UNSUPERVISED NONLINEAR MANIFOLD LEARNING

Matthieu Brucher, Christian Heinrich, Fabrice Heitz

Laboratoire des Sciences de l'Image,
de l'Informatique et de la Télédétection,
Université Louis Pasteur, Strasbourg 1, France
(LSIIT, UMR 7005, CNRS-ULP)

Jean-Paul Armspach

Laboratoire de Neuro-Imagerie in Vivo,
Université Louis Pasteur, Strasbourg 1, France
(LNV, UMR 7004, CNRS-ULP)

ABSTRACT

This communication deals with data reduction and regression. A set of high dimensional data (*e.g.*, images) usually has only a few degrees of freedom with corresponding variables that are used to parameterize the original data set. Data understanding, visualization and classification are the usual goals.

The proposed method reduces data considering a unique set of low-dimensional variables and a user-defined cost function in the multidimensional scaling framework. Mapping of the reduced variables to the original data is also addressed, which is another contribution of this work. Typical data reduction methods, such as Isomap or LLE, do not deal with this important aspect of manifold learning. We also tackle the inversion of the mapping, which makes it possible to project high-dimensional noisy points onto the manifold, like PCA with linear models. We present an application of our approach to several standard data sets such as the SwissRoll.

Index Terms— Unsupervised learning, regression, data reduction, multidimensional scaling

1. INTRODUCTION

Data (or dimensionality) reduction consists in determining a set of reduced dimensionality variables capturing the major features of a given set of high dimensionality data, which in the present case are images. These original (noisy) data are assumed to lie close to a (nonlinear) manifold whose dimension is the dimension of the reduced variables. The reduced variables may be mapped onto the original data. We will call the succession of data reduction and mapping (regression) procedures *manifold learning*.

The usual goals of manifold learning are data understanding (the few degrees of freedom are given a physical interpretation, such as pose angle, giving a better understanding of the data generation process), visualization (a scatter plot of the reduced variables is displayed, where each point is labeled with the original data) and classification (classification is achieved in a more robust manner in the space of reduced variables). Application fields of data reduction and manifold learning are face and character recognition, shape analysis and target classification, to mention a few examples.

Principal Component Analysis (PCA) typically addresses the manifold learning issue, but its limitations are severe and well-known: only linear manifolds may be handled (for example, the SwissRoll [fig.1(a)], which is a nonlinear manifold, can obviously not be described by PCA), and data reduction and regression are achieved in a quadratic framework (*i.e.*, the noise is assumed to be Gaussian [1]). Moreover, it is also well-known that PCA does not always deal satisfyingly with classification problems. Our goal is to propose a method that is not restricted to linear manifolds, whose underlying cost function may be user-defined, and which is flexible enough to adapt to a large class of classification problems.

Let $\boldsymbol{y}$ be any element of the original data set. The goal is to determine the corresponding reduced vector $\boldsymbol{x}$, the mapping $\boldsymbol{f}$ and the noise $\boldsymbol{\varepsilon}$ such that $\boldsymbol{y} = \boldsymbol{f}\left(\boldsymbol{x}\right) + \boldsymbol{\varepsilon}$. An additional hypothesis is that the mapping $\boldsymbol{f}$ preserves distances (which will be defined precisely later). There is no unique solution to this problem, since any isometry $\boldsymbol{\mathcal{I}}$ will yield another solution in the form $\left(\boldsymbol{x}' = \boldsymbol{\mathcal{I}}\left(\boldsymbol{x}\right); \boldsymbol{f}' = \boldsymbol{f} \circ \boldsymbol{\mathcal{I}}^{-1}\right)$. This undetermination has no impact on the usual goals of data reduction (*i.e.*, data understanding, visualization, classification and the computation of means). The noise distribution is assumed to be known up to its parameter values. Determination of $\boldsymbol{x}$ and $\boldsymbol{f}$ will be achieved sequentially (we will estimate piecewise linear $\boldsymbol{f}$'s).

The most straightforward extension of PCA, with a view to handling nonlinear manifolds, is local PCA [2]. The original data are processed groupewise, each group yielding its own PCA-reduced variables. Though this approach is able to adapt to nonlinear manifolds, the question of the determination of the initial groups of variables remains. This determination should be able to adapt to the shape of the manifold. Moreover, the different sets of reduced variables are unrelated, and the method is cast in a quadratic framework. All these features obviously severely hamper the approach. The problem must clearly be addressed globally (*i.e.,* the data cannot be processed groupewise) so that classification may operate simultaneously on all data.

