LOCALLY LINEAR EMBEDDING FOR NONLINEAR DIMENSION REDUCTION IN CLASSIFICATION PROBLEMS: AN APPLICATION TO GENE EXPRESSION DATA

M. Pillati, C. Viroli

1. INTRODUCTION

Although classification is by no means a new subject in the statistical literature, the large and complex multivariate datasets typical of some real problems raise new methodological and computational challenges.

For example, in the last few years gene expression measurements, as currently determined by the microarray technology, have been increasingly used in cancer research to obtain a reliable and precise classification of tumors. The data from such experiments are usually in the form of large matrices containing the expression levels of p genes under n experimental conditions (different times, cells, tissues ...), where n is usually less than 100 and p can easily be several thousands.

The particular condition $p \gg n$ makes most of the standard statistical techniques difficult to employ and dimensionality reduction methods are required.

One possible solution consists in performing a variable selection procedure to avoid the inclusion of not relevant or noisy variables that may degrade the overall performances of the estimated classification rule. There is a vast literature on gene selection for cell classification; a comparative study of several discrimination methods in the context of cancer classification based on filtered sets of genes can be found in Dudoit *et al.* (2002) (see also the more recent proposal in Calò *et al.*, 2005).

As an alternative, the dimensionality reduction can be addressed by mapping the high dimensional data onto a meaningful lower-dimensional latent space. There is a wide class of techniques for the dimensionality reduction task that operate under the hypothesis that the submanifold is embedeed linearly, almost linearly or non-linearly.

In this paper we provide a classification rule based on a supervised version of Locally Linear Embedding (LLE), useful to deal with high dimensional data for which the condition $p \gg n$ holds) lying on non linear structures.

2. THE DIMENSION REDUCTION PROBLEM

In general terms, dimensionality reduction consists in mapping a multidimensional space into a lower dimensional one. Its aim is to obtain compact representations of the data that are essential for higher-level analysis while eliminating unimportant or noisy factors otherwise hiding meaningful relationships and correlations.

If the data fall exactly on a smooth, locally flat subspace then the reduced dimensions are just the coordinates in the subspace and dimensionality reduction can be accomplished without loss of information. More commonly, data are noisy and an exact mapping does not exist. However a relatively small number of relevant directions for describing the true process underlying the data generating mechanism expectantly exists. In doing that we certainly discard some of the original information, but this loss is hopefully less than the gain we could obtain by simplifying the data structure.

Classical techinques for dimensionality reduction are designed to operate when the submanifold is linearly embedded, that is the high dimensional data are assumed to lye close to a hyperplane. Then each data point can be approximated using the vectors that span the hyperplane alone.

If these vectors are the d (with $d \ll p$) eigenvectors corresponding to the largest eigenvalues of the data covariance matrix, the mapping procedure corresponds to the popular Principal Components Analysis (PCA). PCA finds the low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional observed space.

Another popular technique, the Multidimensional Scaling (MDS), finds an embedding that preserves the pairwise distances between data points, which is equivalent to PCA when the distances are Euclidean. Carrying on this line, Linear Discriminant Analysis (LDA) finds the low-dimensional embedding of the data points that best discriminates between two or more groups.

However in many cases of interest, the way the dimensions depend on each other can be very complex and this can lead to data with non linear structure that can be invisible to PCA or MDS. Figure 1 shows a three-dimensional simulated example, the Swiss role data, sampled from a non linear two-dimensional manifold. Note that PCA (in the third graph of the figure) fails to recover the underlying structure of the manifold, since it maps faraway data points to nearby points in the two-dimensional space and vice versa.

