Pattern Recognition 87 (2019) 55–66

Contents lists available at ScienceDirect

Pattern Recognition

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

Nonlinear supervised dimensionality reduction via smooth regular embeddings

Cem Örnek, Elif Vural*

Department of Electrical and Electronics Engineering, METU, Ankara

ARTICLE INFO

Article history:
Received 19 October 2017
Revised 16 August 2018
Accepted 8 October 2018
Available online 10 October 2018

Keywords:
Manifold learning
Dimensionality reduction
Supervised learning
Out-of-sample
Nonlinear embeddings

ABSTRACT

The recovery of the intrinsic geometric structures of data collections is an important problem in data analysis. Supervised extensions of several manifold learning approaches have been proposed in the recent years. Meanwhile, existing methods primarily focus on the embedding of the training data, and the generalization of the embedding to initially unseen test data is rather ignored. In this work, we build on recent theoretical results on the generalization performance of supervised manifold learning algorithms. Motivated by these performance bounds, we propose a supervised manifold learning method that computes a nonlinear embedding while constructing a smooth and regular interpolation function that extends the embedding to the whole data space in order to achieve satisfactory generalization. The embedding and the interpolator are jointly learnt such that the Lipschitz regularity of the interpolator is imposed while ensuring the separation between different classes. Experimental results on several image data sets show that the proposed method outperforms traditional classifiers and the supervised dimensionality reduction algorithms in comparison in terms of classification accuracy in most settings.

1. Introduction

In many data analysis applications, collections of data are acquired in a high-dimensional ambient space; however, the intrinsic dimension of data is much lower. For instance, the face images of a person reside in a high-dimensional space, however, they are concentrated around a low-dimensional manifold that can be parameterized with a few variables such as pose and illumination parameters. An important problem of interest in data analysis has been the learning of low-dimensional models that provide suitable representations of data for accurate classification. Many supervised manifold learning methods have been proposed in the recent years that aim to enhance the separation between training samples from different classes while respecting the geometric structure of data manifolds. However, the generalization capabilities of such methods to initially unavailable novel samples have rather been overlooked so far. In this work, we propose a nonlinear supervised dimensionality reduction method that builds on theoretically established generalization bounds for manifold learning.

Classical methods such as LDA and Fisher's linear discriminant reduce the dimensionality of data by learning a projection so that the between-class separation is increased while the within-class separation is reduced. In the recent years, much research effort has focused on the discovery of low-dimensional structures in data sets, which gave rise to the topic of manifold learning [1–6]. Following these works, many supervised extensions of methods such as the Laplacian eigenmaps algorithm [3] have been proposed. Linear dimensionality reduction methods such as [7–14] learn a linear projection of training samples onto a lower-dimensional domain, where the distance between samples from different classes are increased and the distances within the same class are decreased. Most of these methods include a structure preservation objective as well, which aims to map nearby samples in the original domain to nearby locations in the new domain of embedding. Nonlinear methods such as [15] pursue a similar objective; however, the embedding is given by a pointwise nonlinear mapping instead of a linear projection.

The performance of linear methods depends largely on the distribution of the data in the original ambient space, since the distribution of the data after the embedding is strictly dependent on the original distribution via a linear projection. Nonlinear dimensionality reduction methods such as [15] have greater flexibility in the learnt representation. However, two critical issues arise concerning supervised dimensionality reduction methods: First, most nonlinear methods compute a pointwise mapping only for the initially available data samples. In order to generalize them to new points, an interpolation needs to be done, which is called the out-of-sample extension of the embedding. Second, existing dimension-

* Corresponding author.
E-mail address: velif@metu.edu.tr (E. Vural).

https://doi.org/10.1016/j.patcog.2018.10.006
0031-3203/

ality reduction methods focus on the properties of the computed embedding only as far as the training samples are concerned: Existing algorithms mostly aim to increase the between-class separation and preserve the local structure, however, only for the training data. Meanwhile, the important question is how well these algorithms generalize to test data. This question is even more critical for nonlinear dimensionality reduction methods, as the classification performance of test data will not only depend on the properties of the embedding of the training data, but also on the properties of the interpolator that extends the embedding to the whole space. Several methods have been proposed to solve the out-of-sample extension problem, such as unsupervised generalizations with smooth functions [16–19] or semi-supervised interpolators [20]. These methods intend to generalize an already computed embedding to new data and are constrained by the initially prescribed coordinates for training data. Meanwhile, the best strategy for achieving satisfactory generalization to test data would be to learn the embedding and the interpolator not sequentially, but rather in a joint and coherent manner.

In this work, we propose a nonlinear supervised manifold learning method for classification where the embeddings of training data are learned and optimized in a joint way along with the interpolator that extends the embedding to the whole ambient space. A distinctive property of our method is the fact that it explicitly aims to have good generalization to test data in the learning objective. In order to achieve this, we build on the previous work [21] where a theoretical analysis of supervised manifold learning is proposed. The theoretical results in [21] show that for good classification performance, the separation between different classes in the embedding of training data needs to be sufficiently high, while at the same time the interpolation function that extends the embedding to test data must be sufficiently regular. For good generalization to initially unavailable test samples, a compromise needs to be found between these two important criteria. In this work, we adopt radial basis function interpolators for the generalization of the embedding, and learn the embedding of the training data and the parameters of the interpolator, i.e., the coefficients and the scale parameter of the interpolation function, at the same time with a joint optimization algorithm. The analysis in [21] characterizes the regularity of an interpolator via its Lipschitz regularity. We first derive an upper bound on the Lipschitz constant of the interpolator in terms of the parameters of the embedding. Then, relying on the theoretical analysis in [21], we propose to optimize an objective function that maximizes the separation between different classes and preserves the local geometry of training samples, while at the same time minimizing an upper bound on the Lipschitz constant of the RBF interpolator. We propose an alternating iterative optimization scheme that first updates the embedding coordinates, and then the interpolator parameters in each iteration. We test the classification performance of the proposed method on several real data sets and show that it outperforms the supervised manifold learning methods in comparison and traditional classifiers.

Our contributions with respect to previous works are the following:

- The generalization capability of the classifier resulting from a nonlinear supervised embedding is considered during the learning of the embedding for the first time.
- An embedding along with a continuous interpolator is learnt with an optimization objective based on recent theoretical results on the performance of supervised manifold learning methods.
- We show that enforcing the Lipschitz regularity of the interpolator function in addition to the separation between the different classes improves the accuracy of the classifier in most experimental settings.

The rest of the paper is organized as follows. In Section 2, we overview the related work. In Section 3, we review the recent theoretical results that motivate our method and in Section 4, we formulate the supervised manifold learning problem and present the proposed algorithm. In Section 5, we present results on several face and object data sets. Finally, we conclude in Section 6.

2. Related work

2.1. Unsupervised manifold learning

Manifold learning algorithms aim to compute a low-dimensional representation of data that is coherent with its intrinsic geometry, which is characterized in several different ways via geodesic distances [1], locally linear representations [2], second order characteristics [5], and graph spectral decompositions [3], [4] in previous works. When the underlying manifold model is not analytically known, it is common to represent data with a graph model. Given a set of data samples $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, most manifold learning methods build a data graph such that two samples x_i and x_j are linked with an edge when they are nearest neighbors of each other ($x_i \sim x_j$). The edge weights w_{ij} are typically assigned with respect to a similarity measure between neighboring samples.

Denoting as W the weight matrix containing the edge weights w_{ij}, and defining the diagonal degree matrix D with the i-th diagonal entry given by $d(i) = \sum_{x_j \sim x_i} w_{ij}$, the graph Laplacian matrix is defined as $L = D - W$. The Laplacian eigenmaps algorithm [3] maps each data sample $x_i \in \mathbb{R}^n$ to a sample $y_i \in \mathbb{R}^d$ such that the following optimization problem is solved

$$\min_Y \operatorname{tr}(Y^T L Y) = \min_Y \sum_{i \sim j} \|y_i - y_j\|^2 w_{ij}, \ \text{s.t. } Y^T Y = I \tag{1}$$

where $Y = [y_1\ y_2\ \dots\ y_N]^T$ is the data matrix consisting of the coordinates to be learned and I is the identity matrix. Hence, the Laplacian eigenmaps algorithm formulates the new coordinates of data as the functions that have the slowest variation on the data graph, so that neighboring samples in the original domain are mapped to nearby coordinates in the new domain of embedding. The locality preserving projections (LPP) [4] algorithm has the same objective; however, the new coordinates $y_i = P^T x_i$ are constrained to be given by a linear projection of the original coordinates.

2.2. Supervised manifold learning

Many supervised manifold learning algorithms have been proposed in the recent years, most of which are extensions of the Laplacian eigenmaps method. These methods seek to embed data into new coordinates such that neighboring samples in the same class are mapped to nearby coordinates, while samples from different classes are mapped to distant points. This is often represented as an objective function that minimizes $\operatorname{tr}(Y^T L_w Y)$ while maximizing $\operatorname{tr}(Y^T L_b Y)$, where L_w and L_b are the within-class and between-class Laplacian matrices, derived respectively from the within-class and between-class weight matrices W_w and W_b. The between-class edges in W_b can be set with respect to different strategies in different methods. The supervised dimensionality reduction problem is formulated in [15] as

$$\min_Y \operatorname{tr}(Y^T L_w Y) - \mu \operatorname{tr}(Y^T L_b Y) \ \text{subject to } Y^T Y = I \tag{2}$$

where $\mu > 0$ is a constant that adjusts the weight between the structure preservation and the class-aware discrimination terms.

