Pattern Recognition Letters 33 (2012) 256–262

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An unsupervised data projection that preserves the cluster structure

Lev Faivishevsky *, Jacob Goldberger

School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel

ARTICLE INFO

Article history:
Received 26 September 2010
Available online 2 November 2011
Communicated by G. Borgefors

Keywords:
Unsupervised dimensionality reduction
Mutual information
Clustering

ABSTRACT

In this paper we propose a new unsupervised dimensionality reduction algorithm that looks for a projection that optimally preserves the clustering data structure of the original space. Formally we attempt to find a projection that maximizes the mutual information between data points and clusters in the projected space. In order to compute the mutual information, we neither assume the data are given in terms of distributions nor impose any parametric model on the within-cluster distribution. Instead, we utilize a non-parametric estimation of the average cluster entropies and search for a linear projection and a clustering that maximizes the estimated mutual information between the projected data points and the clusters. The improved performance is demonstrated on both synthetic and real world examples.

1. Introduction

A general task of unsupervised learning is to reveal the hidden structure of unlabeled data sets. The induction of a structure is roughly equivalent to the reduction of the intrinsic complexity of the data set. One common approach to lowering data set complexity is to perform a dimensionality reduction step that produces a low dimensional representation of the data. Such a representation is often easier both for human perception (e.g. visualization by reduction to a two dimensional subspace) and for automatic inference, since a computationally intensive technique may be tractable only for low dimensional data. In addition, usually not all the features are relevant and projecting to a lower dimensional space can be viewed as extracting the most relevant features. Finally, dimensionality reduction can save storage and time when dealing with datasets with huge number of features.

Data projection methods transform data from the original D-dimensional feature space into a new d-dimensional ($d < D$) feature space. Principal component analysis (PCA) is one of the most popular methods for feature transformation. The PCA algorithm implicitly assumes that the data are sampled from a Gaussian distribution and projects the data onto the subspace with the maximum norm covariance matrix. PCA is limited, however, since it assumes that the data distribution is unimodal and does not take into account the cluster structure properties of the data. In addition the variance may not be a proper criterion for optimization when dealing with substantially non-Gaussian data. Non-linear manifold learning algorithms such as multidimensional scaling (MDS), Isomap (Tenenbaum et al., 2000), t-SNE (van der Maaten and Hinton, 2008) and locally linear embedding (LLE) (Roweis and Saul, 2000) map the input data into a single global coordinate system of lower dimensionality that preserves local relations between neighboring points. Like PCA these methods also do not aim to preserve the existing clustering structure in the original space.

The task of dimensionality reduction becomes easier if the data are labeled. In this case in addition to coordinates in R^D each data point is given a label $c \in \{1, \ldots, m\}$. It can be assumed that data points with the same labels are similar and are expected to be close in the projected space. Supervised method such as LDA, NCA (Goldberger et al., 2004) and many others explicitly exploit the known organization of the data into separate classes to obtain low-rank representations that are found to be much better than current unsupervised methods. The goal of supervised dimensionality reduction is to reduce the dimensionality of the input space while preserving the information about the output values. Thus, supervised dimensionality reduction benefits from knowing the data labels that enable to apply powerful optimization criteria. However for the unsupervised methods the exact labels are unavailable. The goal of this paper is to utilize estimated labels for dimensionality reduction to partially benefit from the advantages of a supervised setup even in an unsupervised framework. Such estimated labels are obtained by clustering the projected data. Naturally, the data points assigned to the same cluster are given the same label and vice versa. Both dimensionality reduction and clustering are optimally based on maximizing mutual information.

In addition to low-rank representation, an important way to structure unlabeled data is clustering, which may be considered as a cardinality reduction of the dataset. The clustering step divides all data points into many fewer number of groups. The rationale is

* Corresponding author.
E-mail addresses: levtemp@gmail.com (L. Faivishevsky), goldbej@eng.biu.ac.il (J. Goldberger).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.10.012

that within each group the data points possess nearly the same properties. As a consequence, subsequent calculations may be performed by treating each such group or its representative as a new generalized data point, that again alleviates human and automatic perception. The class of methods that cluster vectors in R^d includes the spectral clustering algorithms (Ng et al., 2002) that have attracted much attention in recent years. Another class of clustering algorithms also admits input in the form of vectors in R^d but in addition implicitly or explicitly assumes certain types of intra-cluster distributions (e.g. applying the EM algorithm to learn a Gaussian mixture density). The key example of such algorithms is k-means (Lloyd, 1982). When the data are arranged in non-convex sets (e.g. concentric circles) these algorithms tend to fail. These problems have now been solved in the recently introduced nonparametric information theoretic clustering algorithm (NIC) (Faivishevsky and Goldberger, 2010). NIC is a clustering algorithm based on maximizing the mutual information between data points and clusters. Unlike previous methods, NIC neither assumes the data are given in terms of distributions nor imposes any parametric model on the within-cluster distribution. Instead, the algorithm utilizes a non-parametric estimation of the average cluster entropies and searches for a clustering that maximizes the estimated mutual information between data points and clusters.