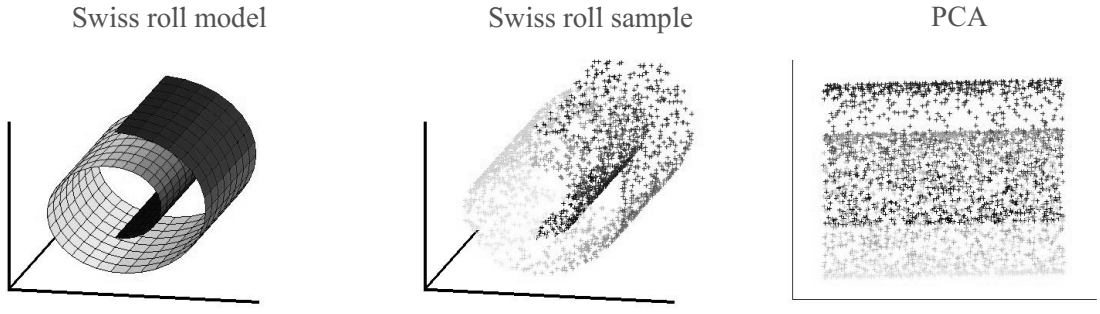

Figure 1 – Dimension reduction with PCA on Swiss role data.

There is a wide class of recently proposed methods by which unsatisfactory linear representations obtained by PCA or MDS may be "improved" towards more successful representations of the data. These techniques include principal curves and surfaces (Hastie and Stuetzle, 1989), generative topographic maps (Bishop *et al.* 1998), self-organizing maps (Kohonen, 1998) or the more recent methods based on non linear multidimensional scaling called Isomap (Tenenbaum *et al.*, 2000).

Another solution in non linear dimension reduction is based on the idea that global structure of the high dimensional data set can be retained in a collection of local structures when projecting the data to a low-dimensional space. One of the methods following this lines is the locally linear embedding of Roweis and Saul (2000), which will be described in details in the next section.

3. LOCALLY LINEAR EMBEDDING

Locally linear embedding is an unsupervised technique for dimensional reduction that looks for an embedding that preserves the local geometry in the neighbourhood of each data point.

In other words, nearby points in the high dimensional space have to remain nearby and similarly co-located with respect to each other in the low dimensional space.

Starting from this intuition each data point $\mathbf{x}_i \in \Re^p$, i=1,..., n, is approximated by a weighted linear combination of its neighbours (from the nature of these local and linear reconstructions the algorithm derives its name).

The neighbourhood of each data point $\mathbf{x}_i$ can be identified in two different ways: *a*) by selecting all the points lying within a hypersphere of radius ε centered at $\mathbf{x}_i$; *b*) by the k nearest neighbours of $\mathbf{x}_i$, as measured by euclidean distance. Both the solutions require a parameter to be set, but in the latter one, the most widely used, the same number of neighbours for every point simplifies the computation.

In its base formulation, the LLE algorithm finds the linear coefficients w_{ij} by minimizing the reconstruction errors

$$\varepsilon(\mathbf{w}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} w_{ij} \mathbf{x}_j \right\|^2 , \tag{1}$$

in which $\|\cdot\|$ is the Euclidean norm.

If the number of neighbours exceeds the input dimensionality ($k > p$), the least square problem (1) does not have a unique solution and a regularisation term must be added to the reconstruction error.

It follows directly from the cost function (1) that for each vector $\mathbf{x}_i$ the weights w_{ij} are invariant to rotations and rescalings of that data point and its neighbours. Moreover, the weights are subject to two constraints, a sparseness constraint and

an invariance constraint. The sparseness contraint means that the weights are enforced to be equal to zero if a point $\mathbf{x}_j$ does not belong to the set of k neighbours of $\mathbf{x}_i$. The second restriction is that for each $\mathbf{x}_i$ the weights w_{ij} are enforced to sum to one. It derives from these two constraints that the weights are also invariant to translations. Thus, the reconstruction weights characterize intrinsic geometric properties of each neighbourhood.

Suppose the data lie on an underlying non-linear manifold of dimensionality $d \ll p$. It is then assumed that there exists a linear mapping, consisting of translations, rotations and rescalings, which maps the high dimensional neighbourhoods to global coordinates on the underlying manifold. As the reconstruction weights are invariant to translation, rotation and rescaling we can expect that the weights that characterize the local geometry in the original data space are equally valid for local pieces of the low dimensional embedding. In other words, the weights that reconstruct the original vectors $\mathbf{x}_i$ of dimensionality p can also be used to reconstruct the underlying manifold in d dimensions.