Similar formulations are adopted in [8,22]; however, under the linear projection constraint $y_i = P^T x_i$. The recent work in [23] is based on a similar objective, which also exploits subclass information by identifying favorable data connections within the same class. A local adaptation of the Fisher discriminant analysis is proposed in [7]. A projection matrix P is sought so that an objective of the form $\mathrm{tr}((P^T S_w P)^{-1} P^T S_b P)$ is maximized over P, where S_w and S_b are the within-class and between-class scatter matrices obtained with the edge weights of samples on the data graph. The methods in [9,10,12,13,24] also optimize a similar Fisher-like objective by maximizing the between-class local scatter and minimizing the within-class local scatter. The method in [25] proposes a scatter discriminant analysis to learn embeddings of local image descriptors. In another recent work [26], the optimization of within- and between-class local scatters is formulated via ℓ_1-norms for robustness against image degradations.

Several supervised linear dimensionality reduction methods are based on preserving locally linear representations of data. The algorithm in [27] provides a supervised extension of the well-known LLE method [2] by introducing a label-dependent distance function; however, it is a nonlinear method without an explicit consideration of the out-of-sample problem. The Neighborhood Preserving Discriminant Embedding method presented in [28] is a linear dimensionality reduction method extending the unsupervised NPE method [29] based on locally linear representations. The Hybrid Manifold Embedding method [30] computes a locally linear but globally nonlinear mapping function by first grouping the data into local subsets via geodesic clustering and then learning a supervised embedding of each cluster. The supervised dimensionality reduction method in [31] partitions the manifold into local regions and takes into account the variation of the embedding along tangent directions of the manifold.

2.3. Continuous embeddings via nonlinear functions

The vast majority of supervised dimensionality reduction methods relies on linear projections, and the methods computing a continuous supervised nonlinear embedding are less common. The generalization of the embedding of a given set of training samples to the whole space via continuous interpolation functions is known as the out-of-sample extension problem. The out-of-sample problem is of critical importance especially for nonlinear supervised manifold learning methods computing a pointwise embedding only at training samples.

The Nyström method [16] proposes an out-of-sample solution for unsupervised manifold learning algorithms that embed training samples to coordinates computed from the eigenvectors of a symmetric similarity matrix M, such that the entries of the similarity matrix are obtained from a kernel function K as $M_{ij} = K(x_i, x_j)$. Let $Y = [y_1 \; y_2 \; \dots \; y_N]^T$ be the matrix consisting of the embeddings $y_i \in \mathbb{R}^d$ of the training samples $x_i \in \mathbb{R}^n$, such that the k-th column Y_k of Y is the k-th eigenvector of M with $MY_k = \lambda_k Y_k$. Then, the Nyström method maps a previously unseen test sample x to the point $y(x) = [y^1(x) \; \dots \; y^d(x)]^T$, such that its k-th coordinate is given by $y^k(x) = \frac{1}{\lambda_k} \sum_{i=1}^N Y_{ik} K(x, x_i)$, which is shown to extend the embedding of the training samples to the whole ambient space. The Nyström extension can be applied to many common unsupervised manifold learning algorithms including ISOMAP [1], LLE [2], and Laplacian eigenmaps [3], by suitably identifying a kernel-induced similarity matrix M associated with these methods. However, the Nyström method is often inappropriate for the out-of-sample extension of supervised manifold learning methods, since in this case the similarity matrix M is often class-dependent and can no longer be induced from a unique kernel function as $M_{ij} = K(x_i, x_j)$.

Another possible way to obtain the out-of-sample generalization of an embedding is to employ linear representations. The out-of-sample extension for a test sample x is obtained in [32], by first computing a sparse representation of x in terms of the training samples as $x \approx \sum_{i=1}^N a_i x_i$, where $a = [a_1 \; \dots \; a_N]^T$ is a sparse coefficient vector. Regarding the magnitudes of the sparse coefficients as a measure of similarity between x and the training samples, the embedding y of the test sample x is then obtained as a linear combination of the embeddings y_i of the training samples as $y = (\sum_{i=1}^N |a_i| y_i) / (\sum_{i=1}^N |a_i|)$.

Besides such unsupervised out-of-sample extension methods, the method in [20] proposes a solution for the out-of-sample problem in a semi-supervised setting. Given a data set containing labeled and unlabeled samples, and the embeddings of the labeled samples learnt via any supervised manifold learning algorithm, the method in [20] first computes an RBF out-of-sample interpolator that fits the learnt embedding to the labeled training data. This RBF interpolator is then gradually refined using the unlabeled training samples, such that the RBF interpolator and the estimated class labels of the unlabeled samples are jointly updated in an iterative learning procedure.

The above out-of-sample extension strategies can be coupled with several supervised and unsupervised nonlinear manifold learning algorithms, and can be used to extend priorly learnt embeddings to the whole ambient space. Meanwhile, there also exist nonlinear dimensionality reduction algorithms that learn a specific embedding along with its interpolation function extending the embedding to the whole space. The unsupervised manifold learning method [17] maps the training samples to a lower-dimensional space with a locally linear reconstruction objective as in the LLE algorithm [2]; however, under the constraint that the embedding coordinates be polynomial functions of data samples. The polynomial coefficients are thus optimized to minimize the reconstruction error of the locally linear representation. Previously unseen test data can then be embedded into the new domain via the learnt polynomials. Finally, the method in [33] can be seen as a supervised extension of ISOMAP [1] that also addresses the problem of extension to novel samples. The training samples are first embedded via a modified version of the ISOMAP algorithm by using a supervised distance function that takes the class labels into account. The learnt embedding is then generalized to the whole space via kernel ridge regression.

The focus of our work is essentially different from that of out-of-sample extension algorithms such as [16,20,32], as these methods seek an extension of an already computed embedding to the whole space, while we also address the question of what the embedding should be. Then, compared to manifold learning algorithms such as [17,33] that learn an embedding along with its extension, the main difference of our method is that it explicitly takes into account the performance of the generalization of the learnt classifier to test data, by formulating a supervised learning objective motivated by the recent theoretical generalization bounds of supervised manifold learners.

A possible solution to get around the limitations of linear embeddings while avoiding the out-of-sample problem of nonlinear embeddings is to employ kernel extensions of linear dimensionality reduction methods. The kernel extensions of many well-known dimensionality reduction methods such as PCA, LDA, ICA exist [34–36]. The construction of continuous functions via smooth kernels is also quite common in Reproducing Kernel Hilbert Space (RKHS) methods [37,38]; however, these methods differ from supervised manifold learning methods in that the learnt mapping often represents class labels of data samples rather than their coordinates in a lower-dimensional domain of embedding as in manifold learning. The choice of the kernel type and parameters can be critical in kernel methods. Several previous works in the semi-supervised

learning literature have addressed the learning of kernels by combining known kernels [39,40]. A two-stage multiple kernel learning method is recently proposed in [41] for supervised dimensionality reduction, which finds a nonlinear mapping by optimizing between-class and within-class distances.

3. Theoretical bounds in supervised manifold learning

Nonlinear dimensionality reduction methods in the literature that minimize objectives as in (2) often yield embeddings where training samples from different classes are linearly separable, and the local neighborhoods on the same manifold are preserved as imposed by the term involving the within-class graph Laplacian. On the other hand, most existing methods fail to consider how well these embeddings generalize to new test data: When a test sample of unknown class label is mapped to the low-dimensional domain of embedding via an interpolator or an out-of-sample extension method, what is critical is how likely the test sample is to be correctly classified. This depends both on the coordinates of the embedding for the training samples and the interpolator used to generalize the embedding to the whole ambient space. In the previous work [21], this problem is theoretically studied. In this section, we overview some main results from [21], which will provide a basis for the proposed manifold learning algorithm.

The classification problem is analyzed in [21] in a setting where each data sample in the training set $X = \{x_i\}_{i=1}^N$ is assumed to belong to one of the classes $\{1, 2, \ldots, M\}$ and the samples of each class m are distributed according to the probability measure ν_m. Let $\mathcal{M}_m$ denote the support of the probability measure ν_m. Denoting as $B_\delta(x)$ an open ball of radius δ around a point x

$$B_\delta(x) = \{u \in \mathbb{R}^n : \|x - u\| < \delta\},$$

the following definition introduces the smallest possible measure for a ball $B_\delta(x)$ of radius δ centered around a point in the support $\mathcal{M}_m$ of the m-th class.

$$\eta_{m,\delta} := \inf_{x \in \mathcal{M}_m} \nu_m(B_\delta(x))$$

Next, we recall the definition of Lipschitz continuity for a function f.

