in the definition of $\mathcal{S}_{\text{P}}$. Finally, we would like to enhance the convexity of the cost function, which is the role of the second factor in the definition of $\mathcal{S}_{\text{P}}$ (in the transient phase of the optimization, when $d_1 - d_2$ is large, the corresponding term in the sum is quadratic). This eliminates many local minima. More precisely, γ is a small real number so that the square root is always differentiable, τ^2 is a threshold indicating whether the cost function is linear or quadratic; and σ is a threshold indicating whether $d_1(\mathbf{y}_i - \mathbf{y}_j) - d_2(\mathbf{x}_i - \mathbf{x}_j)$ should be considered or not (if $d_1(\mathbf{y}_i - \mathbf{y}_j) \ll \sigma$, $d_1/(\sigma + d_1)$ vanishes and $d_1 - d_2$ has no influence on the stress). We set $\gamma = 10^{-7}$, we chose σ as the first percentile of the distances d_1 and τ as the 80th percentile of the distances d_1.

This cost function is highly multimodal and must therefore be handled with care. Two different stochastic approaches are proposed to solve the optimization problem.

The first approach (Algorithm 1) considers all points $\mathbf{x}_i$ simultaneously. We use gradient descent with an exact line search. The algorithm proceeds iteratively and adds a small amount of noise to the points $\mathbf{x}_i$ after each descent step to try to avoid local minima. The $\mathbf{x}_i$'s are centered after each descent step to avoid drifting and to maximize precision. The

Input: original coordinates ($\mathbf{y}_i$, $i \in 1, \cdots I$)
Output: reduced coordinates ($\hat{\mathbf{x}}_i$, $i \in 1, \cdots I$)
begin
  Initialize randomly the reduced coordinates;
  repeat
    Add a small amount of noise to each $\mathbf{x}_i$;
    Optimize wrt all $\mathbf{x}_i$ simultaneously (gradient descent);
    Center the $\mathbf{x}_i$;
  until *convergence*;
end

Algorithm 1: Standard compression algorithm.

second approach (Algorithm 2) proceeds by incorporating successively the $\mathbf{x}_i$'s. The rationale is that a cost function encompassing few points will be less multimodal, thus enabling to reach the global optimum without particular care and computational effort. Incorporating additional points will have negligible effect on the points already optimized and will (hopefully, at least if only one point is incorporated at a time) correspond to a monomodal problem. As for the first algorithm, gradient descent and centering are used.

The solution of the Isomap problem may be obtained analytically and thus does not involve any iterative optimization [20]. Sammon [6] proposed gradient descent with a fixed step size for the stress function $\mathcal{S}_{\text{SAM}}$. Other approaches include optimizations on local distances [21], as opposed to approximated geodesic distances, but this can create folding in the reduced space as may also be observed for LLE [3].

Let us mention that the optimization algorithms approaches mentioned in the literature proved to be inefficient for the data sets we considered. Handling the intricate optimization problem related to compression may thus be considered as a contribution of the present work. In addition to those approaches based on the optimization of a stress function, another class of approaches, known as eigenmap methods, has recently gained interest. Generally speaking, eigenmap methods rely on the computation of eigenvectors and amount indirectly to the optimization of a cost. Some of the eigenvectors are used as a compact representation of the data. This general class encompasses Belkin and Niyogi's Laplacian eigenmaps [8], Donoho and Grime's Hessian eigenmaps [9], and Coifman and Lafon's diffusion maps [10]. Those methods, which attempt to preserve either local or global structure, will not be detailed here, the reader being referred to the cited references.

Let us finally mention that unifying interpretations of several methods have been proposed: Ham et al. propose an interpretation from a kernel point of view [22], Coifman and Lafon propose an interpretation from a diffusion map point of view [10].

2.2. Learning stage II: regression

The mapping $\mathbf{f}$ can be estimated, now that the reduced set of coordinates is known. We choose for $\mathbf{f}$ a piecewise affine function, because affine functions are highly adaptable

and relatively easy to handle. PCA would correspond to a piecewise affine function of only one piece.

The general goal is to estimate a set $\mathbf{W} \triangleq \{\mathbf{W}_k,\ k = 1 \cdots K\}$ of $N_y \times N_x$ regression matrices, where K is the unknown number of pieces of the model approximating the manifold optimally and where N_x (resp., N_y) is the dimension of the compressed (resp., response) variable. The guideline will be the minimization of $\sum_i \|\mathbf{y}_i - \mathbf{W}_l \mathbf{x}_i\|^2$. The index l involved is unknown and depends on i. Such an optimization problem is intricate and we will propose a suboptimal solution, but holding satisfying approximation properties. Let us mention that we enforce a connectivity constraint: no vector $\mathbf{x}_i$ can have a label not represented in its neighborhood and two or more patches with the same label are not authorized. The proposed formulation of the regression problem has a stochastic interpretation as maximum likelihood estimation with Gaussian iid noise, subject to the connectivity constraint.

Let us also mention that nonlinear regression is usually addressed in the literature in the scalar case ($\mathbf{y}_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^{N_x}$) or in the case of the juxtaposition of independent scalar regressions (e.g., see the projection pursuit regression method of Friedman and Stuetzle [23] and the multivariate adaptive regression splines method of Friedman [24]). The reader is also referred to themonograph [25]. More generally, all monographs we are aware of are dealing with regression and smoothing methods in the scalar response case. In the multidimensional response case, our approach is close to the one proposed by Haralick and Harpaz [26], where search for features linear cluster by linear cluster is carried out. In Haralick and Harpaz's approach, no dimension reduction is considered, thus leaving the clusters unconnected, which is a drawback for future classification, visualization, data understanding, and the computation of means. *To our best knowledge, no other work has addressed the general case* $\mathbf{y}_i \in \mathbb{R}^{N_y}$, $N_y > 1$, set aside principal surfaces approaches which we already mentioned.