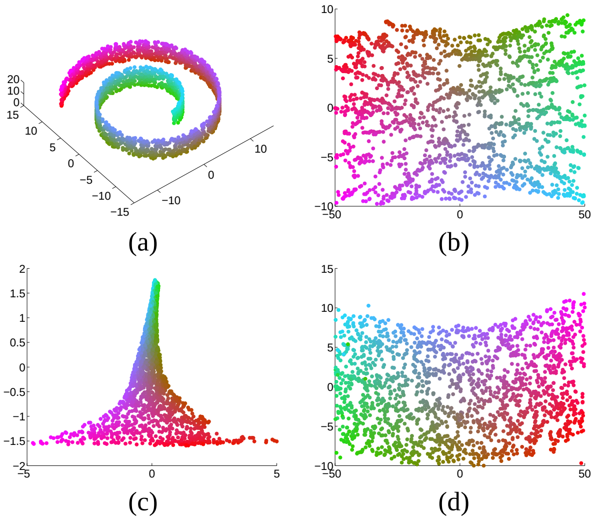

Fig. 1. SwissRoll example. (a) the original SwissRoll, (b) the reduction with Isomap, (c) the reduction with LLE, (d) the reduction with our cost function. A given point has the same color on all graphs.

Typical global compression techniques are Isomap [3] and LLE [4]. Isomap is a geodesic distance-based classical multidimensional scaling (MDS) [5], which is aimed at reproducing a given interpoint distance matrix in low dimensional space. It is to be considered as a global method since all distances, small and large, are reproduced simultaneously in the reduced space. This raises the question of the immunity to noise in the data, due to its quadratic nature [6]. LLE seeks to preserve the local barycentric coordinates, and is to be considered as a local method (it proceeds by trying to preserve the local geometric structure of points).

Several authors [7, 8] reduce the data groupewise and address the linking of local reduced variables in coordination to projection on the manifold. These approaches use a Gaussian probabilistic framework with a fixed number of local linear models and, as with our algorithm, do not require the original sample points after the learning step.

This communication is organized as follows. Our approach is described in section 2. Several application examples are given in section 3. finally the conclusion and future prospects to this work are presented in section 4.

2. DATA REDUCTION AND MANIFOLD LEARNING

In this section, we detail our method, both in its data reduction and manifold learning (regression) aspects.

2.1. Data reduction

The problem here is to determine the $\boldsymbol{x}_i$'s such that the reduced interpoint distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ match the corresponding data interpoint distances $d_{ij} \triangleq \|\boldsymbol{y}_i - \boldsymbol{y}_j\|$. Since the data $\boldsymbol{y}_i$ are supposed to lie close to a nonlinear manifold, we consider geodesic distances in the original space, following the Isomap algorithm. Euclidean distances are considered in the reduced space since the $\boldsymbol{x}_i$'s are supposed to fill this space.

Isomap estimates the $\boldsymbol{x}_i$'s by minimization of a quadratic stress function $\sum_{i,j} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - d_{ij}\right)^2$. We suggest estimating the $\boldsymbol{x}_i$'s by minimization of $\sum_{i,j} f\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$, where

$$f\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sqrt{\varepsilon + \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - d_{ij}\right)^2} \quad \frac{d_{ij}}{\sigma + d_{ij}}, \qquad (1)$$

which fits into the metric multidimensional scaling framework [9, section 9.4.2].