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}^d$ is Lipschitz continuous with constant $L > 0$ if for any $u, v \in \mathbb{R}^n$, the inequality $\|f(u) - f(v)\| \leqslant L\|u - v\|$ holds.

The analysis in [21] considers supervised manifold learning algorithms that compute the embedding $y_i \in \mathbb{R}^d$ of each training sample $x_i \in \mathbb{R}^n$. It is assumed that a test sample x of unknown class label is mapped to $\mathbb{R}^d$ via an interpolation function $f : \mathbb{R}^n \to \mathbb{R}^d$. The following main result from [21] gives a bound on the classification error, when the estimate $\hat{C}(x)$ of the class label $C(x)$ of x is estimated via nearest-neighbor classification in $\mathbb{R}^d$ as $\hat{C}(x) = C(x_i)$, where

$$i = \arg\min_j \|y_j - f(x)\|.$$

Theorem 1. *Let $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$ be a set of training samples such that each x_i is drawn i.i.d. from one of the probability measures $\{\nu_m\}_{m=1}^M$, with ν_m denoting the probability measure of the m-th class. Let $Y = \{y_i\}_{i=1}^N$ be an embedding of X in $\mathbb{R}^d$ such that there exist a constant $\gamma > 0$ and a constant A_δ depending on $\delta > 0$ satisfying*

$$\|y_i - y_j\| < A_\delta, \text{ if } \|x_i - x_j\| \leqslant 2\delta \text{ and } C(x_i) = C(x_j)$$
$$\|y_i - y_j\| > \gamma, \quad \text{if } C(x_i) \neq C(x_j).$$

For given $\epsilon > 0$ and $\delta > 0$, let $f : \mathbb{R}^n \to \mathbb{R}^d$ be a Lipschitz-continuous interpolation function with constant L, which maps each x_i to $f(x_i) = y_i$, such that

$$L\delta + \sqrt{d}\epsilon + A_\delta \leqslant \frac{\gamma}{2}. \tag{3}$$

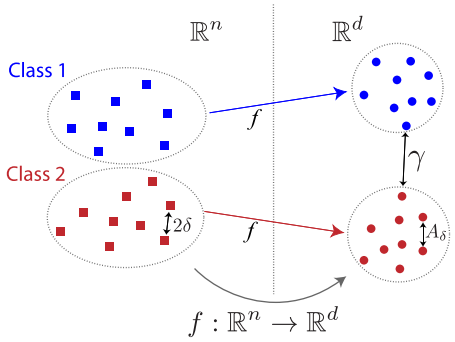


Fig. 1. Illustration of the setting considered in Theorem 1. Samples from the same class having a distance of 2δ are embedded to points at most A_δ apart. Samples from different classes are separated by at least γ.

Consider a test sample x randomly drawn according to the probability measure ν_m of class m. For any $Q > 0$, if X contains at least N_m training samples from the m-th class drawn i.i.d. from ν_m such that $N_m > Q/\eta_{m,\delta}$, then the probability of correctly classifying x with nearest-neighbor classification in $\mathbb{R}^d$ is lower bounded as

$$P\big(\hat{C}(x) = m\big) \geqslant 1 - \exp\left(-\frac{2\,(N_m\,\eta_{m,\delta} - Q)^2}{N_m}\right) - 2\,d\exp\left(-\frac{Q\,\epsilon^2}{2L^2\delta^2}\right). \tag{4}$$

Theorem 1 considers an embedding such that nearby training samples from the same class are mapped to nearby coordinates, while training samples from different classes are separated by a distance of at least γ in the low-dimensional domain of embedding. An illustration of the setting considered in the theorem is given in Fig. 1. The parameter γ can be considered as the separation margin of the embedding. Then for such an embedding, the condition in (3) assumes an interpolator f that is sufficiently regular (with a sufficiently small Lipschitz constant L) compared to the separation margin γ. Finally, a probabilistic classification guarantee is given for this setting in (4), which states that the misclassification probability decreases exponentially with the number of samples under these assumptions. The above result considers the NN classifier in the final stage, which is a simple and widely used classification strategy for which efficient algorithms exist [42]. An extension of this result is also presented in [21] which studies the performance of classification when a linear classifier is used instead of nearest-neighbor classification in the low-dimensional domain. If a linear classifier is used in the domain of embedding, a very similar condition to (3) relating the interpolator regularity to the separation margin is obtained, which yields a similar probabilistic bound on the misclassification error.

While most supervised manifold learning methods in the literature focus on achieving large separation between the training samples from different classes in the embedding, the condition (3) in the above theoretical analysis points to a critical compromise to seek in supervised dimensionality reduction: Achieving high separation between different classes in the training set does not necessarily mean that the classifier will generalize well to test samples. The presence of a sufficiently regular interpolator is furthermore needed, so that the Lipschitz constant L of the interpolator remains below a threshold involving the separation margin γ of the embedding. From this perspective, depending on the data distribution, increasing the separation too much has the risk of forcing the interpolator to be too irregular, which may in turn cause condition (3) to fail. What we propose in this work is to learn the embedding $\{y_i\}_{i=1}^N$ together with the interpolator f in view of condition (3), which is detailed in the next section.

4. Proposed nonlinear supervised smooth embedding method

In this section, we present our proposed supervised dimensionality reduction method. We first formulate the manifold learning problem and define an optimization problem based on the perspectives discussed in Section 3. We then describe our algorithm.

4.1. Formulation of the manifold learning problem

Given training points $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$ from M classes, our purpose is to learn an embedding of data $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^d$ together with a continuous interpolation function $f : \mathbb{R}^n \to \mathbb{R}^d$, such that $f(x_i) = y_i$. The interpolator f will then be used to classify new test points x by mapping x to the low-dimensional domain $\mathbb{R}^d$ as $f(x)$, so that examining $f(x) \in \mathbb{R}^d$ with respect to the embedding Y of the training points with known class labels provides an estimate of the class label of x.

Our method relies on the theoretical results presented in Section 3. Recall from Theorem 1 that, a necessary condition to obtain good generalization performance is

$$L\delta + \sqrt{d}\epsilon + A_\delta \leq \frac{\gamma}{2}.$$

In the sequel, we formulate a manifold learning problem in view of this condition, whose purpose is to make the Lipschitz constant L of the interpolator and the distance A_δ between neighboring points from the same class as small as possible, while making the separation γ between different classes as large as possible, in order to increase the chances that the above condition be met.

Let $f(x) = [f^1(x) \ \dots \ f^d(x)] \in \mathbb{R}^d$, where $f^k(x)$ is the k-th dimension of $f(x)$, with $f^k : \mathbb{R}^n \to \mathbb{R}$. We propose to choose the function f as a radial basis function (RBF) interpolator, as RBF interpolators are a well-studied family of functions [43,44] with many desirable properties such as smoothness and adjustable spread around anchor points. Hence, each component f^k of f is of the form

$$f^k(x) = \sum_{i=1}^{N} c_i^k \phi(\|x - x_i\|) \tag{5}$$

where $\phi : \mathbb{R} \to \mathbb{R}^+$ is an RBF kernel, c_i^k are the coefficients, and x_i are the kernel centers. A common choice for the RBF kernel is the Gaussian kernel $\phi(r) = e^{-r^2/\sigma^2}$, which we also adopt in this work. Under this setting, we now examine our three entities of interest, namely the regularity of the interpolator, the distance between neighboring points from the same class and the separation between different classes.

Interpolator regularity. We begin with proposing a Lipschitz constant for f in terms of the function parameters.

Proposition 1. *Let $L_\phi := \sqrt{2}e^{-\frac{1}{2}}\sigma^{-1}$ and let C be the matrix consisting of the RBF coefficients such that $C_{ij} = c_i^j$. Then the RBF interpolator f satisfies for all $u, v \in \mathbb{R}^n$ the inequality $\|f(u) - f(v)\| \leqslant L\|u - v\|$, where $L := \sqrt{N}L_\phi\|C\|_F$.*

The proof of Proposition 1 is available in the accompanying technical report [45] where more details about our work can be found. When learning an interpolator, we would like to minimize the Lipschitz constant $L = \sqrt{N}L_\phi\|C\|_F$ of $f(x)$. From the form (5) of the interpolator components and the fact that the interpolator values at training points must correspond to the coordinates of the embedding $y_i = f(x_i)$, we get the relation $\Psi C = Y$, where $\Psi \in \mathbb{R}^{N\times N}$ is the matrix consisting of the values of the RBF kernels with $\Psi_{ij} = \phi(\|x_i - x_j\|)$ and $Y = [y_1 \ y_2 \ \dots \ y_N]^T \in \mathbb{R}^{N\times d}$ is the matrix consisting of the coordinates of the embeddings of the training samples. Then the coefficient matrix is given by $C = \Psi^{-1}Y$, so that

$$\|C\|_F^2 = \|\Psi^{-1}Y\|_F^2 = \mathrm{tr}(Y^T\Psi^{-2}Y). \tag{6}$$

In order to keep the Lipschitz constant $L = \sqrt{N}L_\phi\|C\|_F$ of the interpolator small, we need to keep both the Lipschitz constant L_ϕ of the Gaussian kernel and the norm $\|C\|_F$ of the coefficient matrix small. Using the expression of $\|C\|_F^2$ in (6) and recalling that $L_\phi = \sqrt{2}e^{-\frac{1}{2}}\sigma^{-1}$, we thus propose to minimize the following objective for controlling the interpolator regularity

$$\min_{Y,\sigma} \ \mathrm{tr}(Y^T\Psi^{-2}Y) + \frac{\mu}{\sigma^2} \tag{7}$$

where μ is a weight parameter. The objective is chosen proportionally to the squares of the terms $\|C\|_F$ and L_ϕ instead of themselves, due to the convenience of the analytical expression obtained for $\|C\|_F^2$ in (6).

