Learning Exemplar-Represented Manifolds in Latent Space for Classification

Shu Kong and Donghui Wang *

College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China
{aimerykong,dhwang}@zju.edu.cn

Abstract. Intrinsic manifold structure of a data collection is valuable information for classification task. By considering the manifold structure in the data set for classification and with the sparse coding framework, we propose an algorithm to: (1) find exemplars from each class to represent the class-specific manifold structure, in which way the object-space dimensionality is reduced; (2) simultaneously learn a latent feature space to make the mapped data more discriminative according to the class-specific manifold measurement. We call the proposed algorithm Exemplar-represented Manifold in Latent Space for Classification (EMLSC). We also present the nonlinear extension of EMLSC based on kernel tricks to deal with highly nonlinear situations. Experiments on synthetic and real-world datasets demonstrate the merit of the proposed method.

Keywords: Sparse Coding, Dimensionality Reduction, Manifold, Exemplar Selection, Classification

1 Introduction

Among various areas of machine learning, information retrieval and signal processing, one needs to deal with high-dimensional data collections ahead of specific tasks, such as classification focused on in this paper. This has motivated a lot of work in dimensionality reduction, whose goal is to find compact representations of the data that can save memory and computational time, and also enhance the performance of algorithms that deal with the data.

Since datasets often consist of high-dimensional data, most dimensionality reduction methods aim at reducing the feature-space dimension for all the data, *e.g.* PCA [1], LLE [2] and Isomap [3], etc. Among these methods, geometrically motivated approaches are shown to be effective in discrovering the geometrical structure in the data. Meanwhile, manifold-based methods, such as LLE and Isomap and their variants, have attracted considerable attention in data analysis [4–6], and have achieved very encouraging performances in clustering, classification and data visualization. However, these methods separate the manifold-motivated dimensionality reduction and classifier learning apart, in which way, further improved classification performance is prevented.

* Corresponding author. This work is supported by Natural Science Foundations (No.61071218) of China and 973 Program (No.2010CB327904).

On the other hand, since datasets usually contain a large number of data, dimensionality reduction in the object space is a desirable solution [7]. This can be achieved either by learning an adaptive dictionary [8, 9] or finding exemplars [7]. Learning a compact dictionary to represent data (see [10] and therein) and the problem of learning a supervised dictionary for classification have been well studied in literature [9, 11]. But such learned dictionaries intrinsically ignore the data manifold structures. Because, the dictionary atoms almost never coincide with the original data [12, 11]. Specifically, for example, the negative sign of some atoms are hard to interpret, and the unit Euclidean length of the atoms means they just act as bases for reconstruction of data points but not for representing them. This intrinsic problem in dictionary learning has also been recognized in [7]. Therefore, the learned dictionary atoms cannot be considered as good representatives for the collection of data points when meeting various tasks such as classification. In contrast, one can find a small subset of the data to appropriately represent the whole data collection owing to the self-expressiveness property, which has been studied for subspace clustering using sparse representation [13, 7] and low-rank representation [14]. The selected exemplars can naturally represent the manifold structure of the dataset, and thus reduce the dimensionality in the object space. Actually, finding exemplars is of particular significance in large-scale dataset summarization and visualization, and improves memory requirement and computational time of classification on such large-scale datasets. Nevertheless, merely selecting exemplars in the original space is insufficient to cover all the data points for classification task, since these data points distribute along complex manifolds and the exemplars may be neighbors to the data points from different classes. For this reason, it is desirable to learn a latent space in which the selecting exemplars can better serve the classification purpose.

By considering the two ways of dimensionality reduction and their limitations in classification presented above, we propose an algorithm to implement dimensionality reduction along the two directions by considering the manifold structure of each class. The proposed algorithm simultaneously does the following:

- find exemplars from each class to represent the class-specific manifold structure, in which way the object-space dimensionality is reduced;
- learn a latent space in the feature space to make the mapped data more discriminative according to the class-specific manifold measurement.
- carry out classification under a simple sparse coding framework.

We call this algorithm Exemplar-represented Manifolds in Latent Space for Classification (EMLSC). In the classification stage, EMLSC adopts a simple sparse coding framework for classification in a way like Sparse Representation-based Classification (SRC) [15]. But different from employing all training data in SRC, EMLSC only uses the selected exemplars as the bases. As the sparse coding is done in the lower-dimensional latent space and the number bases is far smaller than the whole training set, it is anticipated that the classification is performed faster than the original SRC method, in which the whole training set is used for classification. Furthermore, we present a nonlinear extension of EMLSC via kernel tricks (K-EMLSC) to deal with highly nonlinear situations. Through experimental validation, we can see that (K-)EMLSC not only reduces the dimension of the data and the scale of the dataset, but also improves classification performance.