Two approaches, both stochastic and iterative, are proposed to estimate the $\mathbf{W}_k$'s and the l's. The first one (Algorithm 3) creates a plane in an unlabeled neighborhood of a randomly chosen point. It must be noticed that some points may remain unlabeled at the end of this procedure. We choose not to process them since there are few such points and since their influence is negligible. The second one (Algorithm 4) creates planes considering one point among those which are the most unlikely to belong to an already existing plane (this is done by considering the likelihoods of the point and of its neighborhood). One advantage of the latter algorithm is the control over the quality of the regression allowed by the information criterion. This criterion may also be used to impose the number of pieces of the model or the accuracy of the regression.

2.3. Projection

Projecting a new (i.e., incoming) point $\mathbf{y}$ onto the manifold amounts to estimating the variables $(\hat{\mathbf{x}}, \hat{\boldsymbol{\varepsilon}}, \hat{l})$ parameterizing this point. We will thus have $\mathbf{y} = \mathbf{W}_{\hat{l}}\hat{\mathbf{x}} + \hat{\boldsymbol{\varepsilon}}$. The general goal is to assess the belonging of point $\mathbf{y}$ to the manifold.

This problem will be cast in a Bayesian framework. The unknowns $\hat{\mathbf{x}}$ and $\hat{l}$ ($\hat{\boldsymbol{\varepsilon}}$ is implicitly known, when $\hat{\mathbf{x}}$ and $\hat{l}$ are given) will be chosen as the maximizers of $p(\mathbf{x}, l \mid \mathbf{y})$. All probabilities involved are conditioned on the model previously determined. This conditioning is dropped for the sake of clarity. Maximizing $p(\mathbf{x}, l|\mathbf{y})$ amounts to maximizing $p(\mathbf{y}|\mathbf{x}, l) \cdot p(\mathbf{x}|l) \cdot p(l)$, which amounts to minimizing

$$\frac{1}{2\sigma_\varepsilon^2}\|\mathbf{y} - \mathbf{W}_l\mathbf{x}\|^2 - \log p(\mathbf{x} \mid l) - \log p(l), \tag{6}$$

where σ_ε^2 is the variance of the noise as estimated from the initial (i.e., learning) data set. Probability $p(\mathbf{x} \mid l)$ is determined from the learning data set using Gaussian kernels. Probability $p(l)$ is also estimated from the learning data set.

For $l \in [1, K]$, $\hat{\mathbf{x}}_l$ is computed using a local optimization algorithm with a multistart process, since the cost function may be multimodal. The estimate $\hat{\mathbf{x}}$ is retained as the minimizer of the cost function among all $\hat{\mathbf{x}}_l$'s, whose indices l satisfy the connectivity constraint.

It must benoted that the original data $\mathbf{y}_i$ are not needed for the projection, since only the reduced coordinates $\mathbf{x}_i$ and the matrices $\mathbf{W}_k$ are used in the process. This saves a lot of memory, which is an advantage of the proposed methodology.

2.4. Comment: iterating compression and regression

As an extension to the previous approach, one might also consider iterating compression and regression steps. The rationale is to use the result of the regression step to initialize the neighborhood graph: in addition to its use for the nearest neighbors, Euclidean distance is then used for all pair of points belonging to the same linear piece, thus reducing the errors that might occur in the estimation of those distances.

We implemented this approach on the SCurve. The estimation of the reduced coordinates was not improved. This is due to the fact that the regression step is an approximation to the true function if few planes are used. Thus, errors are introduced in the regression step, which are further propagated in the compression step. If many planes are used, the new initialization of the graph is not different from the one considering the nearest neighbors. Hence, there is no global advantage in iterating between both steps, at least for the SCurve, and it is highly probable that this would also be the case for other data sets. This approach—iterating compression and regression—will not be considered in the sequel.

3. EXPERIMENTAL RESULTS

This section is separatedin two parts. The first part presents results of the compression procedure; the second one addresses the projection procedure. The manifolds used here are standard data sets known as the SwissRoll (Figure 1(a)), the SCurve (Figure 1(c)) (e.g., see [2, 3, 13]) and the COIL-20 database (e.g., see Figure 6, [27]). We mention that both the SwissRoll and the SCurve have two degrees of freedom

Input: original coordinates ($\mathbf{y}_i, \quad i \in 1, \quad \cdots \quad I$)
Output: reduced coordinates ($\hat{\mathbf{x}}_i, \quad i \in 1, \quad \cdots \quad I$)
begin
Consider a random subset of (typically 10) points and optimize using algorithm 1;
repeat
Incorporate a given number of points (default is one, initial reduced coordinates are random);
Optimize wrt the points just incorporated (gradient descent);
Optimize wrt all points simultaneously (gradient descent);
Center the $\mathbf{x}_i$;
until *all points incorporated*;
end

Algorithm 2: Successive incorporation compression algorithm.

Input: original and reduced coordinates ($\mathbf{y}_i, i \in 1, \ldots I$ and $\hat{\mathbf{x}}_i, i \in 1, \ldots I$)
Output: labels (l), regression matrices ($\mathbf{W}_k$) and noise variance (variance of the $\hat{\boldsymbol{\varepsilon}}_i$)
begin
while *exists a point whose neighbors are not labeled* **do**
Pick randomly a point whose neighbors are not labeled;
Compute matrix $\mathbf{W}_k$ regressing this neighborhood (1);
Update labels, matrices and noise variance (2);
Discard any piece of the model (label and matrix) having not enough points (3);
end
Update noise variance.
end
Comments:

(1) Let $\mathbf{Y}_i$ be the matrix encompassing $\mathbf{y}_i$ and all neighbors $\mathbf{y}_j$ of $\mathbf{y}_i$ and let $\mathbf{X}_i$ be its counterpart. Matrix $\mathbf{W}_k$ is estimated from equation $\mathbf{Y}_i \simeq \mathbf{W}_k \mathbf{X}_i$ by least squares.