Distance between neighboring points from the same class. Recall from Theorem 1 that the condition (3) required for good classification performance enforces the term A_δ to be sufficiently small, where A_δ is an upper bound on the distance between the embeddings of nearby samples; i.e., $\|y_i - y_j\| < A_\delta$ whenever $\|x_i - x_j\| \leqslant 2\delta$. It is not easy to study the distance $\|y_i - y_j\|$ in relation with the ambient space distance $\|x_i - x_j\|$ for each pair of samples x_i, x_j. Nevertheless, we adopt a constructive solution here and relax this problem to the minimization of the distance between the embeddings of nearby points from the same class. The total distance between the embeddings of neighboring points from the same class, weighted by the edge weights, is given by

$$\sum_{x_i,x_j:\ C(x_i)=C(x_j)} \|y_i - y_j\|^2 w_{ij} = \mathrm{tr}(Y^T L_w Y).$$

Here $L_w = D_w - W_w$ is the within-class Laplacian matrix associated with the within-class weight matrix W_w, where D_w is the diagonal degree matrix whose entries are given by $(D_w)_{ii} = \sum_j (W_w)_{ij}$. The within-class weight matrix W_w contains the weights w_{ij} of the edges between each pair of neighboring samples $x_i \sim x_j$ from the same class. A common choice for assigning the edge weights is the Gaussian kernel, in which case the matrix W_w is of the form

$$(W_w)_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\beta}}, & \text{if } C(x_i) = C(x_j), \quad x_i \sim x_j \\ 0, & \text{otherwise.} \end{cases}$$

Then, the objective

$$\min_{Y} \ \mathrm{tr}(Y^T L_w Y) \tag{8}$$

used in several previous works is an appropriate choice for our purpose.

Separation between samples from different classes. The last entity to be examined in view of the condition (3) is the separation margin γ. In order to satisfy the condition (3), the separation between the samples from different classes must be sufficiently high. Although the margin γ stands for a lower bound for the distance $\|y_i - y_j\|$ between any pair of samples from different classes in Theorem 1, the examination of the minimum value of $\|y_i - y_j\|$ for all pairs of samples is a relatively hard problem. We propose to relax this and evaluate the total distance between the embeddings of different-class samples. Hence, in order to increase the separation margin γ, we propose to maximize

$$\sum_{C(x_i)\neq C(x_j)} \|y_i - y_j\|^2 = \mathrm{tr}(Y^T L_b Y)$$

where W_b is a between-class weight matrix given by $(W_b)_{ij} = 1$ if $C(x_i) \neq C(x_j)$, and $(W_b)_{ij} = 0$ if $C(x_i) = C(x_j)$. The diagonal between-class degree matrix is defined as $(D_b)_{ii} = \sum_j (W_b)_{ij}$, and $L_b = D_b - W_b$ is the corresponding between-class Laplacian matrix. Thus, the maximization of the separation margin is represented by the objective function

$$\max_{Y} \mathrm{tr}\left(Y^T L_b Y\right). \tag{9}$$

Overall optimization problem. Now, bringing together the objective functions presented in (7)–(9), we propose to solve the following optimization problem in order to learn an embedding Y together with its corresponding interpolator:

$$\min_{Y,\sigma} \operatorname{tr}\left(Y^T L_w Y\right) - \mu_1 \operatorname{tr}\left(Y^T L_b Y\right) + \mu_2 \operatorname{tr}\left(Y^T \Psi^{-2} Y\right) + \frac{\mu_3}{\sigma^2}, \quad \text{s.t. } Y^T Y = I \tag{10}$$

Here μ_1, μ_2, and μ_3 are positive weights that balance the different terms in the objective function, and the normalization condition $Y^T Y = I$ is imposed in order to prevent solutions with arbitrarily small embedding coordinates that might trivially minimize the objective.

4.2. Proposed manifold learning algorithm

The proposed objective function (10) can be made convex with respect to Y if the weight parameters μ_1 and μ_2 are suitably chosen; however, it is not jointly convex with respect to both optimization variables Y and σ. We thus propose to minimize (10) with an alternating iterative optimization algorithm. In each iteration, we first fix the scale parameter σ and optimize the embedding coordinates Y, which is then followed by fixing Y and optimizing σ.

Optimization of Y. When the scale parameter σ is fixed, the minimization of the objective (10) is equivalent to the following optimization problem

$$\begin{aligned} Y^* &= \arg\min_Y \ \operatorname{tr}(Y^T L_w Y) - \mu_1 \operatorname{tr}(Y^T L_b Y) + \mu_2 \operatorname{tr}(Y^T \Psi^{-2} Y) \ \text{ s.t. } Y^T Y = I \\ &= \arg\min_Y \ \operatorname{tr}\left(Y^T (L_w - \mu_1 L_b + \mu_2 \Psi^{-2}) Y\right) \ \text{ s.t. } Y^T Y = I. \end{aligned} \tag{11}$$

The solution to this problem is given by the $N \times d$ matrix Y^* whose k-th column consists of the eigenvector of the matrix

$$A = L_w - \mu_1 L_b + \mu_2 \Psi^{-2} \tag{12}$$

that corresponds to its k-th smallest eigenvalue, for $k = 1, \ldots, d$.

Optimization of σ. Note that the dependence of the objective function (10) on the scale parameter σ is through its third term $\mu_2 \operatorname{tr}(Y^T \Psi^{-2} Y)$ and fourth term μ_3/σ^2. Hence, when the embedding Y is fixed, the optimization of the objective is reduced to the problem

$$\sigma^* = \arg\min_\sigma \ \ \mu_2 \operatorname{tr}(Y^T \Psi^{-2} Y) + \frac{\mu_3}{\sigma^2}. \tag{13}$$

The objective in (13) is not a convex function of σ in general. Nevertheless, a useful observation is the following: As the entries of the matrix Ψ consist of the RBF kernel terms $\phi(\|x_i - x_j\|) = \exp\left(-\|x_i - x_j\|^2/\sigma^2\right)$, the matrix Ψ and its inverse Ψ^{-1} have poor conditioning when σ takes arbitrarily large values. Hence, the first term $\operatorname{tr}(Y^T \Psi^{-2} Y)$ in (13) increases with increasing large values of σ. On the other hand, the term σ^{-2} approaches infinity as σ approaches 0. These observations imply that that there exists a positive kernel scale $\sigma^* > 0$ that minimizes the objective (13). As the problem (13) requires the optimization of a single parameter σ, an optimal value σ^* can be computed easily. In practice, we find σ^* via an exhaustive search procedure, by computing the objective (13) over a sufficiently dense sampling of the σ values within a suitably chosen interval $[\sigma_{\min}, \sigma_{\max}]$ of typical scale parameters. The optimal parameter σ^* is then taken as the σ value at which the objective is minimum.

These steps for the alternating optimization of Y and σ are applied successively until the stabilization of the objective function. Note that if μ_1 is chosen sufficiently small to make the matrix A in (12) positive semi definite, the overall objective function (10) is positive. In this case, since both of the alternating optimization steps in (11) and (13) bring updates that cannot increase the objective function in each iteration, being bounded from below, the objective function is guaranteed to converge.

Once the embedding Y of the training points and the kernel scale σ are computed in this way, the interpolator f is simply obtained as in (5) by computing the coefficients as $C = \Psi^{-1} Y$. We call the proposed method Nonlinear Supervised Smooth Embedding (NSSE) and give its description in Algorithm 1.

Algorithm 1 Nonlinear supervised smooth embedding (NSSE).

1: **Input:**
   $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$: Training samples with known class labels
   d: Embedding dimension
   μ_1, μ_2, μ_3: Weight parameters
2: **Initialization:** Set kernel scale σ^* to a typical positive value.
3: **repeat**
4: Fix $\sigma = \sigma^*$ and optimize Y by solving$Y^* = \arg\min_Y \operatorname{tr}\left(Y^T (L_w - \mu_1 L_b + \mu_2 \Psi^{-2}) Y\right)$ s.t. $Y^T Y = I$
5: Fix $Y = Y^*$ and optimize σ by solving$\sigma^* = \arg\min_\sigma \ \mu_2 \operatorname{tr}(Y^T \Psi^{-2} Y) + \mu_3 \sigma^{-2}$
6: **until** Objective function in (10) is stabilized
7: Compute interpolator coefficients as $C = \Psi^{-1} Y$.
8: **Output**:
   $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^d$: Embedding of training samples
   $f : \mathbb{R}^n \to \mathbb{R}^d$: Interpolation function

4.3. Complexity of the proposed algorithm

We now analyze the computational complexity of the proposed NSSE method. The algorithm is composed of three main stages, which are the initialization stage (calculation of the L_w and L_b matrices), the main loop between steps 3 and 6 of Algorithm 1, and the finalization stage in step 7.