The first factor is a robust (nonquadratic) discrepancy measurement between the candidate distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ and the target distance d_{ij}, where ε is a small real number whose role is to ensure the differentiability of the cost function. We consider a robust discrepancy measurement since some geodesic distances could be estimated with significant error, which should have limited influence on the estimate. The second factor weighs the discrepancy, such that nearby points (corresponding to small d_{ij}'s) play no role in the estimate. The motivation is that small distances d_{ij} are highly contaminated by noise of variance σ^2, and hence are not reliable.

```
Input: original coordinates
Output: reduced coordinates
begin
    for each point on the manifold do
        Compute the (Euclidean) distances to its
        nearest neighbors
    end
    Estimate the geodesic distances using Floyd's
    algorithm;
    Initialize randomly the reduced coordinates;
    while not converged do
        Optimize all points simultaneously
    end
    while not converged do
        Optimize simultaneously some points drawn
        randomly
    end
end
```

Algorithm 1: data reduction algorithm

The optimization of this cost function (referred to as OCF in the following sections) is achieved by a standard damped gradient optimizer, where all points $\boldsymbol{x}_i$ are updated jointly. Unfortunately, the algorithm usually gets trapped in a local minimum. This is why an additional gradient step in which only some points can be moved jointly is needed (see algorithm 1).

2.2. Regression

The estimation of the mapping of the reduced variables to the original variables is an unsupervised nonlinear regression problem, which is quite intricate because of its dimensionality. To the best of our knowledge, this general regression problem (regressing data from $\mathbb{R}^n$ to $\mathbb{R}^m$) is not addressed in the literature. Usual nonlinear regression techniques address much simpler problems, where the target variable is scalar. Spline regression may tackle this issue. But processing each component of the data independently is precluded here because of the computational cost. Therefore, we process all components jointly, using locally linear models. The problem is then to label all $\boldsymbol{x}_i$'s and to estimate the matrices corresponding to the linear models (see algorithm 2).

```
Input: original and reduced coordinates
Output: point labels and model matrices
begin
    while exists a point whose neighbors are not labeled do
        Pick randomly a point whose neighbors are not labeled;
        Compute the matrix $\boldsymbol{W}$ regressing this neighborhood ① ;
        Update all labels and update all matrices ②;
        Prune any model (labels and matrix) having too few points ③;
    end
    for every remaining unlabeled point do
        Search through the neighborhood;
        Assign them to nearby planes ④
    end
end
```

Algorithm 2: piecewise linear regression

Some details are needed for algorithm 2 :

① Let $\boldsymbol{Y}_i$ be the matrix consisting of all $\boldsymbol{y}_j$ that are neighbors of $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ the corresponding matrix of all $\boldsymbol{x}_j$. Matrix $\boldsymbol{W}$ is estimated from the equation $\boldsymbol{Y}_i \simeq \boldsymbol{W}\boldsymbol{X}_i$ by mean 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 linear model. Once all updates are completed, these variances are computed again for all linear models and their assigned points.

③ If a linear model has too few points, the matrix that describes it cannot be computed. In this case, the linear model is discarded.

④ When no linear model can be added, some points may remain unassigned. In this case, they are assigned to the linear model most represented in their neighborhood.

2.3. Projecting a point on the manifold

An important feature of data reduction techniques and manifold learning is the ability to project or to process (*e.g.,* classify) an incoming point without running all computations from scratch. Formally, for a given $\boldsymbol{y}_i$, we must determine the corresponding $\boldsymbol{x}_i$ and ε_i.

For each linear model $\boldsymbol{W}_j$, a candidate $(\boldsymbol{x}_{ij}, \varepsilon_{ij})$ is computed by mean squares. The variance associated to $\boldsymbol{W}_j$ is used to compute the likelihood of $\boldsymbol{x}_{ij}$. The maximum likelihood estimate is retained (see algorithm 3).

```
Input: reduced coordinates, point labels and model matrices
Output:
begin
    for each linear model do
        Compute corresponding reduced coordinates;
        Map the reduced coordinates on the manifold;
        Compute its likelihood;
    end
    Retain the label maximizing the likelihood;
end
```

Algorithm 3: projection of a new point on the manifold

3. APPLICATION

We apply our data reduction and regression methods to several standard data sets. Comparison with other methods is proposed. We also deal with projecting a new point on the manifold.

Two types of data are analyzed, on the one hand, large sets of data (the Swiss Roll) of low extrinsic and intrinsic dimensions, and on the other hand, data sets having far fewer points, but of high extrinsic dimension (the duck images, from the COIL-20 [10] data sets).

3.1. Data reduction

Our data reduction approach is compared to a standard method in the field, namely Isomap. To quantify the quality of the data reduction step, true geodesic distances (computed analytically) are compared to Euclidean distances in the reduced space. Comparison is achieved using correlation. Data are noise corrupted. Results show that our OCF behaves slightly better than Isomap (see table 1).Parameter σ is set to the first percentile of all distances on the manifold. Several runs are achieved to try to escape from local minima.