This paper introduces a novel approach that combines these two forms of the complexity reduction. The proposed algorithm performs dimensionality reduction that takes into account the possible clustering of the data. In particular, we search for the reduced complexity representation of the data in terms of both dimensionality and cardinality simultaneously, which maximizes the mutual information between the projected data and the cluster structure. The output of the algorithm is comprised of both a low-dimensional representation of the data and a clustering of points into groups, though each of them may be further used separately and independently. Here we do not assume any parametric assumptions either about the data points themselves or about their arrangements (e.g. distributions) into clusters, though the method may naturally incorporate such additional information. The rest of the paper is organized as follows. Section 2 introduces the proposed method. Section 3 describes numerical experiments.

2. The projection algorithm

Suppose a dataset X is composed of n data points $x_1,\ldots,x_n \in R^D$. In many cases the cluster structure is the most important feature of the dataset and our goal is to find a lower dimensionality representation of the data ($d < D$) that maximally preserves the cluster structure that exists in the original data. We restrict here the discussion to the simple case of a projection obtained by a linear transformation $A : R^D \to R^d$. In addition to the projection matrix we are also looking for an optimal clustering function $C : AX \to \{1,\ldots,m\}$ of the projected data. Denote the cluster of Ax_i by c_i. Note that the pair $\{A,C\}$ actually defines a reduced complexity representation of the original dataset X, where A accounts for the dimensionality reduction and C for the cardinality reduction (clustering).

We aim to obtain a maximally information preserving representation of the original data. Therefore we want to find the pair $\{A,C\}$ that achieves the maximal mutual information between the projected data and its cluster structure:

$$(\widehat{A},\widehat{C}) = \arg\max_{A}\max_{C} I(AX;C)$$

where A goes over all the linear transformation from $R^D \to R^d$ and C goes over all the clusterings of the set AX into (at most) m clusters. Using well known information-theoretic identities (Cover and Thomas, 1991) the mutual information (MI) score function takes the form:

$$I(AX;C) = H(AX) - H(AX|C) = H(AX) - \sum_{j=1}^{m} \frac{n_j}{n} H(AX|C=j) \qquad (1)$$

where n_j is the number of data points that are assigned by C to the cluster j. To utilize the MI score function (1) we have to tackle the technical issue of computing the within-cluster entropy terms $H(AX|C=j)$ and also the global entropy $H(AX)$. A simple assumption we can impose is that the conditional density $f(Ax|C=j)$ is Gaussian. Since there is a closed-form expression for the entropy of a Gaussian distribution we can compute the cluster score $I(AX;C)$ given the within-cluster Gaussian assumption. In the general case, however, we do not have any prior knowledge on the within-cluster distribution. Assuming that the intra-cluster distribution is Gaussian is not always a good choice since by utilizing a Gaussian distribution to describe the density we implicitly assume a unimodal blob type shape which is not always the case.

In the MI criterion $I(AX;C)$ the relevant term is not the within-cluster distribution but the within-cluster entropy. The key point is that by using a mutual information criterion (1) we do not need to have an explicit representation of the intra-cluster distribution. We only need to compute the cluster entropy. In what follows we utilize a nonparametric estimation of intra-cluster entropy in order to benefit from the MI score function (1).

Classical methods for estimating the mutual information $I(AX;C)$ require the estimation of the joint probability density function of (AX,C). This estimation must be carried out on the given dataset. Histogram- and Kernel-based (Parzen windows) pdf estimations are among the most commonly used methods (Torkkola, 2003). Their use is usually restricted to one- or two-dimensional probability density functions (i.e. pdf of one or two variables). However, for high-dimensional variables histogram- and Kernel-based estimators suffer dramatically from the curse of dimensionality; in other words, the number of samples needed to estimate the pdf grows exponentially with the number of variables. An additional difficulty in Kernel based estimation lies in the choice of Kernel width.