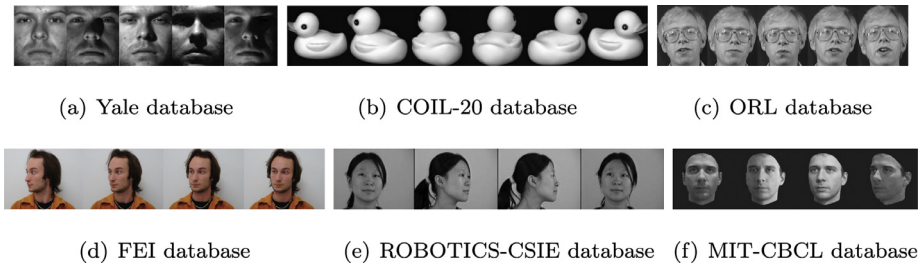
In the initialization step, the complexity of the computation of L_w and L_b is mainly determined by the complexity of computing the within-class and between-class weight matrices W_w and W_b, which is of $O(nN^2)$.

We next consider the main loop of the algorithm. The matrix Ψ in step 4 can be calculated with complexity $O(nN^2)$ and it is inverted with complexity $O(N^3)$ to obtain Ψ^{-1}. As a result, the computation of Ψ^{-2} is of complexity $O(nN^2) + O(N^3)$. In order to find Y^* in step 4, the eigenvectors of $L_w - \mu_1 L_b + \mu_2 \Psi^{-2}$ should be found, which is of complexity $O(N^3)$. Then, the total complexity of step 4 is $O(nN^2) + O(N^3)$. In step 5, the expression $\mu_2 \operatorname{tr}(Y^T \Psi^{-2} Y) + \mu_3 \sigma^{-2}$ must be computed repeatedly to find σ^*, which is of complexity $O(N^3)$. Hence, the complexity of the main loop is found as $O(nN^2) + O(N^3)$.

In step 7, the complexity of the calculation of Ψ^{-1} is $O(N^3)$, and the matrix product $\Psi^{-1} Y$ is of complexity $O(dN^2)$. We may assume $d \ll N$, which then gives the complexity of step 7 as of $O(N^3)$. Combining this with the previous stages, the overall complexity of the algorithm is found as $O(nN^2) + O(N^3)$.

5. Experimental results

In this section, we evaluate the performance of the proposed NSSE method on six real data sets. We first describe the data sets, then study the iterative optimization procedure employed in the proposed method, and then compare the performance of NSSE with that of other supervised manifold learning algorithms and traditional classifiers.



(a) Yale database (b) COIL-20 database (c) ORL database

(d) FEI database (e) ROBOTICS-CSIE database (f) MIT-CBCL database

Fig. 2. Sample images from one class of the used databases.

5.1. Data sets and experimentation setting

We experiment on the data sets listed below. Some sample images from one class of each data set are presented in Fig. 2.

Yale Face Database. The data set consists of 2242 greyscale face images of 38 different subjects, where each subject has 59 images [46]. All images are taken from a single viewpoint with variations in the lighting angles and rates.

COIL-20 Database. The Columbia Object Image Library database consists of 1440 grayscale images of 20 different objects, where each object has 72 images captured by rotation increments of 5 degrees [47].

ORL Database. The database consists of a total of 400 images, with 10 images of each one of the 40 subjects taken in an upright, frontal position [48]. The images contain variations in the lighting, facial expressions and facial details such as glasses.

FEI Database. The FEI database is a face database containing a total of 2800 images, with 14 images for each one of the 200 subjects taken in an upright frontal position with profile rotation of up to about 180 degrees and scale variation of about 10% [49]. We experiment on 50 classes from this database.

ROBOTICS-CSIE Database. The database contains a total of 3330 face images of 90 subjects, with 37 images for each subject captured under rotation increments of 5 degrees [50]. We experiment on 40 classes from this database.

MIT-CBCL Database. The database contains face images of 10 subjects [51]. We experiment on a total of 5240 images, with 524 images per subject captured under rotations of up to 30 degrees and varying illumination conditions.

We experiment on greyscale versions of the images resized to around 25×25 pixels. All experiments are conducted in a supervised setup, by randomly separating the images into a training set and a test set in each repetition of the experiment. In all experiments, the proposed NSSE algorithm is evaluated in a setting where the training images are used to learn a continuous embedding into a low-dimensional domain. The test images are then mapped to the domain of embedding via the learnt interpolator and their class labels are estimated via nearest neighbor classification in the low-dimensional domain. The graph edge weights are set with a Gaussian kernel. In all experiments, the weight parameters μ_1, μ_2, and μ_3 of NSSE are set with cross-validation. The weight parameters are set sequentially, by first initializing them with some typical values and then optimizing one of them at a time via cross validation where the others are kept fixed. When optimizing one weight parameter, the training samples are divided randomly into two sets as training and validation, the algorithm is trained on the training set, and the classification error is measured on the validation set for different values of the weight parameter. We repeat this several times by randomly assigning the training and the validation set, and then finally select the parameter value that gives the smallest average classification error on the validation set. In practice, we have observed that the typical ranges of appropriate μ_1, μ_2, and μ_3 values do not usually vary dramatically between different data sets and setting these parameters to values within the intervals $\mu_1 \in [100, 1000]$, $\mu_2 \in [0.0001, 0.001]$, and $\mu_3 \in [1, 5]$ often yields satisfactory performance.

5.2. Study of the iterative optimization procedure

In this first experiment, we study the iterative optimization procedure employed in the proposed method. As discussed in Section 4.2, the NSSE algorithm follows an alternating optimization scheme by minimizing the objective function in (10) first with respect to the embedding Y of the training samples, and then the scale parameter σ of the RBF kernels.

The results given in Fig. 3 are obtained on the FEI face data set, where an embedding into a $d = 10$ dimensional domain is computed using a total of 100 training samples. Fig. 3(a) shows the variation of the objective function in (10) throughout the iterations. Although the proposed alternating optimization procedure is not theoretically guaranteed to find the global optimum of the objective, it is observed from the figure that the proposed scheme can effectively minimize the objective function, which converges in a small number of iterations. The misclassification rates of the test images in percentage are reported in Fig. 3(b) obtained with the embeddings and interpolators computed in each iteration. The results show that the progressive update of the continuous embedding throughout the iterations improves the classification performance. The comparison of the plots in Figs. 3(a) and (b) reveals that the variations of the objective function and the misclassification rate throughout the iterations are quite similar. This suggests that the choice of the objective function in (10), motivated by theoretical bounds, indeed matches the actual classification error. Fig. 3(c) shows the evolution of the RBF kernel scale parameter σ throughout the iterations. The RBF kernel scale σ is deliberately initialized with a too high value in this experiment in order to study the effect of the initial conditions on the algorithm performance. Despite the initialization of σ with a too large value, the iterative minimization of the objective gradually pulls the kernel scale towards a favorable value that improves the classification performance.

The same experiment is also repeated by initializing the RBF kernel scale this time with a small value, whose results are given in the lower row of Fig. 3. It is observed in Fig. 3(f) that the RBF scale σ is effectively optimized throughout the iterations towards a larger value, which gradually decreases the objective function and improves the classification accuracy in Figs. 3(d) and 3(e). These results suggest that the algorithm performance is not affected much by the initialization of the RBF kernel scale. We have obtained similar results on the other data sets and under different choices of the parameters such as the number of training samples, which we skip here for brevity.

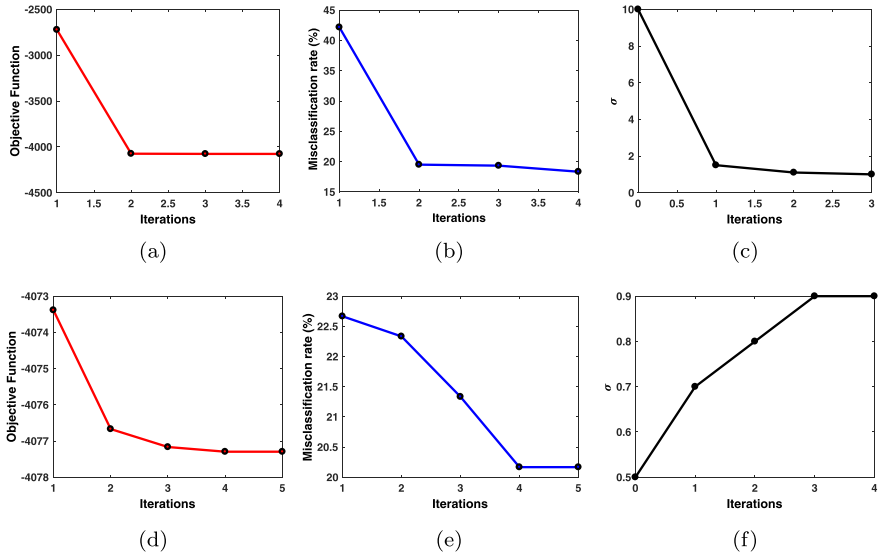


Fig. 3. Algorithm performance throughout iterations. Results in the upper and lower row are respectively obtained by initializing the algorithm with a high and a low RBF kernel scale.