2 Notations and Related Work

Let $\mathbf{X} \in \mathbb{R}^{p\times N}$ denote a training data set which consists of N data points from C classes, and $\mathbf{X}_c \in \mathbb{R}^{p\times N_c}$ denote the subset of data from the c^{th} class, where $N = \sum_{c=1}^{C} N_c$. $\mathbf{I}$ is an identity matrix with appropriate size.

Under SRC framework [15], Ngugen *et al.* propose a unified method called Latent Sparse Embedding Residual Classifier (LASERC) to learn dictionary and a latent space [16]. In detail, for each class, LASERC jointly learn an adaptive dictionary $\mathbf{D}_c$ and a latent space defined by a projection $\mathbf{W}_c$ through:

$$\begin{aligned}\min_{\mathbf{W}_c,\mathbf{D}_c,\mathbf{A}_c} &\|\mathbf{W}_c^T\mathbf{X}_c - \mathbf{D}_c\mathbf{A}_c\|_F^2 + \lambda\|\mathbf{X}_c - \mathbf{W}_c\mathbf{W}_c^T\mathbf{X}_c\|_F^2,\\ &\text{s.t. } \mathbf{W}_c^T\mathbf{W}_c = \mathbf{I}, \|\mathbf{A}_c\|_1 \le T,\end{aligned} \tag{1}$$

where $\mathbf{A}_c$ is the coefficient matrix and $\|\mathbf{A}_c\|_1$ is the sum of ℓ_1 norms of all columns in $\mathbf{A}_c$. LASERC uses the projection to reduce the dimensionality of the data, and adopts a reconstruction error based classifier for the final classification. However, even though the method can be extended to nonlinear version via kernel tricks, it fails to consider the discrimination power among the separately learned class-specific dictionaries $\mathbf{D}_c$'s, such that it is not guaranteed to produce improved classification performance.

Elhamifar *et al.* propose to find exemplars in the dataset to reduce the dimensionality in the object space [17], so that computational cost and memory requirements are significantly reduced. They use nearest neighbor to do classification, and achieve comparable results with exemplars to that with all the training data. In [7], the authors also propose Sparse Modeling Representative Selection (SMRS) to find exemplars for classification with different classifiers. SMRS first selects exemplars by solving the following objective function over row-sparse coefficient matrix $\mathbf{A}$:

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2, \quad \text{s.t. } \|\mathbf{A}\|_{1,q} \le \tau, \mathbf{1}^T\mathbf{A} = \mathbf{1}^T, \tag{2}$$

where $\|\mathbf{A}\|_{1,q} = \sum_{i=1}^{N} \|\mathbf{a}^i\|_q$ denotes the sum of ℓ_q norms[1] of the rows of coefficient matrix $\mathbf{A} = [\mathbf{a}^1; \ldots; \mathbf{a}^N] \in \mathbb{R}^{N\times N}$; $\tau > 1$ is an appropriately chosen parameter to make the optimization program in Eq. 2 convex; and the affine constraint $\mathbf{1}^T\mathbf{A} = \mathbf{1}^T$ means invariance of the selected exemplars w.r.t global translation of the data. As the $\ell_{1,q}$-norm vanishes rows of $\mathbf{A}$, the exemplars can be found according to nonzero rows in $\mathbf{A}$. SMRS learns different classifiers over the selected exemplars, and experimental results demonstrate that well-chosen exemplars can not only reduce the scale of the training set, but also produce very good classification performances with far fewer data points. Despite the effectiveness of SMRS, it separately finds the exemplars in the original space and learns the classifier. Therefore, the learned classifiers are not optimal for classification based on the selected exemplars. Moreover, SMRS simply selects exemplars from all the classes, hence using the exemplars as a subset of the training data in the original space can significantly change the inner and intra class distances of the training data, such that good classification performance is not guaranteed as discussed in [7].

[1] In this paper, we merely set $q = 2$, *i.e.* using an $\ell_{1,2}$-norm regularizer in the objective functions presented latter.

3 Exemplar-Represented Manifold in Latent Space for Classification (EMLSC)

As reviewed previously, finding exemplars in the original space directly from all classes is not optimal for classification, and separately learning class-specific dictionaries also limits the discrimination power of the dictionaries. Since we understand the importance of selecting exemplars opposed to learning adaptive dictionaries in representing the class-specific manifold structure, it is worth simultaneously finding exemplars in each class and learning a latent space with consideration of the discrimination power.