Let $\mathbf{y}_i \in \Re^d$ be the i^{th} coordinate in the embedding. The n coordinates are then estimated by minimizing the cost function:

$$\Phi(\mathbf{y}) = \sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j=1}^{n} w_{ij} \mathbf{y}_j \right\|^2 . \tag{2}$$

Then the embedding is calculated directly from the w_{ij} without reference to the original imputs $\mathbf{x}_i$.

Two constraints are imposed: the center of gravity of the data is set in the origin, *i.e.* $\sum_{i=1}^{n} \mathbf{y}_i = 0$, to ensure the uniqueness of the solution and the covariance matrix is supposed to be equal to the identity matrix, *i.e.* $1/n\ \mathbf{YY'}=\mathbf{I}$, otherwise $\mathbf{Y} = \mathbf{0}$ would minimise (2).

While the cost function (1) can be minimized by solving a set of constrained least square problems, the embedding cost function (2) can be optimized, after having introduced the two constraints above, by solving a $n \times n$ eigenvalue problem which is a global operation over all the data points.

The optimal solution is given by the eigenvectors corresponding to the smallest eigenvalues of the matrix (I-W)′(I-W), where W is the $n \times n$ sparse matrix containing the weights w_{ij}. As the last eigenvector is the unit vector with equal entries and with eigenvalue equal to 0, we need to compute the eigenvectors corresponding to the bottom $(d+1)$ eigenvalues of the matrix and discard the last one to obtain an embedding centered at the origin.

To find a good LLE mapping, two parameters will have to be set: the dimensionality d of the reduced space and the neighbourhood size k. Incorrect choices for these parameters may degrade the results of the analysis: in fact, if d is set too high, the mapping will enhance noise; if it is set too low, distinct parts of the data set might be mapped one onto each other. On the other hand, if k is set too

small, the mapping will not reflect any global properties; if it is too high, the mapping will loose its nonlinear character as the entire data set is seen as local neighbourhood (some methods for the setting of these parameters are discussed in de Ridder and Duin, 2002).

4. SUPERVISED LOCALLY LINEAR EMBEDDING

Let X be a $p \times n$ data matrix, in which the n units belong to G different classes and $p \gg n$.

As traditional LLE is an unsupervised dimension reduction method, it does not make use of the class membership of each point to be mapped. But if groups are not well separated in the high-dimensional space, they will remain so in the embedding. This is perhaps the reason why the application of LLE as a dimension reduction method before performing a discriminant analysis does not always give good results (de Ridder *et al.*, 2003).

A supervised version of LLE, useful to deal with data sets containing multiple manifolds, corresponding to different classes, has been proposed in the literature (see de Ridder and Duin, 2002).

For each $\mathbf{x}_i$ from a class g ($1 \le g \le G$) a set of k neighbours is defined by selecting the closest points (in terms of Euclidean distance) to $\mathbf{x}_i$ belonging to the same class g. The mapping of the n units of a training set into a low-dimensional space follows the LLE procedure described in Section 3, but the local neighbourhood is made up of observations belonging to the same class. This procedure leads to a perfect separation of the n points in the low-dimensional space.

Unsupervised and supervised version of LLE can be viewed as particular cases of a more general approach that can be obtained by defining the neighborhood of a units $\mathbf{x}_i$ as the set of k points nearest to $\mathbf{x}_i$ in terms of the following modified distance:

$$d_{ij}^{*} = d_{ij} + \alpha\lambda_{ij} \max(d_{ij})_{i,j=1,\ldots,n} \qquad 0 \le \alpha \le 1 \tag{3}$$

where d_{ij} is the Euclidean distance, λ_{ij} is equal to zero if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class and to 1 otherwise. This means that the distance of units belonging to different classes is increased, and the amount of the increase is controlled by the α parameter. The distance between elements of the same class remains unchanged.