5.3. Variation of the classification performance with the embedding dimension

We now study the classification performance of the proposed algorithm in relation with the dimension d of the embedding. The proposed NSSE method is compared to some other dimensionality reduction algorithms listed below.

- The Supervised Laplacian Eigenmaps (SUPLAP) method proposed in [15] computes a nonlinear low-dimensional embedding of the training samples by minimizing the objective in (2). We extend the embedding of the training samples given by the SUPLAP method to the whole space via an RBF interpolator of the same form as in NSSE. We then embed the test samples into the low-dimensional domain with this interpolation function.
- The Local Fisher Discriminant Analysis (LFDA) method proposed in [7] is a supervised manifold learning algorithm computing a linear embedding with a Fisher-type cost with additional locality preservation objectives.
- The Local Discriminant Embedding method (LDE) [22] is a manifold learning method that optimizes a similar objective as in the SUPLAP method; however, learns a linear projection.
- Linear Discriminant Analysis (LDA) is a classical dimensionality reduction technique that maximizes the between-class scatter while minimizing the within-class scatter.

The dimensionality reduction methods are applied on training samples to compute a d-dimensional embedding, which is then used to classify test samples via nearest neighbor classification in the domain of embedding. The algorithms are evaluated for a range of d values. The parameters of the other methods in comparison are adjusted to attain their best performance.

The variation of the misclassification rates of test samples in percentage with the dimension d of the embedding is presented in Figs. 4 and 5. The results are the average of 20 random realizations of the experiments with different training and test sets, with 10, 10, 2, 2, 7, and 10 training images per class (chosen proportionally to the total number of samples) respectively for the Yale, COIL-20, ORL, FEI, ROBOTICS-CSIE and the MIT-CBCL databases. Most of the tested methods are based on solving a generalized eigenvalue problem and the rank of the involved matrices may be different for each method depending on the number of training samples and the number of classes. Hence, the maximum possible dimension of the embedding may vary between different methods, as well as the best range of dimensions where the methods perform well. For this reason, the results on each data set are presented in two figures with different d ranges for better visual clarity.

The results in Figs. 4 and 5 show that the classification accuracy of the proposed NSSE algorithm compares quite favorably to those of the other methods, as NSSE often yields the smallest misclassification rate at the optimal dimension. The misclassification rate of LDA is observed to decrease monotonically with the dimension d and its best performance is attained when d reaches the number of classes. The LDE and LFDA algorithms exhibit their best performances at much higher dimensions compared to the other algorithms. The error rates of these algorithms usually decrease as the embedding dimension increases; however, in some datasets a local optimum for d can also be observed.

Among all methods, the nonlinear NSSE and SUPLAP methods often perform better than the linear LDA, LFDA, and LDE methods. This shows that the flexibility of nonlinear methods when learning an embedding is likely to bring an advantage in computing better representations for data. It is then interesting to compare the performances of the two nonlinear methods; NSSE and SUPLAP. The SUPLAP algorithm attains its best performance when the dimension d of the embedding is close to the number of classes, while the optimum value of d for the proposed NSSE algorithm is smaller in most data sets. Interestingly, the optimal dimension of NSSE is much smaller than that of SUPLAP in data sets with a low intrinsic dimension such as COIL-20, FEI, and ROBOTICS-CSIE, which are generated by the variation of only one or two camera angle parameters. Similarly, in data sets of larger intrinsic dimension such as MIT-CBCL due to several pose and lighting parameters, the optimal dimension of NSSE is higher and closer to that of SUPLAP. This may suggest that the embedding computed with NSSE tries to capture the intrinsic geometry of data and provides a better representation when the embedding dimension is chosen proportionally to the intrinsic dimension of data.

The reduction of the embedding dimension is desirable especially regarding the complexity of the classification of test samples in a practical application. Another advantage of NSSE over SUPLAP is that NSSE is less sensitive to the choice of the dimension, as the misclassification performance is less affected for non-optimal values of d. Such benefits of the proposed NSSE algorithm mainly

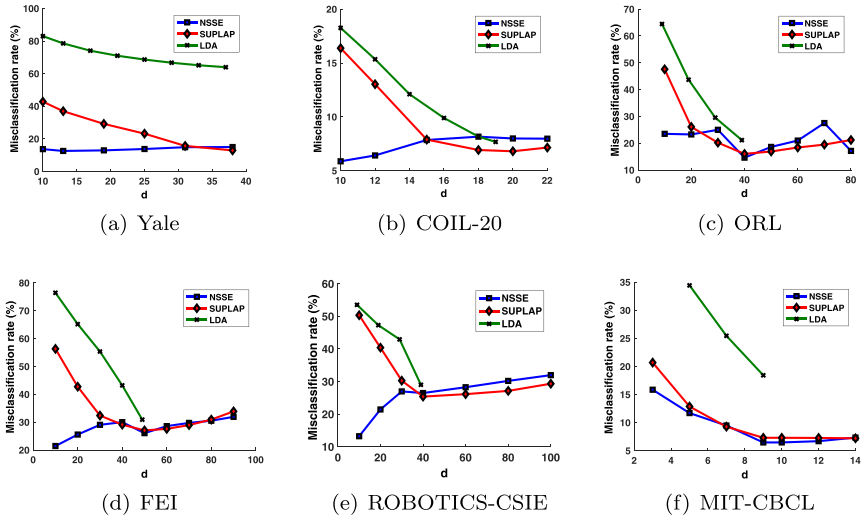


Fig. 4. Variation of the misclassification rates of the NSSE, SUPLAP and LDA methods with the embedding dimension in various data sets.

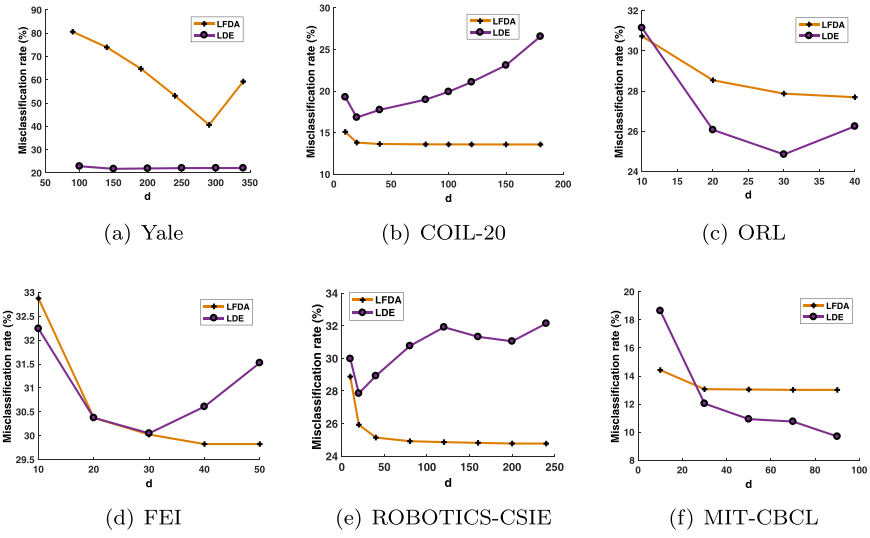


Fig. 5. Variation of the misclassification rates of the LFDA and LDE methods with the embedding dimension in various data sets.

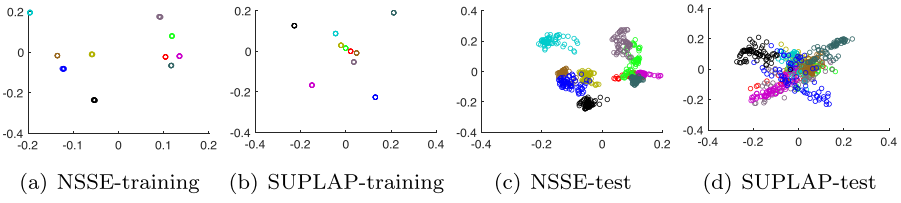


Fig. 6. Visual comparison of the embeddings given by the NSSE and SUPLAP algorithms.

result from the fact that the Lipschitz continuity of the interpolator is imposed in the learning objective. Consequently, the training samples are embedded more evenly in the low-dimensional space so as to allow the construction of a regular interpolator, which in return reduces the required number of dimensions or the sensitivity to the non-optimal choice of d.