(2) A point is assigned to a linear model if the norm of the reconstruction error is less than a given factor times the standard deviation associated to the overall model (the noise is supposed to be iid). Moreover, connectivity must be preserved (i.e., a point can be assigned a label only if this label is represented in the point's neighborhood). Once all updates are completed, the variance of the overall model is re-estimated.

(3) If a linear piece has not enough points, the matrix that describes it cannot be computed. In this case, this piece is discarded.

Algorithm 3: Piecewise linear mapping 1 (PLM 1).

and that neither of them can be described by a 2-dimensional reduced space in a linear framework in a manner that preserves their intrinsic structure. This precludes PCA for a compact representation of these data.

3.1. Compression

First of all, it should be mentioned that comparing compression methods is an intricate problem because of the choice of the evaluation criterion which should not favor one method over the others. As benchmarks, we consider the SwissRoll and the SCurve which were compressed using different paradigms and stress functions (PCA, Isomap, $\mathcal{S}_{\mathrm{P}}$, $\mathcal{S}_{\mathrm{SAM}}$, $\mathcal{S}_{\mathrm{CCA}}$, Laplacian eigenmaps, diffusion maps, Hessian eigenmaps, and locally linear embedding). We consider a reduced space of dimension 2, which is the true intrinsic dimensionality of the data (we will return on the determination of the intrinsic dimensionality in the sequel).

From a qualitative point of view, we observe that PCA does not behave well, since the neighborhood structure is not

Input: original and reduced coordinates ($\mathbf{y}_i, i \in 1, \ldots I$ and $\hat{\mathbf{x}}_i, i \in 1, \ldots I$)
Output: labels (l), regression matrices ($\mathbf{W}_k$) and noise variance (variance of the $\hat{\varepsilon}_i$)
begin
  Create a single matrix encompassing the entire data set;
  repeat
    Determine the points which are the most unlikely to belong to the pieces (matrices) already created;
    Pick one point among the previously determined set;
    Compute matrix $\mathbf{W}_k$ regressing the neighborhood of the point considered;
    Label the points of the neighborhood as belonging to this new piece;
    repeat
      Discard any piece of the model (label and matrix) having not enough points;
      Update matrices and labels;
      Label every unassigned point with the most likely piece;
    until *convergence*;
  until *the information criterion is minimized*;
end

Algorithm 4: Piecewise linear mapping 2 (PLM 2).

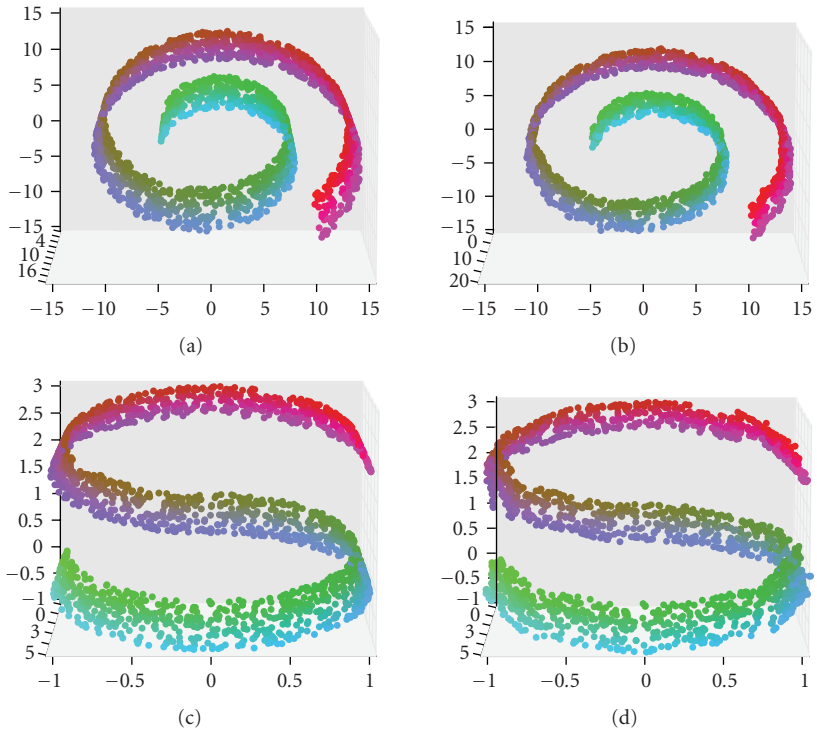

Figure 1: (a) Original SwissRoll, (b) regressed SwissRoll (108 planes), (c) original SCurve, (d) regressed SCurve (28 planes).

preserved in the reduced space (see Figure 2 and compare the colors of the reduced data to the colors of the original data displayed Figure 1(a)) (see also Figure 4). Locally linear embedding, Laplacian eigenmaps, and diffusion maps do not behave satisfyingly either. All other approaches behave quite well in the nonnoisy SCurve case. In the noisy case, as expected, all methods are affected by noise, though to a different extent (see Figure 3). Besides all not very satisfying behaviors mentioned above, we notice that the Hessian eigenmap compression degrades in the presence of noise, as mentioned in the literature [9]. The Laplacian eigenmap compression tends to introduce holes, which correspond to