3.1 Derivation of EMLSC Objective Function

In SMRS, solving Eq. 2 means finding exemplars from all classes in the original space, and it cannot serve classification purpose well. Therefore, it is desirable to finding exemplars in a simultaneously learned latent space, in which the exemplars can effectively represent the data points according to their class-specific manifold structure. Suppose a linear projection $\mathbf{W} \in \mathbb{R}^{p\times m}$ defines this m-dimensional latent space, then we have the new data in the latent space as $\mathbf{W}^T\mathbf{X}$. By replacing the original data set $\mathbf{X}$ in Eq. 2 with $\mathbf{W}^T\mathbf{X} \in \mathbb{R}^{m\times N}$, and constraining $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, we have:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{W}} \|\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{A}\|_F^2, \\ &\text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}, \|\mathbf{A}\|_{1,q} \le \tau, \mathbf{1}^T\mathbf{A} = \mathbf{1}^T. \end{aligned} \tag{3}$$

The constraint of $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ not only leads to a computationally efficient scheme for optimization as we see in the next subsection, but also allows the extension of our proposed method to the nonlinear version as demonstrated in Section 4.

Moreover, it is essential to guarantee that the exemplars from a specific class can well represent all the data of this class. Specifically, for the c^{th} class $\mathbf{X}_c$, we should also minimize the following constraint:

$$\|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}_c\mathbf{A}_c^{(c)}\|_F^2, \tag{4}$$

where $\mathbf{A}_c^{(c)}$ is the c^{th} part of coefficient matrix $\mathbf{A}_c = [\mathbf{A}_c^{(1)}; \ldots; \mathbf{A}_c^{(i)}; \ldots; \mathbf{A}_c^{(C)}]$ corresponding to $\mathbf{X}_c$, *i.e.* $\mathbf{X}_c \approx \mathbf{X}\mathbf{A}_c = \sum_{i=1}^{C} \mathbf{X}_i\mathbf{A}_c^{(i)}$. For brevity, we introduce a selection operator $\mathbf{Q}_c = [\mathbf{q}_1^c, \ldots, \mathbf{q}_j^c, \ldots, \mathbf{q}_{K_c}^c] \in \mathbb{R}^{N\times N_c}$, in which the j^{th} column of $\mathbf{Q}_c$ is of the following form:

$$\mathbf{q}_j^c = [\underbrace{0, \ldots, 0}_{\sum_{i=1}^{c-1} N_i}, \overbrace{\underbrace{0, \ldots, 0}_{j-1}, 1, 0, \ldots, 0}^{N_c}, \underbrace{0, \ldots, 0}_{\sum_{i=c+1}^{C} N_i}]^T. \tag{5}$$

Therefore, we have $\mathbf{Q}_c^T\mathbf{Q}_c = \mathbf{I}$, $\mathbf{X}_c = \mathbf{X}\mathbf{Q}_c$, and $\mathbf{A}_j^{(c)} = \mathbf{Q}_c^T\mathbf{A}_j \in \mathbb{R}^{N_j}$ means the c^{th} part of coefficient matrix $\mathbf{A}_j$ corresponding to $\mathbf{X}_c$. Now, we can rewrite Eq. 4 as:

$$\|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}\mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c\|_F^2. \tag{6}$$

Let $\tilde{\mathbf{Q}}_c = [\mathbf{Q}_1, \ldots, \mathbf{Q}_{c-1}, \mathbf{Q}_{c+1}, \ldots, \mathbf{Q}_C]$, then we have $\mathbf{X}\tilde{\mathbf{Q}}_c = [\mathbf{X}_1, \ldots, \mathbf{X}_{c-1}, \mathbf{X}_{c+1}, \ldots, \mathbf{X}_C]$, and $\tilde{\mathbf{Q}}_c^T\mathbf{A}_c = [\mathbf{A}_c^{(1)}; \ldots; \mathbf{A}_c^{(c-1)}; \mathbf{A}_c^{(c+1)}; \ldots; \mathbf{A}_c^{(C)}]$. To guarantee that exemplars from other classes do not contribute to representing the data from class c, we should also minimize the following:

$$\|\mathbf{W}^T\mathbf{X}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2. \tag{7}$$

This term measures how much the unrelated exemplars (from undesirable classes) contribute to the representation of the data points from a specific class. Thus, minimizing this term means drawing apart the exemplars belonging to different classes [11], in which way the data points from different classes are better separated.

Considering the above three terms, *i.e.* Eq. 3, Eq. 6 and Eq. 7, we have our objective function as below:

$$\begin{aligned}\min_{\mathbf{A}_c,\mathbf{W}} \sum_{c=1}^{C} \Big\{ &\|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}\mathbf{A}_c\|_F^2 + \alpha\|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}\mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c\|_F^2 \\ &+ \beta\|\mathbf{W}^T\mathbf{X}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2 \Big\} \\ \text{s.t. } &\mathbf{W}^T\mathbf{W} = \mathbf{I}, \|\mathbf{A}_c\|_{1,q} \le s, \mathbf{1}^T\mathbf{A}_c = \mathbf{1}^T, \text{ for } \forall c.\end{aligned} \tag{8}$$

In the objective function, α and β are two parameters to balance relative importance of the three terms, and s denotes the sparse level of the coefficient $\mathbf{A}_c$. There are other possible ways to add discrimination power to the latent space and exemplars, such as the methods based on Linear Discriminative Analysis [18] and Maximum Margin Criterion [19]. But it is worth noting the way of improving discrimination in Eq. 8 has an intrinsic relation to the classifier adopted in this paper. As described in Section 5, since our classifier is based on sparse coding technique, this discrimination-enhancing method in Eq. 8 can benefit the classifier a lot.