Other methods used to estimate the mutual information are based on k-nearest neighbor statistics (see e.g. Victor, 2002; Wang et al., 2009). A nice property of these estimators is that they can be easily utilized for high dimensional random vectors and no parameters need to be predefined or separably tuned for each clustering problem (other than determining the value of k). There are a number of non-parametric techniques for the (differential) entropy estimation of random vectors in R^d which are all variants of the following estimator (Kozachenko and Leonenko, 1987):

$$H_k = \frac{d}{n}\sum_{i=1}^{n} \log \epsilon_{ik} + \text{const(k)} \qquad (2)$$

where ϵ_{ik} is the Euclidean distance from i-th vector to its k-th nearest neighbor. The constant in Eq. (2) is:

$$\psi(n) - \psi(k) + \log(c_d)$$

where $\psi(x)$ is the digamma function (the logarithmic derivative of the gamma function) and c_d is the volume of the d-dimensional unit ball. The H_k entropy estimator is consistent in the sense that both the bias and the variance vanish as the sample sizes increase. The consistency of the 1-NN estimator was proven in (Kozachenko and Leonenko, 1987) and the consistency of the general k-NN version was shown in (Goria et al., 2005).

In the task of dimensionality reduction we need to optimize $I(AX;C)$ as a function of matrix A. The non-parametric kNN estimators (2) rely on order statistics. This would make the analytical calculation of the gradient of $I(AX;C)$ cumbersome and thus would complicate the optimization process. It also leads to a certain lack of smoothness of the estimator value as a function of the sample coordinates.

Here we utilize the MeanNN differential entropy estimator (Faivishevsky and Goldberger, 2009) due to its smoothness with respect to the coordinates of data points. Also, it was recently shown (Faivishevsky and Goldberger, 2010) that the MeanNN is more suitable for clustering than estimators that are based on kNN. The MeanNN estimator exploits the fact that the kNN estimation is valid for every k and therefore averaging estimators (2) for all possible values of k itself leads to a new estimator of the differential entropy:

$$\begin{aligned} H_{mean}(X) &= \frac{1}{n-1}\sum_{k=1}^{n-1} H_k(X) \\ &= \frac{d}{n(n-1)}\sum_{i\neq l}\log\|x_i - x_l\| + \text{const} \end{aligned} \tag{3}$$

This estimator computes the entropy based on the pair-wise distances between all the given data points and thus eliminates the calculation of nearest neighbors. Applying the estimator to the projected data yields:

$$H(AX) \approx H_{mean}(AX) = \frac{d}{n(n-1)}\sum_{i\neq l}\log\|A(x_i - x_l)\| \tag{4}$$

and applying this estimator to the intra-cluster entropy of the projected data yields:

$$\begin{aligned} H(AX|C=j) &\approx H_{mean}(AX|C=j) \\ &= \frac{d}{n_j(n_j-1)}\sum_{i\neq l|c_i=c_l=j}\log\|A(x_i - x_l)\| \end{aligned} \tag{5}$$

Plugging the H_{mean} estimations (4) and (5) into the MI score function (1) yields the following form of the quality measure:

$$\begin{aligned} S(A,C) =& H_{mean}(AX) - \sum_j \frac{n_j}{n} H_{mean}(AX|C=j) \\ =& \frac{1}{n(n-1)}\sum_{i\neq l}\log(\|A(x_i - x_l)\|^2) \\ & -\sum_j \frac{1}{n(n_j-1)}\sum_{i\neq l|c_i=c_l=j}\log(\|A(x_i - x_l)\|^2) \end{aligned} \tag{6}$$

The optimization of the score function (6) is performed iteratively in an alternating way. Each iteration τ includes two parts. First, for a given dimensionality reduction matrix $A_{\tau-1}$ we find an optimal clustering assignment C_τ (e.g. applying the sequential greedy optimization NIC algorithm Faivishevsky and Goldberger, 2010) to the unlabeled data $A_{\tau-1}X$. Next, we find the optimal dimensionality reduction matrix A_τ given the assignment C_τ. Note that for a fixed assignment of labels C the score function $S(A,C)$ is a smooth function of dimensionality reduction matrix A and the analytical expression for gradient is readily computed.

$$\frac{\partial S(A,C)}{\partial A} = \frac{A}{n}\left(\frac{1}{n-1}\sum_{i\neq j}\frac{(x_i-x_l)(x_i-x_l)^\top}{\|A(x_i-x_l)\|^2} - \sum_j \frac{1}{n_j-1}\sum_{i\neq l|c_i=c_l=j}\frac{(x_i-x_l)(x_i-x_l)^\top}{\|A(x_i-x_l)\|^2}\right) \tag{7}$$