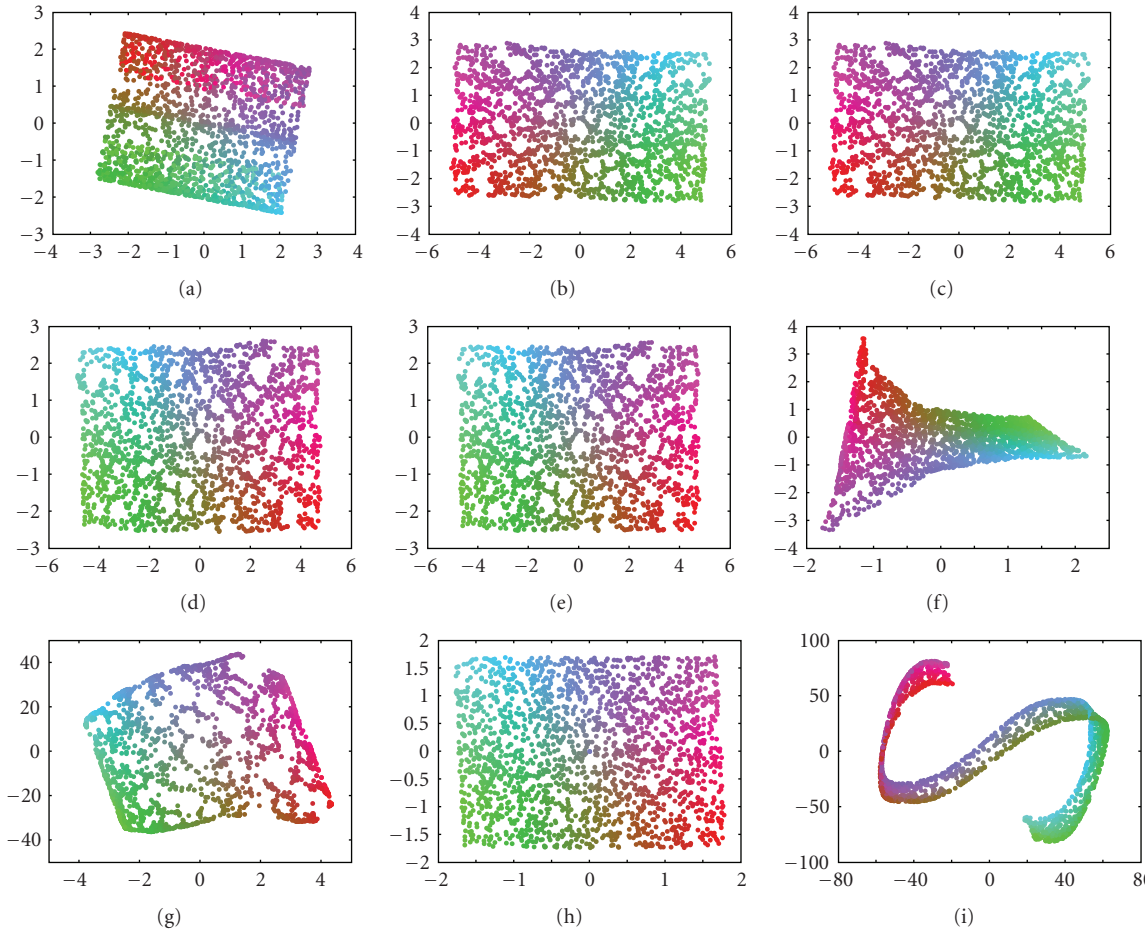

FIGURE 2: Reduced space computed by (a) PCA, (b) Isomap, (c) $\mathcal{S}_P$, (d) $\mathcal{S}_{SAM}$, (e) $\mathcal{S}_{CCA}$, (f) locally linear embedding, (g) Laplacian eigenmap, (h) Hessian eigenmap, and (i) diffusion maps for the nonnoisy SCurve.

TABLE 1: Distances between the original reduced coordinates and the estimated ones for a 2000-point SCurve. For $\mathcal{S}_{CCA}$, the threshold was set so that 5% of the lowest distances d_2 were considered. LEM, DM, HEM, and LLE stand for Laplacian eigenmap, diffusion map, Hessian eigenmap, and locally linear embedding, respectively. In the cases of rows 3 and 4 (Gaussian and Laplacian noise), noise is affecting the coordinates of the $\mathbf{y}_i$'s. The noise variances (2% and 5%) are quantified with respect to the variance of the noise-free data. In case of row 5, the distances d_1 are perturbated with a Laplacian noise affecting 12.5% of the d_1's. The noise variance (200%) is quantified with respect to the variance of the noise-free $d_1 - d_2$'s.

Noise	PCA	Isomap	$\mathcal{S}_P$	$\mathcal{S}_{SAM}$	$\mathcal{S}_{CCA}$	LEM	DM	HEM	LLE
none	43.6	3.01	2.29	2.61	3.01	21.13	67.50	3.05	40.1
Gaussian noise 5%	43.6	8.55	2.94	2.60	6.22	23.51	67.76	18.57	90.2
Laplacian noise 2%	44.4	7.01	6.46	6.10	4.70	23.47	67.54	20.51	69.2
Impulsive perturbation	na	3.80	2.93	3.22	3.09	na	na	na	na

its known property of emphasizing clusters [8]. Sammon's cost function and the proposed cost function exhibit the most satisfying behaviors. Similar conclusions may be drawn with the SwissRoll test case. These differences should be quantified, which we did using two different tests.

The rationale of the first test is to try to recover the true (initial) $\mathbf{x}_i$'s (which will be denoted $\mathbf{x}_i^\star$'s) from the data $\mathbf{y}_i$'s. For this first test to make sense, we have to consider a manifold whose local magnification factor is 1 (the distances between the $\mathbf{y}_i$'s must be the same as the ones between the $\mathbf{x}_i^\star$'s since the distances between the $\mathbf{y}_i$'s are reproduced in the distances between the $\hat{\mathbf{x}}_i$'s). The SCurve was retained for this reason. Since the $\hat{\mathbf{x}}_i$'s are determined up to an isometry, the isometry putting the $\mathbf{x}_i^\star$'s and the $\hat{\mathbf{x}}_i$'s into correspondence is first computed and applied on the $\hat{\mathbf{x}}_i$'s. The sum of the squared distances between both sets of coordinates is then