When $\alpha = 0$, equation (3) gives the unsupervised LLE, *i.e.* the basic version of the method, while when $\alpha = 1$, the result is the supervised LLE.

For $0<\alpha<1$ a mapping is found which preserves some of the manifold structure but introduces separation between classes.

The parameter α introduced in this "partially" supervised version of LLE controls the amount to which class information should be incorporated.

The effect of different values of α can be shown in Figure 2, in which a gene expression data set is mapped onto a two-dimensional space obtained by LLE.

As shown, the separation between classes increases with α. In the forth picture, which corresponds to α=1, the cells belonging to the same class are mapped on the same point. This is due to the fact that the eigenvectors related to the last G eigenvalues assign the same value to the elements of the same class, which corresponds to the mean. It follows that all the members of the same class are mapped on a single point in $\Re^{G-1}$, and G-1 is therefore the optimal embedding dimension for α=1. For $\alpha \neq 1$ this is not necessarily optimal.

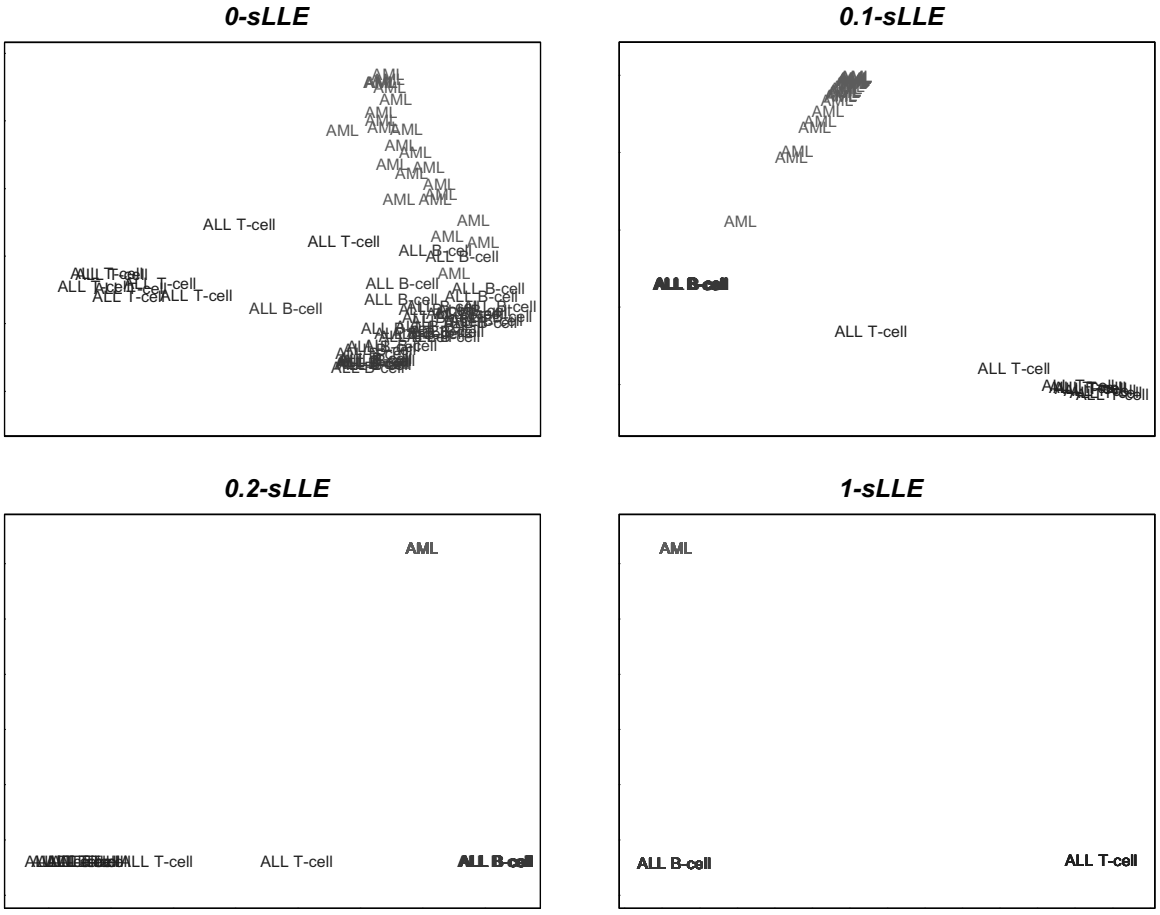

Figure 2 – α-supervised LLE on Leukemia data set (Golub *et al.*, 1999).

The perfect separation between the classes in the low dimensional space for α=1 and $d=G-1$ holds only in the training set, and the best solution in terms of class separation for previously unobserved units is not necessarily obtained with those particular choices for α and d. The use of locally linear embedding in a supervised classification context makes the mapping of a new observation a crucial issue, and in the next section we propose a solution suitable to handle the problem in presence of high-dimensional data.

5. MAPPING AND CLASSIFICATION OF A NEW OBSERVATION

The locally linear embedding is an exploratory data analysis method that requires a whole set of points as an input in order to map it into the embedded space and in the original works of Roweis and Saul (2000) the issue of how a new data point can be mapped was not taken into account, because of the exploratory nature of the method. In presence of new data points, the best way to map them should be to pool old and new points and rerun locally linear embedding again.