For each given clustering C the optimal dimensionality reduction matrix A may be efficiently found, e.g. by the conjugate gradient method (Faivishevsky and Goldberger, 2010). It is often computationally beneficial to put a restriction on a form of the matrix A. For purposes of the dimensionality reduction we need to find an optimal subspace. The subspace is characterized by an orthonormal basis lying in it. Therefore the matrix A may be restricted to have d orthonormal columns, i.e. to be a $d * D$ submatrix of a $D * D$ rotation matrix W. We may parameterize the rotation matrix W by givens rotations (see e.g. Peltonen and Kaski, 2005). In this parametrization a rotation matrix $W \in R^{D\times D}$ is represented by a product of $D(D-1)/2$ plane rotations:

$$W = \prod_{s=1}^{D-1}\prod_{t=s+1}^{D} G_{st} \tag{8}$$

where G_{st} is a rotation matrix corresponding to a rotation in the st plane by an angle λ_{st}. It is the identity matrix except that its elements (s,s), (s,t), (t,s), (t,t) form a two-dimensional rotation matrix by

$$\begin{bmatrix} G_{st}(s,s) & G_{st}(s,t) \\ G_{st}(t,s) & G_{st}(t,t) \end{bmatrix} = \begin{bmatrix} \cos(\lambda_{st}) & \sin(\lambda_{st}) \\ -\sin(\lambda_{st}) & \cos(\lambda_{st}) \end{bmatrix} \tag{9}$$

This way the score function (6) takes the form $S(A(\lambda),C)$. The gradient of a single rotation matrix G_{st} with respect to λ_{st} is a zero matrix except for elements (s,s), (s,t), (t,s), (t,t) for which

$$\frac{\partial}{\partial\lambda_{st}}\begin{bmatrix} G_{st}(s,s) & G_{st}(s,t) \\ G_{st}(t,s) & G_{st}(t,t) \end{bmatrix} = \begin{bmatrix} -\sin(\lambda_{st}) & \cos(\lambda_{st}) \\ -\cos(\lambda_{st}) & -\sin(\lambda_{st}) \end{bmatrix} \tag{10}$$

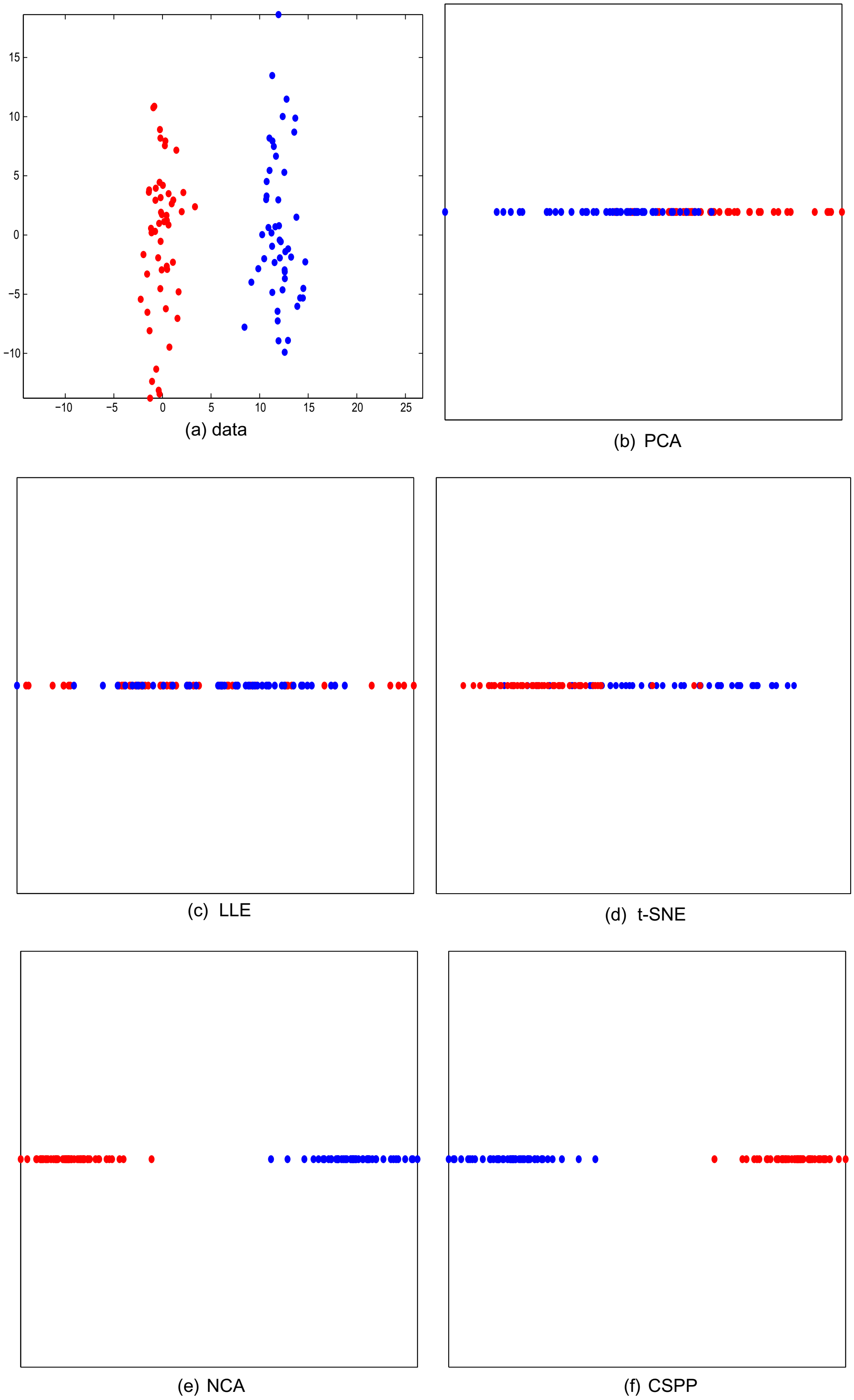
We may then compute the gradient of the score function $S(A(\lambda),C)$ as