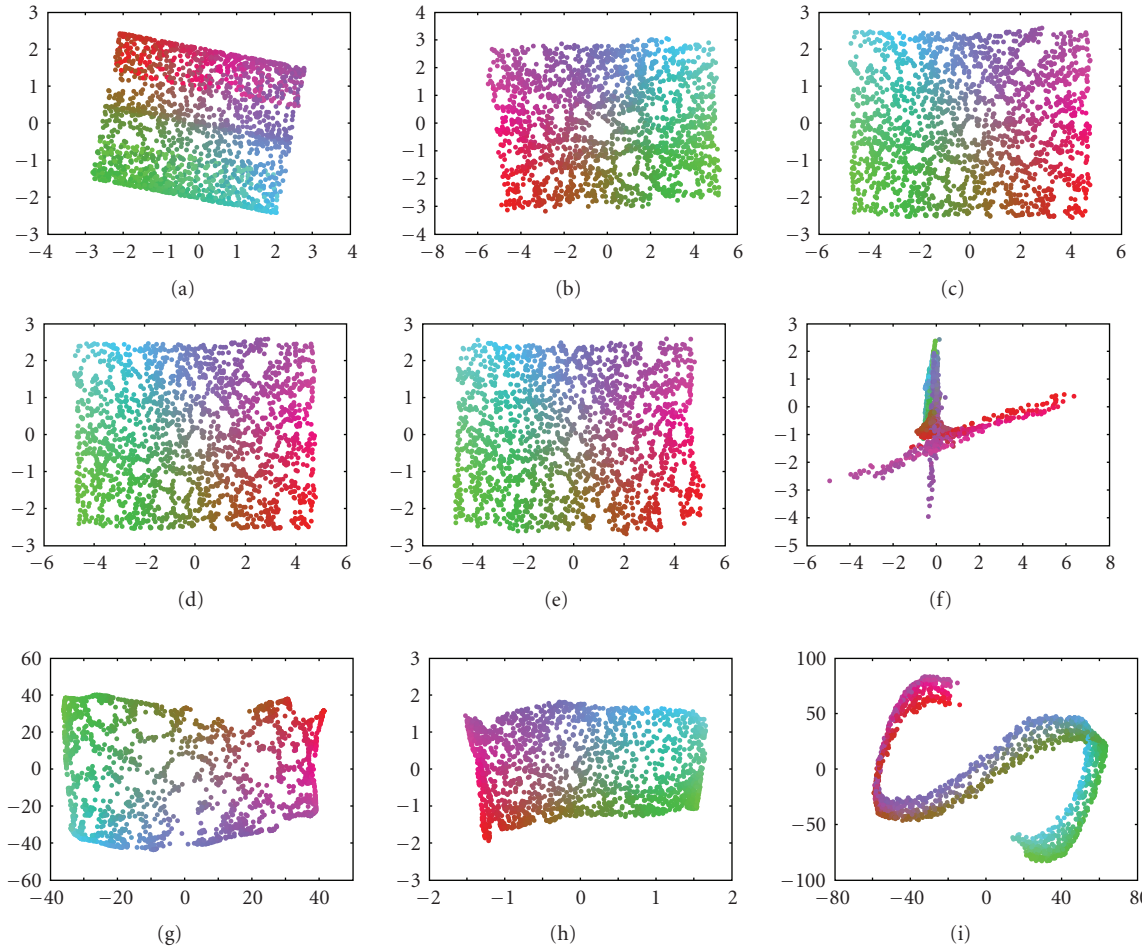

FIGURE 3: Reduced space computed by (a) PCA, (b) Isomap, (c) $\mathcal{S}_P$, (d) $\mathcal{S}_{SAM}$, (e) $\mathcal{S}_{CCA}$, (f) locally linear embedding, (g) Laplacian eigenmap, (h) Hessian eigenmap, and (i) diffusion maps for the noisy SCurve (Gaussian noise, 5% variance).

computed (notice that many other choices for the criterion quantifying the discrepancy between both sets may have been done. We chose this particular one because it is the most used). Examining Table 1, we may assert that $\mathcal{S}_{SAM}$ and $\mathcal{S}_P$ behave better, which is coherent with the qualitative observations. Besides, it should be mentioned that the cost function of Demartines and Hérault ($\mathcal{S}_{CCA}$) behaves quite well and exhibits good robustness to noise. As expected, $\mathcal{S}_P$ yields better results in the presence of outliers (see the last line of Table 1).

In the second test, we computed the linear correlation between the true geodesic distances and the Euclidean distances computed from the estimated reduced coordinates. This was achieved on the SwissRoll. We will not further comment this benchmark since it was not able to clearly discriminate between Isomap, $\mathcal{S}_{SAM}$, $\mathcal{S}_{CCA}$, and $\mathcal{S}_P$. Let us notice that other choices might have been made for the comparison criterion (linear correlation is in essence a quadratic criterion).

The determination of the optimal dimension of the reduced space is tackled by the scree test [28], as is classically done in the literature. Optimizations are achieved for several dimensions of the reduced space and the stress values at the optima are compared. The dimension retained is the value after which the stress level does not decrease significantly any more. This can be seen on Figure 5 for several data sets (the SwissRoll and two sets from the COIL-20 database). Different noise levels for the SwissRoll indicate that low noise level has no influence on the number of dimensions. In each case, the optimal dimension is 2.

To conclude this section, we comment on both proposed algorithms. Algorithm 1 is less demanding from a CPU-time point of view, but requires several runs since it may get stuck at local minima. The computational burden of Algorithm 2 is larger, but the correct solutions were obtained from single runs on the data we processed. A 500-point SwissRoll requires 10-minute CPU time with Algorithm 1 (one run), 20 minutes with Algorithm 2, 11 seconds with Isomap. A 2000-point SwissRoll requires 75 minutes with Algorithm 1 (one run), 30 hours with Algorithm 2, between 30 seconds and 5 minutes with eigenmaps methods, 12 minutes with Isomap, 25 minutes with Sammon's cost function, and