In their 2002 paper, De Ridder and Duin suggest however that, after embedding, a new observation $\mathbf{x}^*$ can be mapped quickly by calculating the weights for reconstructing it by its k nearest neighbours (as in the first step of LLE, but now the minimization involves only one units). The optimal weights are then used to derive the coordinates in the low dimensional space by a linear combination of the embedding of the k nearest neighbors.

This solution, that provides a significant reduction in the processing time, is not suitable in the context of supervised LLE, since for a new observation $\mathbf{x}^*$, whose class membership is unknown, the neighbourhood can not be identified.

As an alternative, a global approximation of the mapping between the high dimensional coordinates and the low-dimensional ones is proposed in this paper. After the embedding has been recovered, the problem can be addressed by finding an approximation of the mapping between the two spaces, in order to derive the embedding projection $\mathbf{y}^*$ of $\mathbf{x}^*$. The simplest solution is to consider the linear approximation, but in doing this we have to face the problem of high dimensional data for which the condition $p \gg n$ holds.

Thus, after embedding, we look for the best linear transformation between the two spaces that minimizes the following least square problem:

$$\min_{A} \sum_{i=1}^{n} \left\| \mathbf{x}_i - A\mathbf{y}_i \right\|^2$$

where $\hat{A} = XY'(YY')^{-1}$ is a $p \times d$ matrix. Thus a new observation $\mathbf{x}^*$ can be mapped in the latent space by the pseudo-inverse of the matrix Â:

$$\mathbf{y}^* = (A'A)^{-1} A' \mathbf{x}^*$$

Thanks to the use of class information in the feature extraction step, as explained in Section 4, in the allocation phase the employment of a simple classification rule suffices, as suggested in De Ridder *et al.* (2003). Thus the new observation $\mathbf{x}^*$ is assigned to the class having the nearest centroid in the reduced d-dimensional space.

6. EMPIRICAL ANALYSIS ON GENE EXPRESSION DATA

We applied our proposal on four publicy available data sets: the lymphoma data set of Alizadeh *et al.* (2000), the leukemia data set of Golub *et al.* (1999), the

mammary data set of Wit and McClure (2004) and the small round blue cell tumor data set of Khan *et al.* (2001). The main features of these real data sets are summarized in Table 1.