3.2. Regression

To assess the quality of the regression, the samples used to learn the linear models are projected onto the approximated manifold. Fig. 2 displays the result for the SwissRoll.

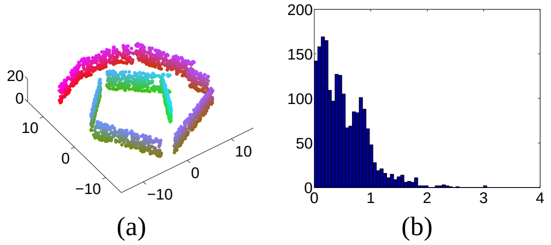

Fig. 2. SwissRoll regression. (a) regression of the SwissRoll, (b) histogram of the norm of the reconstruction error

Noise	Isomap	OCF
None	0.9997	0.9997
Gaus. 2.5%	0.9988	0.9992
Exp. 0.3%	0.9996	0.9997

Table 1. Correlation between real distances and estimated distances for the SwissRoll for Isomap and OCF with an 8-neighborhood for compression. The percentage indicates the value of the variance of the data divided by the variance of the noise.

Low-dimension coordinates computed with Isomap and with OCF lead to comparable reconstruction errors although increasing noise levels leads to lower errors with OCF. The piecewise linear regression algorithm is executed several times to try to escape from local optima. A higher number of neighbors can generate a higher variance, but a smoother global manifold, while a lower number of neighbors will lead to a lower reconstruction error, but can generate a noisy manifold.

We also addressed the reconstruction of an image of the Coil-20 database (see Fig. 3). We reconstructed one of the duck images with PCA (6 and 20 principal vectors) and with our technique (with 2 coordinates only). Our approach clearly exhibits more flexibility than PCA. Moreover, physical interpretation of the data generating process is impossible in dimension 20. This interpretation is possible in dimension 2, which is the dimension of the reduced variables of our approach.

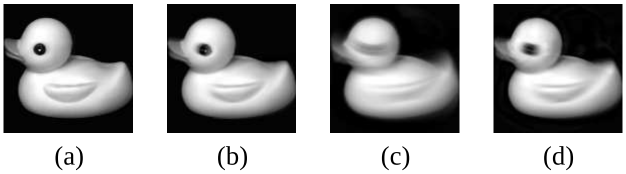

Fig. 3. Reconstruction example. (a) Original duck image, (b) Projected duck image with the proposed method (reduced variable of dimension 2) (c, d) Projected duck image with 6 and 20 principal vectors (PCA)

4. CONCLUSION AND FUTURE WORK

We have presented a comprehensive framework for learning a nonlinear manifold and for projecting new points on this manifold. The framework is divided into two main steps, the first being a dimensionality reduction process that enables learning reduced coordinates, and the second being a piecewise linear mapping of the manifold with the reduced coordinates, leading to an efficient projection on the manifold.

Work is in progress regarding manifolds linked to shape representations. The ultimate goal of our study is learning and classification of brain structures in medical images.

5. REFERENCES

[1] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B*, vol. 61, no. 3, pp. 611–622, 1999.

[2] N. Kambhatla and T.K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, pp. 1493–1516, 1997.

[3] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2.22, Dec 2000.

[4] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec 2000.

[5] I. Borg, *Modern Multidimensional Scaling*, Springer, 2nd edition, Aug 2005.

[6] M. Balasubramanian, E.L. Schwartz, J.B. Tenenbaum, Vin de Silva, and John C. Langford, "The isomap algorithm and topological stability," *Science*, vol. 295, pp. 7a, Jan 2002.

[7] S.T. Roweis, L.K. Saul, and G.E. Hinton, "Global coordination of local linear models.," in *Advances in Neural Information Processing Systems 14*. 2001, pp. 889–896, MIT Press.

[8] M. Brand, "Charting a manifold," Tech. Rep. TR-2003-13, MERL, Mar 2003.

[9] A.Webb, *Statistical Pattern Recognition*, Wiley, 2002.

[10] S.A. Nene, S.K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Tech. Rep. CUCS-005-96, Columbia University Computer Science, Feb 1996.