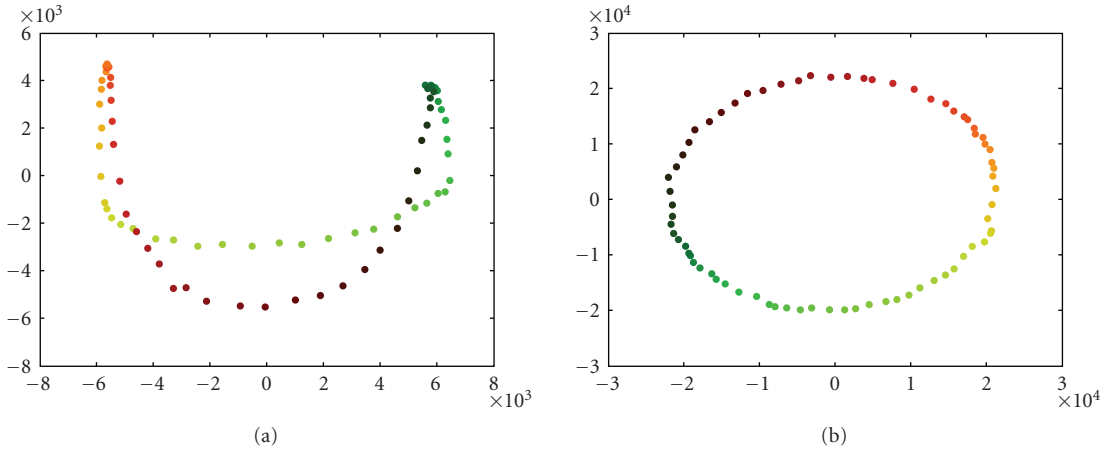

FIGURE 4: Reduced space computed by (a) PCA and (b) $\mathcal{S}_P$ for the duck of the COIL-20 database.

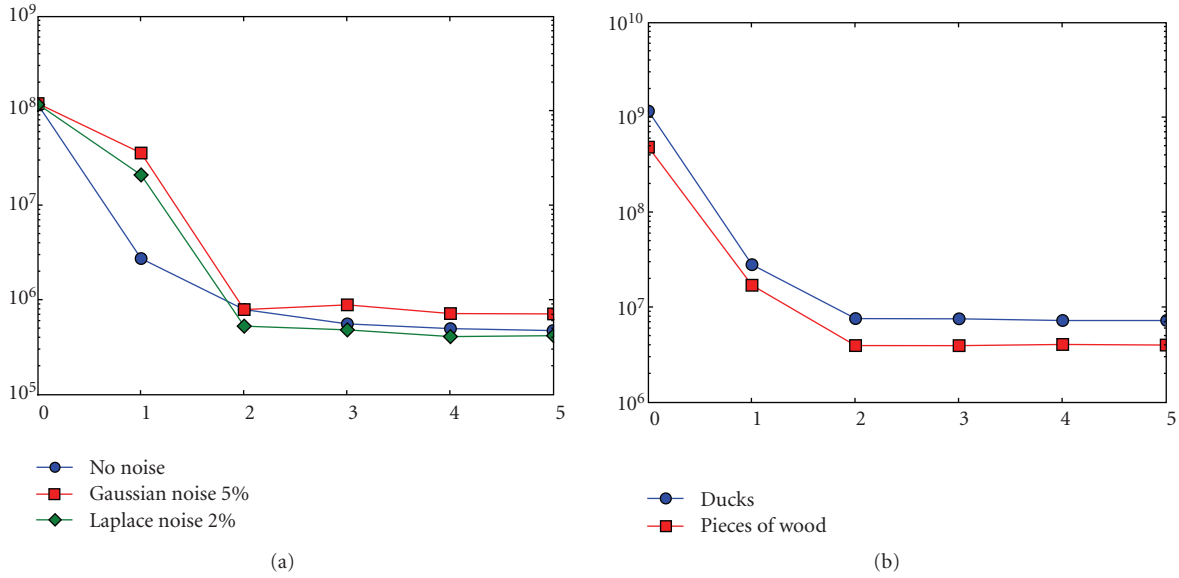

FIGURE 5: (a) Scree plots for $\mathcal{S}_P$ on the SwissRoll, (b) the duck and the piece of wood from the COIL-20 database. The abscissa axis is the dimension of the reduced space; and the ordinate axis is the value of the cost function at the solution.

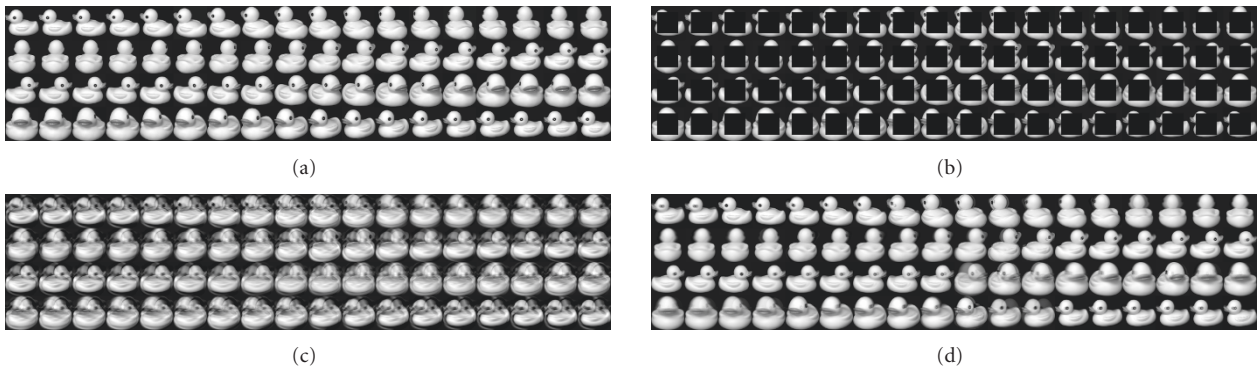

FIGURE 6: The 72 images of the duck set of the COIL-20 database with (a) 0% and (b) 40% occlusion; (c) reconstruction from PCA-15 compression, (d) reconstruction from $\mathcal{S}_P$-2 compression. The occlusion percentage is measured with respect to the entire image and would be larger if measured with respect to the foreground.