TABLE 1

Dataset description

Dataset	*N. of classes*	*N. of variables*	*N. of samples*	*Origin*
Leukemia	3	2,226	72	Golub *et al.* (1999)
Small blue cell tumor	4	2,308	63	Khan *et al.* (2001)
Lymphoma	3	4,026	62	Alizadeth *et al.* (2000)
Mammary	4	12,488	54	Wit and McClure (2004)

According to LLE, different classification rules have been obtained from several values for α, for each of which different values of the neighbourhood size k and of the embedding dimension d are explored.

In order to check the validity of LLE, and in particular of the proposed solution, we consider also alternative dimension reduction methods, i.e. PCA and ICA.

Given the small number of cells in each data set, the misclassification rate of each classifier has been estimated by averaging the error rates obtained on 100 balanced cross validations sets.

In the first two columns of Table 2 the estimated error rates for classifiers based on PCA and ICA are reported.

TABLE 2

Cross validated misclassification rates of the best classifiers for each considered data set

Dataset	*PCA*	*ICA*	*α-LLE*					
			Allocations by LLE weights			***Allocation by Global Linear Approximation***		
			α=0	*α=0.1*	*α=1*	*α=0*	*α=0.1*	*α=1*
Leukemia	0.042	0.042	0.051	0.067	0.079	0.050	0.042	**0.035**
Small blue cell tumor	0.064	0.048	0.116	0.081	0.092	0.051	0.041	**0.012**
Lymphoma	0.016	0.016	0.028	0.016	0.016	0.028	0.025	**0.005**
Mammary	0.148	0.148	0.154	0.093	0.086	0.084	0.074	**0.070**

The other columns show the estimated misclassification rates of classification rules based on LLE, for different values of α. More precisely, the first three columns show the results obtained by mapping test observations by minimizing the reconstruction error for each of them, as suggested in De Ridder and Duin (2002). The last three columns show the estimated rates obtained following the solution we have proposed in the previous section.

Different values of k and d have been considered and the cross-validated error rates of the classifiers with the best performance in the four training sets are reported.

The minimum error rate, for each data set, is the one obtained following our strategy, even if the superiority of the corresponding classifiers with respect to the others differs in the four data sets.

For the Leukemia data set, for example, the superiority of LLE with respect to the linear reduction methods is not so evident, but for the other data sets this is no longer true. This may depend on the different degrees of non linearity of the data structure in the observed space.

The supervised version of LLE generally outperforms the original one, but the amount of the improvement still strictly depends on the class structure of data in the high-dimensional space.

The smaller error rates obtained with the proposed solution seem to suggest that the mapping of new observation based only on the reconstruction weights looses relevant information about the relations between the two spaces.

7. CONCLUSIONS

In this paper we propose to deal with classification problems with high dimensional data, through the so-called locally linear embedding. The goal of this methodology is to recover low-dimensional, neighbourhood preserving embedding of high dimensional data.

We consider the supervised version of the method in order to take into account of class information in the feature extraction phase.

We propose a solution to the problem of mapping a new observation suitable to handle high-dimensional data for which the condition $p \gg n$ holds.

The proposed discriminant strategy is applied to the problem of cell classification using gene expression data.

In the four considered data set, it leads to classifiers with small misclassification rates, that are competitive with the ones obtained with other techniques on the same data sets (see, among the others, Calò *et al.* (2005) in which the nearest shrunken centroid method of Tibshirani *et al.* (2002) have been applied to the same data).

As the preliminary results on these real data sets show, the proposed strategy seems to represent a useful tool for supervised classification when the number of variables is greater than the number of units.

However, some aspects deserve further analysis. In particular, the issues concerning the choice of both the neighbourhood size k and the dimension d of the reduced space should be examined in more depth.

Finally, the employment of different approximations for the mapping between the observed space and the embedding in order to project new observations in the latent space could be explored.

Dipartimento di Scienze Statistiche MARILENA PILLATI
Alma Mater Studiorum – Università di Bologna CINZIA VIROLI

REFERENCES

A. ALIZADETH, M. EISEN, R. DAVIS (2000), *Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,* "Nature", 403, pp. 503-511.