$$\frac{\partial S(A(\lambda),C)}{\partial\lambda_{st}} = \sum_{q,r}\left[\frac{\partial S}{\partial A}\right]_{qr}\left[\prod_{u=1}^{D-1}\prod_{v=u+1}^{D}\widetilde{G}_{uv}\right]_{qr} \tag{11}$$

where $\widetilde{G}_{uv} = \frac{\partial}{\partial\lambda_{uv}}G_{uv}$ if both $u = s$ and $v = t$, and $\widetilde{G}_{uv} = G_{uv}$ otherwise.

```
Input: Data vectors $X = \{x_1, x_2, ..., x_n\} \subset R^D$, number of clusters $m$, target dimension $d$.
Output: Dimensionality reduction matrix $A : R^D \rightarrow R^d$, clustering $C : X \rightarrow \{1, ..., m\}$.
Method:
  1. Randomly initialize $D(D-1)/2$ rotation angles $\lambda_0$, assignment $C_0(X)$.
  2. Compute dimensionality reduction matrix $A_0 = A(\lambda_0)$.
  3. Do until convergence
      • Cluster projected data points $A_{\tau-1}X$ by the NIC algorithm [4] to obtain an assignment $C_\tau$.
      • Find optimal rotation angles $\lambda_\tau$ by maximization of score function $S(A(\lambda_\tau), C_\tau)$ applying conjugate gradient method.
      • Compute dimensionality reduction matrix $A_\tau = A(\lambda_\tau)$.
```

Fig. 1. The cluster structure preserving projection algorithm (CSPP).