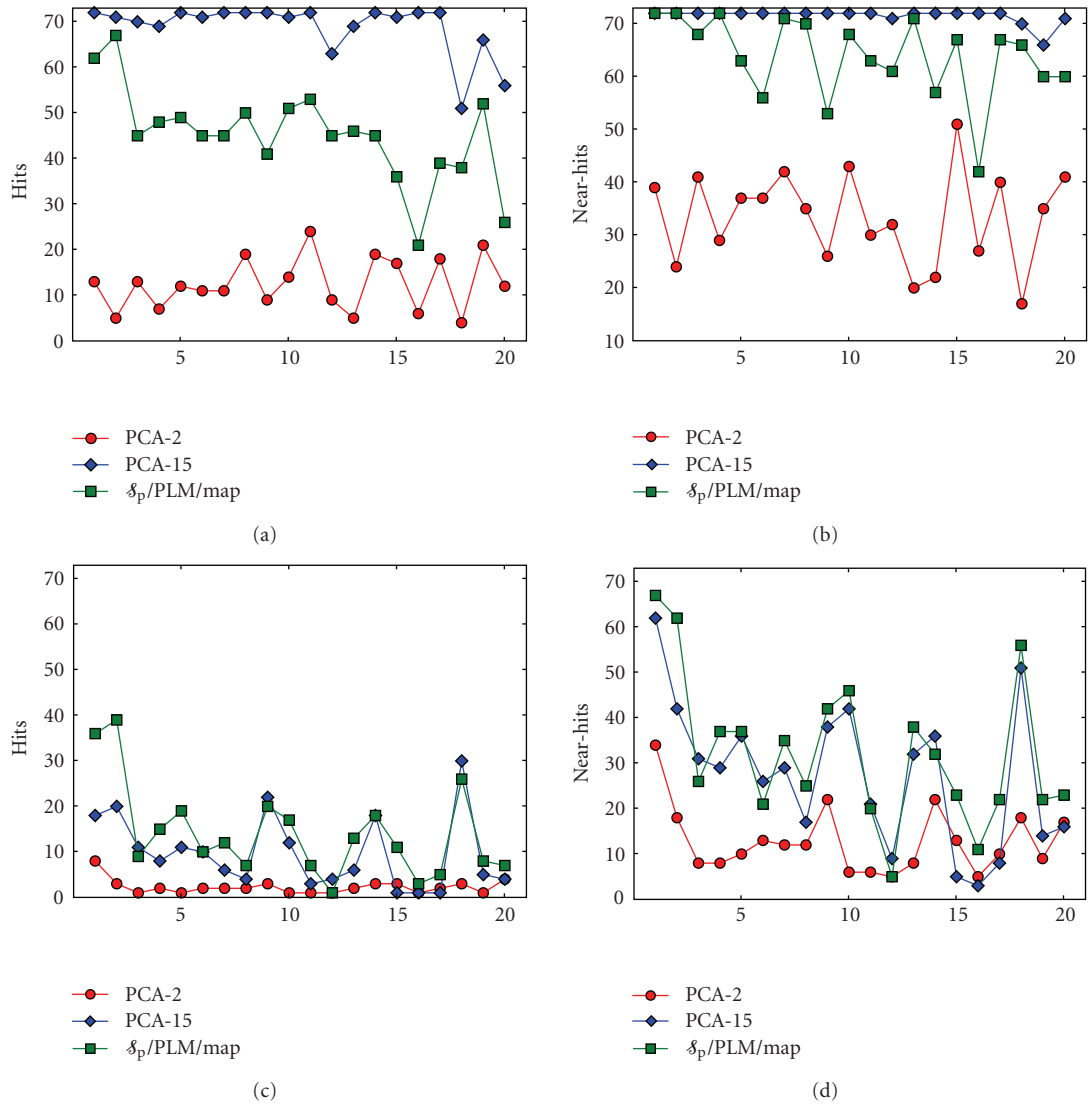

Figure 7: Projection of the sets of the COIL-20 database (abscissa represents the index of the data set, ranging from 1 to 20). The curves represent (a) hits with 0% occlusion, (b) near hits with 0% occlusion, (c) hits with 40% occlusion, (d) near hits with 40% occlusion for different paradigms (PCA and 2-dimensional $\mathcal{S}_P$).

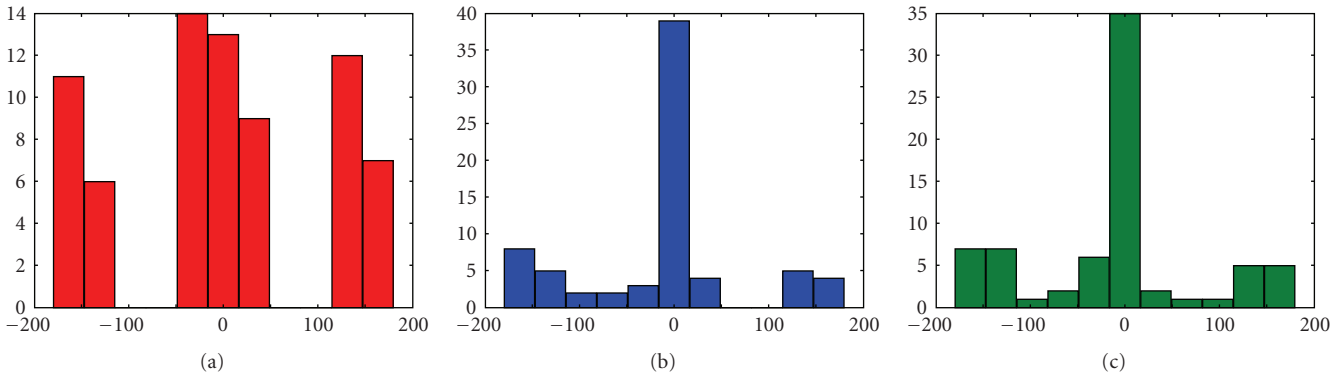

Figure 8: Histograms of the pose estimation error in degrees of one representative example from the COIL-20 database (40% occlusion): (a) PCA 2, (b) PCA 15, (c) 2-dimensional $\mathcal{S}_P$.