3.2 Numerical Solution

Even though the optimization problem in Eq. 8 is a non-convex problem with two matrix variables $\mathbf{W}$ and $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_C]$, we still can derive effective solutions through iterative minimization, as demonstrated by experiments in Section 6. In this subsection, we present the detailed optimization of each variable matrix.

Update projection $\mathbf{W}$ By omitting the terms which are independent to $\mathbf{W}$, we have the following:

$$\begin{aligned}\mathbf{W}^* = &\operatorname*{argmin}_{\mathbf{W}} \|\mathbf{W}^T(\mathbf{X} - \mathbf{X}\mathbf{A})\|_F^2 + \alpha\|\mathbf{W}^T(\mathbf{X} - \mathbf{X}\hat{\mathbf{A}})\|_F^2 + \beta\|\mathbf{W}^T(\mathbf{X}\tilde{\mathbf{A}})\|_F^2, \\ &\text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I},\end{aligned} \tag{9}$$

where $\hat{\mathbf{A}} = [\mathbf{Q}_1\mathbf{Q}_1^T\mathbf{A}_1, \ldots, \mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c, \ldots, \mathbf{Q}_C\mathbf{Q}_C^T\mathbf{A}_C]$, $\tilde{\mathbf{A}} = [\tilde{\mathbf{Q}}_1\tilde{\mathbf{Q}}_1^T\mathbf{A}_1, \ldots, \tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c, \ldots, \tilde{\mathbf{Q}}_C\tilde{\mathbf{Q}}_C^T\mathbf{A}_C]$, and $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_c, \ldots, \mathbf{A}_C]$. Through simple derivation, we have the following concise function:

$$\mathbf{W}^* = \operatorname*{argmin}_{\mathbf{W}} \mathsf{tr}(\mathbf{W}^T\boldsymbol{\Omega}\mathbf{W}), \quad \text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I}, \tag{10}$$

where $\boldsymbol{\Omega} = (\mathbf{X} - \mathbf{X}\mathbf{A})(\mathbf{X} - \mathbf{X}\mathbf{A})^T + \alpha(\mathbf{X} - \mathbf{X}\hat{\mathbf{A}})(\mathbf{X} - \mathbf{X}\hat{\mathbf{A}})^T + \beta(\mathbf{X}\tilde{\mathbf{A}})(\mathbf{X}\tilde{\mathbf{A}})^T$. Therefore, to derive the optimal $\mathbf{W}^*$ with fixed $\mathbf{A}$, we can simply solve this eigenvalue decomposition problem, and choose the eigenvectors w.r.t the m smallest eigenvalues as the columns of $\mathbf{W}^*$.

Update the coefficient matrix $\mathbf{A}_c$ Specifically, by fixing the projection $\mathbf{W}$, we focus on updating $\mathbf{A}_c$ for demonstration as below:

$$\begin{aligned}\mathbf{A}_c^* &= \underset{\mathbf{A}_c}{\operatorname{argmin}} \|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}\mathbf{A}_c\|_F^2 + \alpha\|\mathbf{W}^T\mathbf{X}_c - \mathbf{W}^T\mathbf{X}\mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c\|_F^2 \\ &\qquad + \beta\|\mathbf{W}^T\mathbf{X}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c\|_F^2 \\ &= \underset{\mathbf{A}_c}{\operatorname{argmin}} \|\bar{\mathbf{X}}_c - \bar{\mathbf{X}}\mathbf{A}_c\|_F^2 \\ &\text{s.t. } \|\mathbf{A}_c\|_{1,q} \le s, \mathbf{1}^T\mathbf{A}_c = \mathbf{1}^T,\end{aligned} \tag{11}$$

where $\bar{\mathbf{X}}_c = \begin{pmatrix} \mathbf{W}^T\mathbf{X}_c \\ \sqrt{\alpha}\mathbf{W}^T\mathbf{X}_c \\ \mathbf{0} \end{pmatrix}$ and $\bar{\mathbf{X}} = \begin{pmatrix} \mathbf{W}^T\mathbf{X} \\ \sqrt{\alpha}\mathbf{W}^T\mathbf{X}\mathbf{Q}_c\mathbf{Q}_c^T \\ \sqrt{\beta}\mathbf{W}^T\mathbf{X}\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T \end{pmatrix}$. In this paper, by using Lagrange multipliers on $\|\mathbf{A}_c\|_{1,q}$, we turn to an Alternating Direction Method of Multipliers optimization framework [20] to solve the above problem.