In fact, Fig. 6 provides a visual comparison of the embeddings obtained with the NSSE and the SUPLAP algorithms. Panels (a) and (b) show the two-dimensional embeddings of 70 training samples from 10 classes of the ROBOTICS-CSIE data set, respectively with NSSE and SUPLAP. The embeddings of training samples look similar between the two methods, although different classes are more regularly spaced in NSSE. The performance difference between these two methods becomes much clearer when the embeddings of the test samples in panels (c) and (d) are observed. Even at this very small embedding dimension of 2, NSSE separates test samples from different classes much more successfully than SUPLAP, which is due to the inclusion of the interpolator parameters in the learning objective in order to attain good generalization performance.

5.4. Overall comparison with several classification methods

We now provide an overall comparison of the proposed NSSE method with baseline classifiers and other manifold learning methods. In addition to the supervised manifold learning algorithms used in the experiments of Section 5.3, we compare NSSE

Table 1
Misclassification rates (%) of compared methods on Yale database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
6	22.09	23.10	29.43	63.48	26.89	35.86	20.05	**19.47**	63.81	22.58	19.89
10	12.60	12.92	15.17	52.89	40.63	63.97	21.79	**11.88**	53.19	12.58	12.36
15	7.52	7.95	9.09	43.57	10.78	57.87	7.95	7.84	43.41	7.11	**6.90**
20	5.02	5.60	6.14	37.51	7.42	52.80	5.16	6.43	38.31	4.61	**4.50**
30	2.56	2.57	2.99	30.13	3.22	46.43	3.04	4.63	32.46	2.38	**2.35**

Table 2
Misclassification rates (%) of compared methods on COIL-20 database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
7	**8.09**	10.97	10.38	13.90	17.93	11.84	20.86	13.87	13.90	11.17	8.49
10	**4.97**	6.81	6.93	10.22	13.59	7.68	16.84	9.38	10.22	7.07	5.44
15	**2.79**	3.85	4.60	6.88	11.32	4.22	14.01	5.70	6.88	3.96	3.05
20	**1.25**	2.04	3.23	4.51	9.53	2.29	12.64	3.34	4.54	2.00	1.31
30	**0.53**	0.80	2.27	2.31	7.08	0.99	13.28	1.56	2.44	0.79	0.73

Table 3
Misclassification rates (%) of compared methods on ORL database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
2	**14.11**	16.04	19.74	19.34	27.70	21.18	24.92	17.03	19.34	14.81	14.85
3	**8.00**	9.49	10.70	12.96	14.89	13.13	12.74	11.06	12.96	8.63	8.38
5	**3.90**	5.32	4.35	6.92	8.10	7.74	7.05	6.90	6.92	4.23	4.13

Table 4
Misclassification rates (%) of compared methods on FEI database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
2	**20.86**	27.07	35.38	32.13	29.83	30.93	30.05	31.91	32.13	26.42	25.03
4	**8.05**	12.46	12.85	19.45	12.90	12.56	10.80	19.20	19.45	11.06	10.77
7	**5.00**	6.42	9.09	10.86	9.74	5.40	7.77	11.53	11.23	5.03	5.40

Table 5
Misclassification rates (%) of compared methods on ROBOTICS-CSIE database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
7	**13.56**	27.23	23.97	34.46	24.87	29.43	25.13	34.87	34.53	24.29	20.32
14	**4.38**	11.74	8.78	17.80	11.97	14.15	9.74	17.36	17.84	9.04	5.86
21	**2.83**	6.52	4.77	10.09	6.99	10.57	5.88	9.85	9.76	4.81	3.08

Table 6
Misclassification rates (%) of compared methods on MIT-CBCL database.

# Tr	NSSE	SUPLAP	SVM	NN	LFDA	LDA	LDE	NYS	SPE	IsoKRR	MReg
10	**6.48**	7.31	9.91	14.43	12.32	18.44	9.69	15.03	14.43	6.53	6.55
20	**2.49**	3.38	4.18	5.65	8.36	8.38	6.02	6.06	5.66	2.50	2.85
40	0.77	1.22	1.52	1.46	5.29	3.18	2.97	1.84	2.05	**0.71**	0.97

with the SVM classifier in the original domain, the nearest neighbor (NN) classifier in the original domain, the out-of-sample generalization of the Laplacian eigenmaps embedding with the Nyström method (NYS) [16], the out-of-sample generalization of Laplacian eigenmaps with sparse coding (SPE) [32], the IsoKRR method proposed in [33] which computes a supervised nonlinear embedding and generalizes it with kernel ridge regression, and the supervised manifold regularization algorithm (MReg) proposed in [37] based on Reproducing Kernel Hilbert Spaces. The embedding dimensions and other algorithm parameters of the manifold learning methods are set to their optimal values. The classification errors over test samples are studied by varying the training/test ratio and the results are averaged over 20 realizations of the experiments under different random choices of the training and test sets.

The misclassification rates of test samples in percentage are presented for the compared methods for different training data sizes in Tables 1–6 for the tested data sets. The leftmost columns of the tables show the number of training samples per class. Experiments are conducted over a suitable range of number of training samples for each data set, considering the total number of samples in the data set. The smallest classification errors are shown in bold.

The proposed NSSE method is observed to outperform the other methods in most data sets. In Table 1, out-of-sample generalization with the Nyström method NYS [16] is seen to be one of the two best performing methods along with MReg [37], while its performance is behind many others in the other data sets. The extreme illumination changes in the Yale data set lead to degeneracies in the data manifold due to the very high local curvatures and non-differentiability, which seems to pose a challenge for the proposed NSSE method. Meanwhile, the global structure of this data set can in fact be approximated with linear subspace models fairly well, thanks to which an unsupervised out-of-sample extension method such as Nyström achieves good performance on this data. In Tables 2–6, the proposed NSSE method is seen to yield the best

Table 7
Running times of the NSSE algorithm observed for several data sizes on three data sets. The data size stands for the total number of training images.

	COIL-20	FEI	ROBOTICS-CSIE
Data size–Running time	140–.81 sec	100–0.82 sec	280–2.22 sec
	300–1.45 sec	200–1.52 sec	560–8.33 sec
	600–5.45 sec	350–3.74 sec	840–14.99 sec

performance in most settings, and the algorithms closest in performance to NSSE are the nonlinear and supervised SUPLAP, IsoKRR, and the MReg methods. Among the supervised manifold learning algorithms, the nonlinear methods seem to outperform the linear ones in general. The linear manifold learning algorithms LFDA, LDA, and LDE exhibit variable performance depending on the data set. As the performances of the algorithms improve with the increase in the number of training samples, these linear manifold learning methods may get outperformed by the baseline SVM and NN classifiers especially when the number of samples is sufficiently high. The performance gap between NSSE and the other nonlinear and supervised MReg, IsoKRR, and SUPLAP methods is more significant in the FEI and ROBOTICS-CSIE datasets containing a large number of classes, especially when the number of training samples is limited. The lack of training samples compared to the large number of classes is likely to lead to degenerate embeddings in nonlinear methods computing a pointwise embedding as in SUPLAP, while the regularization term enforcing the regularity of the interpolator in NSSE proves effective for the prevention of such degeneracies and ensuring the preservation of the overall geometric structure of data in the embedding. For the particular case of initially few labeled samples, the extension of our study to an active learning framework [52] remains as a potential future direction.

Note that, unlike complex classifiers involving rich models with many parameters to learn, the classifiers obtained with the proposed method consist of a relatively simpler model with fewer parameters to learn. Based on models particularly fit to the priors on the data geometry and dimensionality, the proposed method attains satisfactory classification accuracy on data sets conforming to such low-dimensional models, even when the number of training samples is very limited. The accuracy of the proposed method would inevitably degrade if applied directly to data collections registered under highly uncontrolled settings violating the low-dimensional manifold assumption, e.g., data sets of complex backgrounds, with many different and dissimilar objects belonging to the same class, etc. Nevertheless, the learning of representations that extract the useful and essential information from such data sets registered under challenging conditions is still an open problem. Referring the reader to [53] for a recent comparison of several feature descriptors, we note that the proposed method can potentially be coupled with progressing representation learning techniques that can capture the data geometry invariantly to acquirement conditions.

Finally, we report the observed computation times for jointly learning an embedding and an interpolator with the proposed NSSE algorithm. The running times obtained for a single run of the NSSE algorithm with a non-optimized MATLAB implementation on a laptop computer are given in Table 7 for three data sets, for different data sizes. The observed running times seem to be consistent with the complexity analysis of the method provided in Section 4.3.

6. Conclusion

We have proposed a nonlinear supervised manifold learning method that learns an embedding of the training data jointly with a smooth RBF interpolation function extending the embedding to the whole space. The embedding and the interpolator parameters are jointly optimized with the purpose of good generalization to initially unavailable data, based on recent theoretical results on the performance of supervised manifold learning methods. In particular, the embedding and the RBF paramaters are learnt such that the interpolator has sufficiently good Lipschitz regularity while different classes are separated as much as possible. Experiments have shown that the proposed method often yields better classification performance while requiring a smaller number of dimensions in comparison with other approaches. Thanks to the priors on the Lipschitz regularity of the interpolator, the proposed method can learn efficient representations even under limited availability of training samples, and is relatively robust to conditions such as the non-optimal choice of the embedding dimension and unfavorable initialization. The proposed method can find use in a variety of applications concerning the classification and analysis of data, especially conforming to low dimensional models. The extensions of our study to multi-view or active learning settings remain as possible future directions.