(a) data

(b) PCA

(c) LLE

(d) t-SNE

(e) NCA

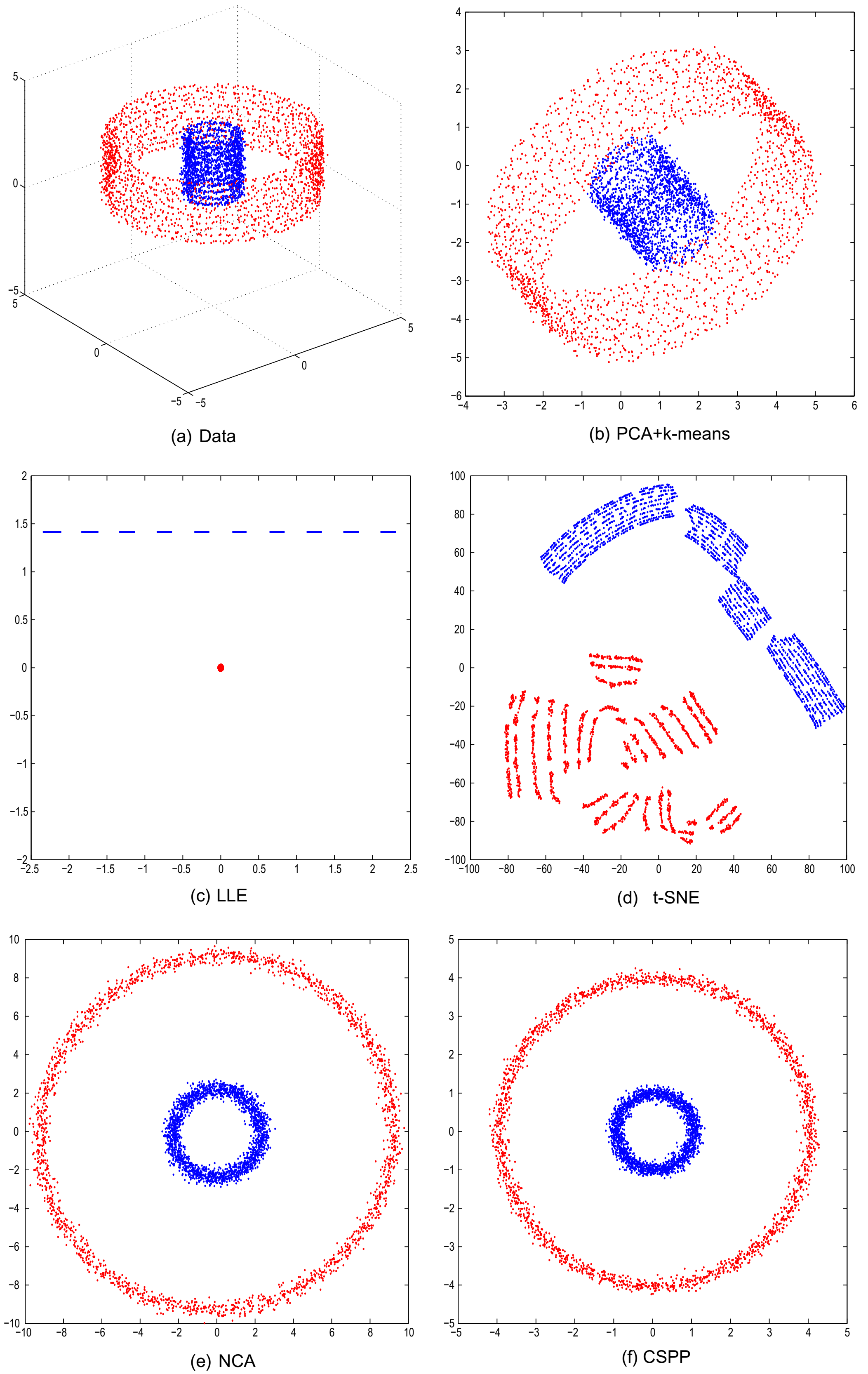
(f) CSPP

Fig. 2. Example of applying dimensionality reduction algorithms on a 2D dataset.

For a fixed clustering C_τ the optimal rotation angles λ_τ are found by a conjugate gradient method. The algorithm, which we dub the cluster structure preserving projection (CSPP), is summarized in Fig. 1.

3. Experiments

In this section we concentrate on two types of experiments. First we conducted experiments on synthetic data. The purpose of these



(a) Data

(b) PCA+k-means

(c) LLE

(d) t-SNE

(e) NCA

(f) CSPP

Fig. 3. Example of dimensionality reduction of a 3D dataset.

experiments is to emphasize the ideas behind the proposed method. Then we proceed with experiments on real world datasets from the UCI machine learning repository (Frank and Asuncion, 2010) in order to compare the method to other approaches.

3.1. Synthetic data

We begin by a simple motivating example for the proposed CSPP approach of simultaneous dimensionality reduction and clustering.