In sum, the overall optimization procedure iterates the two steps, updating $\mathbf{W}$ in Eq. 10 and updating $\mathbf{A}$ in Eq. 11. It stops when meeting a predefined condition, *i.e.* reaching a maximum number of iterations or the difference between two consecutive the projection $\mathbf{W}$ is small enough. Finally, we choose the data as selected exemplars according to nonzero rows of the coefficient matrix $\mathbf{A}$.

4 Kernel EMLSC

Even if the proposed EMLSC exploit the data manifolds in the learned latent space based on the selected exemplars, it may fail to discover the intrinsic geometry when the data manifold is highly nonlinear. In this section, we discuss how to perform EMLSC in Reproducing Kernel Hilbert Space (RKHS), which gives rise to kernel version of EMLSC, denoted as K-EMLSC.

4.1 An Equivalent Objective Function

Before deriving the K-EMLSC, we provide an equivalent objective function to the original one in Eq. 8. We first present the following proposition.

Proposition 1 *With fixed $\mathbf{A}$, there exists an optional solution $\mathbf{W}^*$ to Eq. 8 that has the following form:*

$$\mathbf{W}^* = \mathbf{X}\mathbf{P} \tag{12}$$

for some $\mathbf{P} \in \mathbb{R}^{N\times m}$.

The proof of this proposition is given in Appendix A. As a corollary of Proposition 1, it is sufficient to seek an optimal solution for the optimization in Eq. 8 through $\mathbf{P}$ and coefficient matrix $\mathbf{A}$. By substituting Eq. 12 into Eq. 8, we have:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{P}} \|\mathbf{P}^T\mathbf{K}(\mathbf{I}-\mathbf{A})\|_F^2 + \alpha\|\mathbf{P}^T\mathbf{K}(\mathbf{I}-\hat{\mathbf{A}})\|_F^2 + \beta\|\mathbf{P}^T\mathbf{K}\tilde{\mathbf{A}}\|_F^2, \\ &\text{s.t. } \mathbf{P}^T\mathbf{K}\mathbf{P} = \mathbf{I}, \|\mathbf{A}_c\|_{1,q} \leq s, \mathbf{1}^T\mathbf{A}_c = \mathbf{1}^T, \text{ for } \forall c, \end{aligned} \tag{13}$$

where $\mathbf{K} = \mathbf{X}^T\mathbf{X}$, $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_c, \ldots, \mathbf{A}_C]$, $\hat{\mathbf{A}} = [\mathbf{Q}_1\mathbf{Q}_1^T\mathbf{A}_1, \ldots, \mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c, \ldots, \mathbf{Q}_C\mathbf{Q}_C^T\mathbf{A}_C]$, and $\tilde{\mathbf{A}} = [\tilde{\mathbf{Q}}_1\tilde{\mathbf{Q}}_1^T\mathbf{A}_1, \ldots, \tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c, \ldots, \tilde{\mathbf{Q}}_C\tilde{\mathbf{Q}}_C^T\mathbf{A}_C]$.

To derive the optimal $\mathbf{P}$, we have the following proposition.

Proposition 2 *The optimal solution of Eq. 13 when* $\mathbf{A}$ *is fixed is:*

$$\mathbf{P}^* = \mathbf{U}\mathbf{S}^{-\frac{1}{2}}\mathbf{G}^*, \tag{14}$$

where $\mathbf{U}$ *and* $\mathbf{S}$ *come from the SVD of* $\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^T$*, and* $\mathbf{G} \in \mathbb{R}^{N\times m}$ *is the optimal solution of the following minimum eigenvalue problem:*

$$\mathbf{G}^* = \underset{\mathbf{G}}{\operatorname{argmin}}\, \mathrm{tr}\mathbf{G}^T\tilde{\mathbf{H}}\mathbf{G}, \quad \textit{s.t. } \mathbf{G}^T\mathbf{G} = \mathbf{I}, \tag{15}$$

where $\tilde{\mathbf{H}} = \mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}$ *in which* $\mathbf{H} = (\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T + \alpha(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T + \beta\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$.

The proof of this proposition is provided in Appendix B. From the equivalence illustrated by Proposition 2, we can derive the optimal $\mathbf{W} = \mathbf{X}\mathbf{P}$ after having the optimal $\mathbf{P}$. It is worth noting the following remark.

Remark 1. With fixed coefficient matrix $\mathbf{A}$, we can derive the optimal projection $\mathbf{W}$ either through solving Eq. 10 or Eq. 12 (Eq. 14 and Eq. 15 are used). The difference between these two ways can be beneficial in different situations. Particularly, when the number of training data $N \gg p$, which stands for the dimensionality of the data, we can choose the first way to derive the optimal $\mathbf{W}$, as the complexity of the eigenvalue problem is $\mathcal{O}(p^3)$. When $p \gg N$, we can use the second strategy to calculate the optimal $\mathbf{W}$ with the $\mathcal{O}(N^3)$ complexity of the eigenvalue problem .