C. BISHOP, M. SVENSEN, C. WILLIAMS (1998), *The generative topographic mapping,* "Neural Computation", 10, pp. 215-235.

D.G. CALÒ, G. GALIMBERTI, M. PILLATI, C. VIROLI (2005), *Variable selection in cell classification problems : a strategy based on independent component analysis*, in M. VICHI, P. MONARI, S. MIGNANI, A. MONTANARI (eds.) *New Developments in Classification and Data Analysis*, Springer, Heidelberg, pp. 21-29.

D. DE RIDDER, R. DUIN, (2002), *Locally linear embedding for classification*, Technical Report PH-2002-1, Delf University of Technology, pattern Recognition Group.

D. DE RIDDER, O. KOUROPTEVA, O. OKUN, M. PIETIK AINEN, AND R.P.W. DUIN (2003), *Supervised locally linear embedding*, "Proceeding ICANN/ICONIP 2003, Lecture Notes in Computer Science", pp. 333-341, Springer-Verlag.

S. DUDOIT, J. FRIDLYAND, T. SPEED, (2002), *Comparison of discrimnation methods for the classification of tumors using gene expression data*, "Journal of the American Statistical Association", 457, pp. 77-87.

T. GOLUB, D. SLONIM, P. TAMAYO, (1999), *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, "Science", 286, pp. 531-537.

T. HASTIE, W. STUETZLE, (1989), *Principal curves,* "Journal of the American Statistical Association", 84, pp. 502-516.

J. KHAN, J. WEI, M. RINGNER, (2001), *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networs,* "Nature Medicine", 7, pp. 673-679.

T. KOHONEN, (1988), *Self-organization and Associative Memory,* Springer-Verlag, Berlin.

J. NILSSON, T. FIORETOS, M. HOGLUND, M. FONTES, (2004), *Approximate geodesic distances reveal biologically relevant structures in microarray data,* "Bioinformatics", 20, pp. 874-880.

S. ROWES, L. SAUL, (2000), *Non linear dimensionality reduction by locally linear embedding,* "Science", 290, pp. 2323-2326.

R. TIBSHIRANI, T. HASTIE, B. NARASIMHAN, G. CHU, (2002), *Diagnosis of multiple cancer types by shrunken centroids of gene expression,* "Proceedings of the National Accademy of Sciences", 99, 6567-6572.

E. WIT, J. MCCLURE, (2004), *Statistics for Microarrays: Design, Analysis and Inference,*Wiley, Chichester.

RIASSUNTO

Locally linear embedding per la riduzione delle dimensioni in problemi di classificazione: un'applicazione a dati di espressione genica

In alcuni problemi reali, quali il riconoscimento d'immagini o l'analisi di dati di espressione genica, si dispone dell'osservazione di un numero molto elevato di variabili in un esiguo numero di unità. La soluzione di problemi di classificazione in tale contesto non sempre può avvalersi dei metodi tradizionali per difficoltà sia di ordine analitico che interpretativo.

In questo lavoro si propone di affrontare il problema della classificazione in presenza di un numero di osservazioni inferiore al numero delle variabili attraverso il ricorso a una tecnica di riduzione delle dimensioni, detta *locally linear embedding*. L'impiego di una particolare versione del metodo consente di tener conto delle informazioni sulla classe di appartenenza nella fase di riduzione delle dimensioni. La strategia discriminante proposta è stata impiegata per la classificazione di tessuti mediante dati di espressione genica.

SUMMARY

Locally linear embedding for nonlinear dimension reduction in classification problems: an application to gene expression data

Some real problems, such as image recognition or the analysis of gene expression data, involve the observation of a very large number of variables on a few units. In such a context conventional classification methods are difficult to employ both from analytical and interpretative points of view.

In this paper we propose to deal with classification problems with high-dimensional data, through a non linear dimension reduction technique, the so-called locally linear embedding. We consider a supervised version of the method in order to take into account of class information in the feature extraction phase. The proposed discriminant strategy is applied to the problem of cell classification using gene expression data.