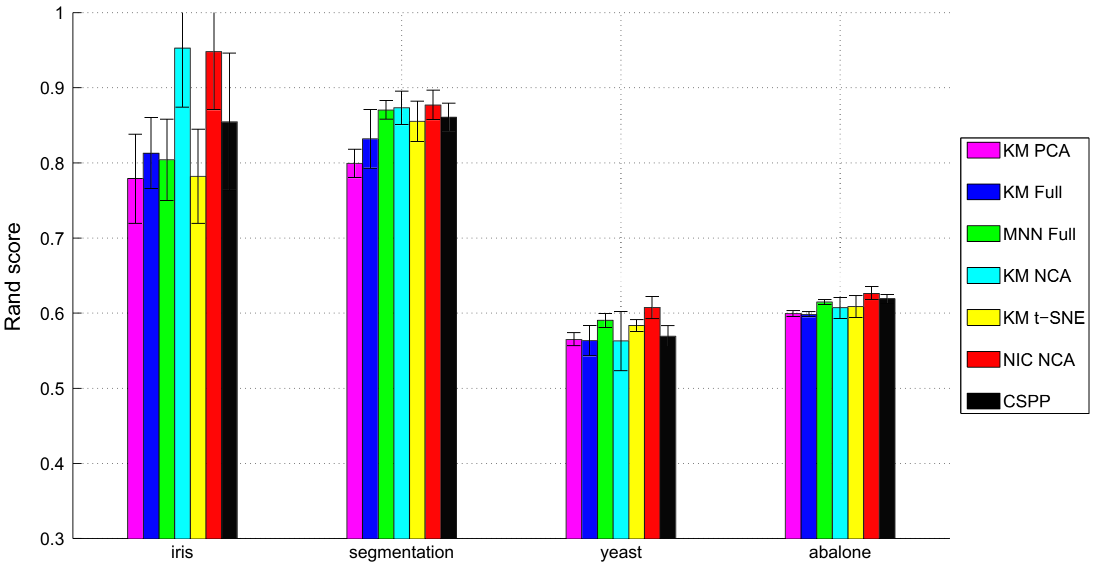


Fig. 4. Effect of simultaneous dimensionality reduction and clustering on UCI data sets. Statistics are shown for 10 repetitions.

Fig. 2a shows 2D Gaussians with most of the in-cluster variability in axis y whereas the clusters are naturally separable in axis x. In such a case the PCA method projects the data in one-dimensional subspace in the direction of in-cluster variability that eliminates the intrinsic structure of the original dataset and makes subsequent treatment practically impossible, see Fig. 2b. Alternatively, we may first apply clustering on the original data and use the obtained labels for supervised dimensionality reduction. Applying a 2-cluster k-means on the original data shown in Fig. 2a yields a data splitting based on the y-axis which is just a large variance noise that is uncorrelated with the clustering structure. Hence, clustering the original data results in wrong labels that mislead the supervised dimensionality reduction algorithm. The unsupervised non-linear manifold learning algorithm LLE also fails to project correctly the data because this approach can not account for multi cluster structure of the dataset, see Fig. 2c. The t-SNE algorithm (van der Maaten and Hinton, 2008) for visualization of high-dimensional data also does not reveal the correct structure of the data either, see Fig. 2d. On the other hand, supervised dimensionality reduction methods such as NCA (Goldberger et al., 2004) projects the data into another one-dimensional subspace preserving intrinsic structure, see Fig. 2e. Despite being unsupervised, the proposed method is able to provide the correct dimensionality reduction as a supervised method, see Fig. 2f.

The above example shows that the dimensionality reduction stage based on a variance criterion leads to inferior results compared to a mutual information based approach like CSPP. On the other hand, one could use a variant of the proposed algorithm with a variance based method like k-means in the clustering stage of the algorithm instead of mutual information based NIC clustering. However, this would lead to worse performance in the cases of non-convex data arrangements, where the in-cluster data distribution is far from Gaussian. We generated a synthetic example for such a case where the data are given as two concentric 3D cylinders, see Fig. 3a. Here the vertical dimension of the data is of less importance and therefore we perform a dimensionality reduction from $3D$ to $2D$. A variant of the PCA algorithm followed by a k-means segmentation fails to identify and to eliminate the non-informative dimension, see Fig. 3b. The unsupervised LLE method does not reveal correctly the two cluster structure, Fig. 3c and the unsupervised visualization technique t-SNE produces less intuitive results, Fig. 3d. However both a supervised method, based on NCA, Fig. 3e and the proposed unsupervised algorithm CSPP, Fig. 3f correctly perform the dimensionality reduction such that the remaining 2D representation is easy to perceive as a two cluster arrangement.

3.2. Datasets from the UCI repository

Next we compared the proposed CSPP method with other techniques on four real world datasets from the UCI machine learning repository (Frank and Asuncion, 2010). We used the Iris, Segmentation, Yeast and Abalone datasets.[1] The Iris dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. Each instance is described by four features. The segmentation dataset contains 7 classes with total 2310 instances. Each instance is presented by 19 features, it corresponds to a 3 × 3 region of one of 7 outdoor images. The Yeast dataset contains 10 classes with total 1462 instances. Each instance is composed of 8 predictive attributes and defines a localization site. Finally the Abalone dataset is composed of 3 classes with total 4177 instances. Each instance is given by 8 attributes and predicts the age of abalone from physical measurements.