4.2 Derivation of Kernel EMLSC

Since we have $\mathbf{x}_i \in \mathbb{R}^p$ where $\mathbf{x}_i$ is the i^{th} training sample, we consider the problem in a feature space $\mathcal{H}$ induced by some nonlinear mappings:

$$\phi : \mathbb{R}^p \to \mathcal{H}. \tag{16}$$

For a proper chosen ϕ, an inner produce $\langle\cdot,\cdot\rangle$ can be defined on $\mathcal{H}$ which makes for a reproducing kernel Hilbert space (RKHS). More specifically,

$$\langle\phi(\mathbf{x}), \phi(\mathbf{y})\rangle = \mathcal{K}(\mathbf{x}, \mathbf{y}) \tag{17}$$

holds where $\mathcal{K}(\cdot,\cdot)$ is a positive semi-definite kernel function. Several popular kernel functions are Gaussian kernel $\mathcal{K}(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2)$, polynomial kernel $\mathcal{K}(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}^T\mathbf{y})^\alpha$, and Sigmoid kernel $\mathcal{K}(\mathbf{x},\mathbf{y}) = \tanh(\mathbf{x}^T\mathbf{y}+\alpha)$.

Let $\boldsymbol{\Phi}$ denote the data matrix in RKHS:

$$\boldsymbol{\Phi} = [\phi(\mathbf{x}_i),\ldots,\phi(\mathbf{x}_N)]. \tag{18}$$

Now, the problem Eq. 13 in RKHS can be written as below:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{P}} \|\mathbf{P}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\mathbf{I}-\mathbf{A})\|_F^2 + \alpha\|\mathbf{P}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\mathbf{I}-\hat{\mathbf{A}})\|_F^2 + \beta\|\mathbf{P}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\tilde{\mathbf{A}}\|_F^2, \\ &\text{s.t. } \mathbf{P}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{P} = \mathbf{I}, \|\mathbf{A}_c\|_{1,q} \le s, \mathbf{1}^T\mathbf{A}_c = \mathbf{1}^T, \text{ for } \forall c. \end{aligned} \tag{19}$$

Denote the kernel matrix by $\boldsymbol{\mathcal{K}} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$, in which $\boldsymbol{\mathcal{K}}_{ij} = \mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) = \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$. Then, we have:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{P}} \|\mathbf{P}^T\boldsymbol{\mathcal{K}}(\mathbf{I}-\mathbf{A})\|_F^2 + \alpha\|\mathbf{P}^T\boldsymbol{\mathcal{K}}(\mathbf{I}-\hat{\mathbf{A}})\|_F^2 + \beta\|\mathbf{P}^T\boldsymbol{\mathcal{K}}\tilde{\mathbf{A}}\|_F^2, \\ &\text{s.t. } \mathbf{P}^T\boldsymbol{\mathcal{K}}\mathbf{P} = \mathbf{I}, \|\mathbf{A}_c\|_{1,q} \le s, \mathbf{1}^T\mathbf{A}_c = \mathbf{1}^T, \text{ for } \forall c. \end{aligned} \tag{20}$$

The resulting kernelized objective function in Eq. 20 can be solved in the same way as in the linear case. Note that in the nonlinear case, the dimension m of the output space can be higher than the dimension p of the input space, and is only upper bounded by the number of training samples. For a sample datum $\mathbf{x}$ either from the training set or a testing one, we have the corresponding mapped point $\mathbf{z}$ in RKHS as $\mathbf{z} = \mathbf{P}^T\mathcal{K}(\mathbf{X},\mathbf{x})$.

5 Classification Scheme

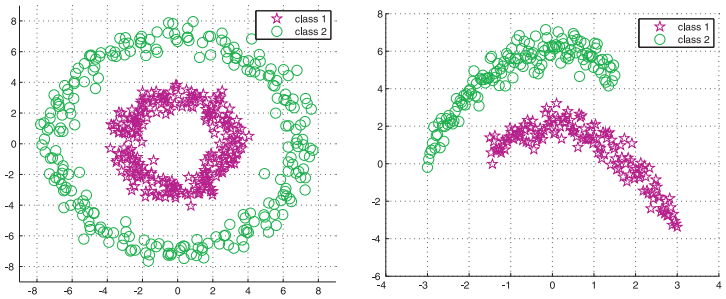
After learning the projection and selecting the exemplars, we have the linear or nonlinear mapped exemplars as $\mathbf{Z} = [\mathbf{Z}_1,\ldots,\mathbf{Z}_C] \in \mathbb{R}^{m\times M}$, in which $\mathbf{Z}_c \in \mathbb{R}^{m\times M_c}$ is the mapped exemplars selected from the c^{th} class ($\sum_{c=1}^{C} M_c = M$). Specifically, in the linear version, we have $\mathbf{Z}_c = \mathbf{W}^T\mathbf{X}_c$, while in nonlinear situation, we have $\mathbf{Z}_c = \mathbf{P}^T\mathcal{K}(\mathbf{X},\mathbf{X}_c)$. When comes a query datum $\mathbf{x}$, we have the mapped point $\mathbf{z} \in \mathbb{R}^m$. To classify the query, we follow the classification framework of SRC [15]. In detail, we first solve the following sparse coding problem:

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{z}-\mathbf{Z}\boldsymbol{\alpha}\|_F^2 + \gamma\|\boldsymbol{\alpha}\|_1. \tag{21}$$

Then, we calculate the reconstruction error of each class by: $e_c = \|\mathbf{z}-\mathbf{Z}_c\boldsymbol{\alpha}_c^*\|_F^2$, where $\boldsymbol{\alpha}_c$ is the part in coefficient $\boldsymbol{\alpha}^*$ corresponding to $\mathbf{Z}_c$. Finally, we classify the query to class c^* such that $c^* = \operatorname{argmin}_c e_c$.

6 Experimental Results

In this section, we evaluate the performance of EMLSC with its kernel version denoted by K-EMLSC on both synthetic and two real-world datasets. We compare our



(a) (b)

Fig. 1. Two synthetic datasets and each one has two classes: (a) circle-like distributed data and (b) parabola-like distributed data.

proposed (K-)EMLSC with several state-of-the-art methods that are based on Sparse Representation-based Classification (SRC) framework for classification on the learned dictionaries or the selected exemplars. The standard SRC method (SRC) acts as a baseline method, which uses all the training data without learning dictionaries or finding exemplars. As our EMLSC can also learn a project along the feature-space dimension, for comparing the effect of the dimensionality on the feature-space dimension, we first apply random projection, PCA and LPP [6] to reduce the feature-space dimensionality, and then use SRC for classification. We call these schemes rSRC, pSRC and lSRC respectively. Moreover, we compare two closely related methods for comparison, they are Sparse Modeling Representative Selection (SMRS) [7] and Latent Sparse Embedding Residual Classifier (LASERC) [16]. SMRS merely selects exemplars in the original space, and LASERC simultaneously learns the projection and adaptive dictionaries for each class. Both SMRS and LASERC use SRC framework for classification. For all experiments, we simply set $\alpha = \beta = 1$ for EMLSC and K-EMLSC. As for K-EMLSC, we choose Gaussian kernel with $\sigma = 1.7$.

6.1 Experiments with Synthetic Data

We first evaluate K-EMLSC for its ability of discriminating class-specific manifolds in the learned latent space. We synthesize two datasets of two-dimensional data points, and each set includes two classes of data. These two datasets consist of circle-like and parabola-like distributed data as illustrated by Fig. 1. We can easily see there are highly nonlinear manifold structures in the data. Our K-EMLSC jointly learns a latent space and selects exemplars, therefore, we can draw the mapped data points of the two classes in the learned latent space. Among the compared methods, only LASERC can jointly learn a latent space, but learns dictionary for each class individually. Therefore, we focus on the comparison of K-EMLSC and LASERC, and choose Gaussian kernel with $\sigma = 1.7$ for both of the methods.

Fig. 2 displays the mapped data of the synthetic datasets in the latent space by LASERC and K-EMLSC. (a) to (d) are the mapped data from the circle-like and parabola-like datasets by LASERC and K-EMLSC, respectively; as well, we also plot the learned

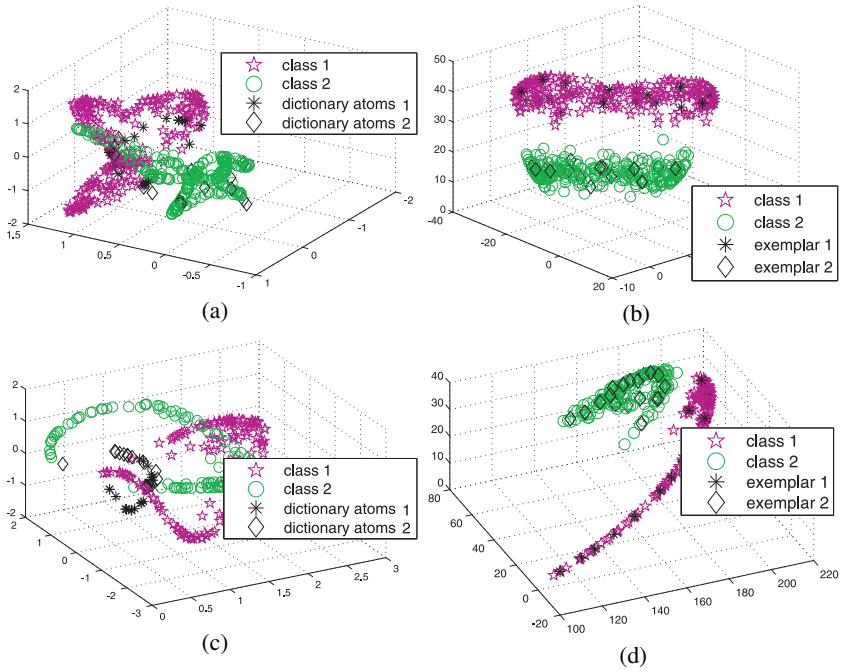


Fig. 2. Synthetic example: circle-like distributed data in the original 2D space and the latent space with the learned dictionary atoms or selected exemplars are presented in (a) and (b), respectively; circle-like distributed data in the original 2D space and the latent space with the learned atoms or selected exemplars are illustrated by (c) and (d), respectively.