48 hours with $\mathcal{S}_{\text{CCA}}$. Nevertheless, though quite demanding, a compression has to be achieved only once for a given data set. For data sets with a very large number of points, other approaches known as landmark approaches, where compression is done with a subset of the initial data set, might be more adapted. The remaining data are then compressed with respect to the fixed landmarks. The regression phase requires from about 3- to 4-hour CPU time.

3.2. Projection

Two test casesare considered to evaluate the projection method (and in fact the whole procedure, since the projection method cannot be isolated).

The first test case is the computation of the projection error for the nonnoisy SwissRoll and SCurve. Points $\mathbf{y}_i$ are projected onto the regressed manifold. The average error is less than 1%, when measured with respect to the standard deviation of the $\mathbf{y}_j$'s.

The second test case is conducted by projecting occluded COIL-20 images (see Figure 6) onto each underlying manifold (i.e., a model is learnt for each data set). In the learning phase, all images (nonoccluded) of a given set are considered. Regression is achieved using both algorithms (PLM1 and PLM2), the best solution being retained.

The goal is to recover, for each occluded image, the original image. A hit is obtained when the image closest to the projection is the nonoccluded image. A near hit is obtained when the image closest to the projection is one of the five images closest to the original one. This assesses the robustness of the procedure. Because of the occlusion, the Gaussian noise assumption in the projection (see (6)) is replaced by a generalized Gaussian noise, with exponential rate of decay of 1.2. The results are displayed in Figure 7. It may be observed that the proposed approach yields better results than PCA-2. PCA-15 yields better results than the proposed approach, particularly in the case displayed in Figure 7(a), but at the cost of a significant increase of the dimension of the underlying reduced space. Even if PCA-15 behaves well from a hit or miss point of view, the distance in the reduced space is very large (see the reconstructed ducks, Figure 6(c)), much larger than the distance yielded by the proposed approach (Figure 6(d)). As a complement to this hit and miss point of view, we compute the pose estimation errors (see Figure 8). The analyses of those errors are consistent with the analysis corresponding to the hit and miss framework.

Moreover,we emphasize once more that the true degrees of freedom are not captured by PCA-15, thus hampering visualization and data understanding. We remind the reader that the computation of means is an important goal of the present work, with applications to brain imaging. The Fréchet means [29, chapter 9], [30] will be computed in the reduced space and then lifted into the original space. It is thus important to capture the true degrees of freedom of the data in order to avoid spurious effects induced by extra coordinates. We will thus retain the proposed $\mathcal{S}_{\text{P}}$ compression approach, along with the regression and Bayesian projection methods detailed in this article. Finally, let us mention that the computational cost of the projection phase is negligible (from 1 to 10 seconds).

4. CONCLUSION

In this article, we have introduced an original metric multidimensional scaling-based nonlinear manifold learning framework, allowing efficient and robust reduction of high-dimensional data. The approach is composed of compression, regression, and projection. The original data do not need to be stored hence saving significant memory space once the model is learnt.

Data understanding, visualization, classification, and the computation of means are possible even in the case of highly nonlinear manifolds. The classical solution to the general problem addressed here is PCA, which is clearly outperformed by the proposed method. Application of this method to analysis and classification of shapes in brain imaging is currently investigated.

ACKNOWLEDGMENT

The authors are grateful to the reviewers and associate editor for their questions and comments which lead to a significant improvement of the paper.

REFERENCES

[1] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[2] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[4] I. Borg, *Modern Multidimensional Scaling*, Springer, New York, NY, USA, 2nd edition, 2005.

[5] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, p. 7, 2002.

[6] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, no. 5, pp. 401–409, 1969.

[7] A. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, New York, NY, USA, 2002.

[8] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[9] D. Donoho and C. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 10, pp. 5591–5596, 2003.

[10] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[11] S. Roweis, L. Saul, and G. Hinton, "Global coordination of local linear models," in *Proceedings of the 14th Annual Conference on Advances in Neural Information Processing Systems (NIPS '01)*, vol. 14, pp. 889–896, MIT Press, Vancouver, BC, Canada, December 2001.

[12] M. Brand, "Charting a manifold," in *Proceedings of the 15th Annual Conference on Advances in Neural Information Processing Systems (NIPS '02)*, vol. 15, pp. 985–992, MIT Press, Vancouver, BC, Canada, December 2002.

[13] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[14] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[15] B. Kégl, A. Krzyzak, T. Linder, and K. Zeger, "Learning and design of principal curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 281–297, 2000.

[16] M. LeBlanc and R. Tibshirani, "Adaptive principal surfaces," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 53–64, 1994.

[17] A. Elgammal and C.-S. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 681–688, Washington, DC, USA, June-July 2004.

[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[19] P. Demartines and J. Hérault, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 148–154, 1997.

[20] T. Cox and M. Cox, *Multidimensional Scaling*, Chapman & Hall, Boca Raton, Fla, USA, 2001.

[21] D. K. Agrafiotis and H. Xu, "A self-organizing principle for learning nonlinear manifolds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 25, pp. 15869–15872, 2002.

[22] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, pp. 369–376, Banff, Canada, July 2004.

[23] J. Friedman and W. Stuetzle, "Projection pursuit regression," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.

[24] J. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, Berlin, Germany, 2003.

[26] R. Haralick and R. Harpaz, "Linear manifold clustering in high dimensional spaces by stochastic search," *Pattern Recognition*, vol. 40, no. 10, pp. 2672–2684, 2007.

[27] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (coil-20)," Tech. Rep. CUCS-005-96, Deparment of Computer Science, Columbia University, New York, NY, USA, 1996.

[28] R. Cattell, "The scree test for the number of factors," *Multivariate Behaviorial Research*, vol. 1, no. 2, pp. 245–276, 1966.

[29] D. Kendall, D. Barden, T. Carne, and H. Le, *Shape and Shape Theory*, John Wiley & Sons, New York, NY, USA, 1999.

[30] A. Kume and H. Le, "On Fréchet means in simplex shape spaces," *Advances in Applied Probability*, vol. 35, no. 4, pp. 885–897, 2003.