The proposed CSPP performs both dimensionality reduction and clustering therefore a real world numerical experiment should define settings for both techniques. We reduced each dataset to a 2D subspace and measured the quality of the resulting clustering. The performance was measured in terms of the Rand score (Rand, 1971) which is a standard non-parametric measure of clustering quality. Let C_1 and C_2 be two clusterings of the same set (not necessarily with the same number of clusters). Then:

$$\text{Rand Score}(C_1, C_2) = \frac{n_{diff} + n_{same}}{n_{pairs}}$$

where n_{diff} is the number of pairs of points that belong to different clusters in both C_1 and C_2, n_{same} is the number of pairs of points that belong to the same cluster in both C_1 and C_2 and n_{pairs} is the total number of data pairs. In our case we compute the Rand score between the clustering obtained by the algorithm and the true clustering formed by the data labels.

We compared several methods for the intrinsic structure exploration of these data sets. First we used PCA dimensionality reduction followed by the k-means method for the clustering (KM PCA).

[1] http://archive.ics.uci.edu/ml.

Table 1
Mean run time (in second), for each method appears in Fig. 4, applied on the Iris dataset.

Method	Run time
KM PCA	0.02
KM Full	0.02
MNN Full	1.13
KM NCA	1.94
KM t-SNE	21.16
NIC NCA	3.06
CSPP	0.23

We also applied the *k*-means and NIC methods in the original space (KM Full and MNN full respectively). In addition we used the supervised dimensionality reduction by NCA (Goldberger et al., 2004) followed by k-means and the NIC algorithm followed by NCA. Finally we applied the t-SNE algorithm (van der Maaten and Hinton, 2008) followed by *k*-means. In each run we randomly chose 90% of original data points. To provide statistics we ran 10 repetitions per each dataset. The results are summarized in Fig. 4. In all cases the proposed CSPP method resulted in performance that was comparable to that of the methods running in the full original space and t-SNE and the supervised method NCA and better than that of the other algorithms running in the projected space. A comparative mean running time (in seconds) over the Iris dataset is shown in Table 1. Algorithms that show better performance than CSPP are either supervised (i.e. based on the data labels) or the running time is much higher.

4. Conclusion

We proposed a new approach to unsupervised manifold learning that performs both dimensionality reduction and clustering simultaneously. The method is based on revealing and preserving the intrinsic structure of the unlabeled data which yields maximal information about the whole high dimensional set. Improved performance is demonstrated in clustering application for both synthetic and real data sets. The strength of our approach is that, unlike other methods, we assume no explicit structure and no parametric probabilistic description of the clusters.

Future research may concentrate on nonlinear dimensionality reduction extensions of the proposed approach. In addition, the Kernel methods may be applicable for the proposed framework of simultaneous dimensionality reduction and clustering. Another possible research direction is modifying current manifold learning algorithms such as LLE and Isomap such that in addition to learning the local manifold structure, they will also preserve the more global cluster structure. In datasets that are organized in separate clusters this can significantly improve the performance of those learning algorithms.

References

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley-Interscience.

Faivishevsky, L., Goldberger, J., 2009. ICA based on a smooth estimation of the differential entropy. Adv. Neural Inform. Process Syst. 21.

Faivishevsky, L., Goldberger, J. 2010. A mutual information based dimensionality reduction with application to non-linear regression, IEEE Internat. Workshop on Machine Learning for Signal Processing.

Faivishevsky, L., Goldberger, J., 2010. A nonparametric information theoretic clustering algorithm. Int. Conf. Mach. Learn..

Frank, A., Asuncion, A. 2010. UCI machine learning repository.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighbourhood components analysis. Adv. Neural Inform. Process. Syst. 16.

Goria, M., Leonenko, N., Mergel, V., Inverardi, P., 2005. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. J. Nonparam. Statist., 277–297.

Kozachenko, L., Leonenko, N., 1987. On statistical estimation of entropy of random vector. Probl. Inform. Transmiss. 23 (2).

Lloyd, S.P., 1982. Least squares quantization in PCM. Special issue on quantization. IEEE Trans. Inform. Theory, 129–137.

Ng, A.Y., Jordan, M., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. Adv. Neural Inform. Process. Syst. 14.

Peltonen, J., Kaski, S., 2005. Discriminative components of data. IEEE Trans. Neural Netw. 16 (1).

Rand, W., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66, 846–850.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. J. Mach. Learn. Res..

van der Maaten, L.J.P., Hinton, G.E., 2008. Visualizing high-dimensional data using t-SNE. J. Mach. Learn Res., 2579–2605.

Victor, J.D., 2002. Binless strategies for estimation of information from neural data. Phys. Rev..

Wang, Q., Kulkarni, S.R., Verdu, S., 2009. Divergence estimation for multidimensional densities via *k*-nearest-neighbor distances. IEEE Trans. Inform. Theory, 2392–2405.