References

[1] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction., Science 290 (5500) (2000) 2319–2323.
[2] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[3] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[4] X. He, P. Niyogi, Locality preserving projections, Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.
[5] D.L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. U.S.A. 100 (10) (2003) 5591–5596.
[6] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, SIAM J. Sci. Comput. 26 (2005) 313–338.
[7] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, J. Mach. Learn. Res. 8 (2007) 1027–1061.
[8] Q. Hua, L. Bai, X. Wang, Y. Liu, Local similarity and diversity preserving discriminant projection for face and handwriting digits recognition., Neurocomputing 86 (2012) 150–157.
[9] W. Yang, C. Sun, L. Zhang, A multi-manifold discriminant analysis method for image feature extraction, Pattern Recognit. 44 (8) (2011) 1649–1657.
[10] Z. Zhang, M. Zhao, T. Chow, Marginal semi-supervised sub-manifold projections with informative constraints for dimensionality reduction and recognition, Neural Netw. 36 (2012) 97–111.
[11] B. Li, J. Liu, Z. Zhao, W. Zhang, Locally linear representation fisher criterion, in: The 2013 International Joint Conference on Neural Networks, 2013, pp. 1–7.
[12] Y. Cui, L. Fan, A novel supervised dimensionality reduction algorithm: graph-based fisher analysis, Pattern Recognit. 45 (4) (2012) 1471–1481.
[13] R. Wang, X. Chen, Manifold discriminant analysis, in: CVPR, 2009, pp. 429–436.
[14] M. Yu, L. Shao, X. Zhen, X. He, Local feature discriminant projection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1908–1914.
[15] B. Raducanu, F. Dornaika, A supervised non-linear dimensionality reduction approach for manifold learning, Pattern Recognit. 45 (6) (2012) 2432–2444.
[16] Y. Bengio, J.F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, M. Ouimet, Out-of-sample extensions for LLE, ISOMAP, MDS, Eigenmaps, and Spectral Clustering, in: Adv. Neural Inf. Process. Syst., 2004, pp. 177–184.
[17] H. Qiao, P. Zhang, D. Wang, B. Zhang, An explicit nonlinear mapping for manifold learning, IEEE T. Cybernetics 43 (1) (2013) 51–63.
[18] G.H. Chen, C. Wachinger, P. Golland, Sparse projections of medical images onto manifolds, in: Proc. 23rd Int. Conf. Inf. Proc. Medical Imag., 2013, pp. 292–303.
[19] B. Peherstorfer, D. Pflüger, H.J. Bungartz, A sparse-grid-based out-of-sample extension for dimensionality reduction and clustering with laplacian eigenmaps, in: Proc. 24th Australasian Joint Conf. Advances in Artificial Intelligence, 2011, pp. 112–121.
[20] E. Vural, C. Guillemot, Out-of-sample generalizations for supervised manifold learning for classification, IEEE Trans. Image Process. 25 (3) (2016) 1410–1424.
[21] E. Vural, C. Guillemot, A study of the classification of low-dimensional data with supervised manifold learning, J. Mach. Learn. Res. 18 (157) (2018) 1–55.
[22] H. Chen, H. Chang, T. Liu, Local discriminant embedding and its variants, in: IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2005, pp. 846–853.
[23] A. Maronidis, A. Tefas, I. Pitas, Subclass graph embedding and a marginal fisher analysis paradigm, Pattern Recognit 48 (12) (2015) 4024–4035.
[24] Q. Gao, J. Ma, H. Zhang, X. Gao, Y. Liu, Stable orthogonal local discriminant embedding for linear dimensionality reduction, IEEE Trans. Image Proc. 22 (7) (2013) 2521–2531.
[25] M. Yu, L. Shao, X. Zhen, X. He, Local feature discriminant projection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1908–1914.

[26] S. Chen, J. Wang, C. Liu, B. Luo, Two-dimensional discriminant locality preserving projection based on l1-norm maximization, Pattern Rec. Let. 87 (2017) 147–154.
[27] S. Zhang, Enhanced supervised locally linear embedding, Pattern Recognit. Lett. 30 (13) (2009) 1208–1218.
[28] Y. Pang, A.T.B. Jin, F.S. Abas, Neighbourhood preserving discriminant embedding in face recognition, J. Visual Com. Image Rep. 20 (8) (2009) 532–542.
[29] X. He, D. Cai, S. Yan, H. Zhang, Neighborhood preserving embedding, in: 10th IEEE Int. Conf. Computer Vision, 2005, pp. 1208–1213.
[30] Y. Liu, Y. Liu, K.C.C. Chan, K.A. Hua, Hybrid manifold embedding, IEEE Trans. Neural Netw. Learning Syst. 25 (12) (2014) 2295–2302.
[31] Y. Zhou, S. Sun, Manifold partition discriminant analysis, IEEE Trans. Cybern. 47 (4) (2017) 830–840.
[32] F. Dornaika, B. Raducanu, Out-of-sample embedding for manifold learning applied to face recognition, in: IEEE Conf. Computer Vision and Pattern Recognition, CVPR Workshop, 2013, pp. 862–868.
[33] C. Orsenigo, C. Vercellis, Kernel ridge regression for out-of-sample mapping in supervised manifold learning, Expert Syst. Appl. 39 (9) (2012) 7757–7762.
[34] B. Schölkopf, A.J. Smola, K.R. Müller, Kernel principal component analysis, in: 7th Int. Conf. Artificial Neural Networks, 1997, pp. 583–588.
[35] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[36] F.R. Bach, M.I. I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (2002) 1–48.
[37] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Lear. Res. 7 (2006) 2399–2434.
[38] N. Aronszajn, Theory of reproducing kernels, Trans. Am. Math. Soc. 68 (3) (1950) 337–404.
[39] G. Dai, D.Y. Yeung, Kernel selection for semi-supervised kernel machines, in: Prof. 24th Int. Conf. Machine Learning, 2007, pp. 185–192.
[40] A. Argyriou, M. Herbster, M. Pontil, Combining graph laplacians for semi-supervised learning, in: Advances in Neural Information Processing Systems 18, 2005, pp. 67–74.
[41] A. Nazarpour, P. Adibi, Two-stage multiple kernel learning for supervised dimensionality reduction, Pattern Recognit. 48 (5) (2015) 1854–1862.
[42] S. Vajda, K.C. Santosh, A fast k-nearest neighbor classifier using unsupervised clustering, in: Int. Conf. Recent Trends in Image Proc. and Pattern Rec., 2016, pp. 185–193.
[43] B.J.C. Baxter, The interpolation theory of radial basis functions Ph.D. thesis, Cambridge University, Trinity College, 1992.
[44] C. Piret, Analytical and numerical advances in radial basis functions Ph.D. thesis, University of Colorado, 2007.
[45] C. Örnek, E. Vural, Nonlinear supervised dimensionality reduction via smooth regular embeddings. [Online]. Available:. arXiv:1710.07120 2018.
[46] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intelligence 23 (6) (2001) 643–660.
[47] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-20), Technical Report, 1996.
[48] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, in: Proc. 2nd IEEE Workshop on Appl. Comp. Vision, 1994, pp. 138–142.
[49] C.E. Thomaz, G.A. Giraldi, A new ranking method for principal components analysis and its application to face image analysis, Image Vis. Comput. 28 (6) (2010) 902–913.
[50] Robotics CSIE database for face detection, Available: http://robotics.csie.ncku.edu.tw/Databases/FaceDetect_PoseEstimate.htm.
[51] MIT-CBCL face recognition database, Available: http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html.
[52] M.R. Bouguelia, S. Nowaczyk, K.C. Santosh, A. Verikas, Agreeing to disagree: active learning with noisy labels without crowdsourcing, Int. J. Mach. Learn. Cybern. 9 (8) (2018) 1307–1319.
[53] S. Candemir, E. Borovikov, K.C. Santosh, S.K. Antani, G.R. Thoma, RSILC: Rotation- and scale-invariant, line-based color-aware descriptor, Image Vis. Comput. 42 (2015) 1–12.

Cem Örnek got his B.S. degree in Electrical and Electronics Engineering in 2014 from Hacettepe University, Ankara, Turkey. He got his M.S. degree in Electrical and Electronics Engineering from Middle East Technical University, Ankara, Turkey in 2018. His research areas are signal processing and machine learning.

Elif Vural got her B.S. degrees in Electrical and Electronics Engineering and in Mathematics in 2006, and her M.S. degree in 2008 from Middle East Technical University (METU) in Turkey. She got her Ph.D. degree from Ecole Polytechnique Fédérale de Lausanne in Switzerland in 2013, after which she worked as a postdoctoral researcher at INRIA Rennes, France until 2015. She has been an assistant professor at METU since 2015. Her research lies in the intersection of signal processing and machine learning, and mainly focuses on the analysis of data with low-dimensional models.