dictionary atoms of LASERC and the selected exemplars of our K-EMLSC. Since LASERC learns the class-specific dictionaries and projection matrices of each class individually, we plot the data points of the two classes in one figure, as illustrated in (a) and (c). Then, we can see the data are mixed up, as well as the dictionary atoms. This means the dictionary atoms cannot represent the data points well. This observation coincides with what we analyze previously — dictionary atoms act bases to reconstruct data points, and thus the atoms cannot be directly used to represent the data. On the contrary, our K-EMLSC separates the two classes and preserves the class-specific manifolds clearly as expected as demonstrated in (b) and (d), because K-EMLSC considers the discrimination power and representation power of the selected exemplars in the mapped space. Moreover, the selected exemplars can fully represent the data of each class, specifically, the exemplars in the learned latent space can reflect the class-specific data manifolds clearly.

6.2 Experiments with Real Data

Now, we examine the classification performances of the proposed method on two real-world datasets, USPS digit database [21] and Extended-YaleB face database [22]. We

Table 1. USPS digit recognition accuracy (%) with different reduced feature-space dimension.

	10	20	40	60	80
rSRC	89.3 ± 0.6	90.1 ± 0.7	92.5 ± 0.4	93.8 ± 0.4	95.6 ± 0.5
pSRC	93.2 ± 0.4	95.9 ± 0.5	97.3 ± 0.6	98.1 ± 0.3	98.6 ± 0.4
LASERC	85.6 ± 1.6	86.3 ± 1.3	86.9 ± 1.2	87.2 ± 1.1	87.9 ± 0.9
lSRC	93.6 ± 0.6				
SRC	98.9 ± 0.7				
SMRS	91.7 ± 0.6				
EMLSC	95.8 ± 0.7	96.2 ± 0.6	97.1 ± 0.7	97.8 ± 0.5	98.2 ± 0.4
K-EMLSC	96.1 ± 0.3	96.5 ± 0.4	97.3 ± 0.8	97.9 ± 0.4	98.4 ± 0.5

Table 2. Extended-YaleB face recognition accuracy (%) with different reduced feature-space dimension.

	30	60	100	150	200
rSRC	82.7 ± 1.3	91.6 ± 1.4	94.6 ± 1.1	95.8 ± 1.2	96.4 ± 0.9
pSRC	86.4 ± 0.7	91.8 ± 0.7	93.4 ± 0.9	93.8 ± 0.8	94.6 ± 0.6
LASERC	83.3 ± 1.8	87.5 ± 1.5	89.8 ± 1.4	91.4 ± 1.6	92.1 ± 1.3
lSRC	87.4 ± 0.6				
SRC	98.2 ± 0.8				
SMRS	93.1 ± 0.7				
EMLSC	93.6 ± 0.8	95.5 ± 0.6	96.3 ± 0.5	97.9 ± 0.5	98.5 ± 0.3
K-EMLSC	94.2 ± 0.3	95.9 ± 0.4	96.7 ± 0.4	98.2 ± 0.6	98.7 ± 0.5

show that class-specific manifolds commonly exist in real-world data sets, and our (K-)EMLSC can achieve very promising classification results by simultaneously learning the latent space and selecting the exemplars with consideration of the class-specific manifolds.

In USPS/Extended-YaleB dataset, we randomly select 1000 (USPS) / 51 (YaleB) samples of each class for training, and restrict our (K-)EMLSC to select 20 (USPS) / 7 (YaleB) exemplars in each class. As well, SMRS also selects the same number of exemplars as (K-)EMLSC, and LASERC learns the same number of dictionary atoms for each class. The recognition accuracy of each method is averaged over 10 runs. As for dimensionality reduction along the feature-space dimension, rSRC, pSRC, LASERC and our (K-)EMLSC reduce the data to the same dimension; while lSRC reduces the data to $C - 1$ dimension, where C is the number of classes; SRC and SMRS directly perform on the original data without dimensionality reduction process. In this evaluation, only K-EMLSC uses a Gaussian kernel.

Table 1 and Table 2 report the averaged classification accuracies with standard deviations for USPS and Extended-YaleB respectively, and the results are obtained from each method after running 10 times. Fig.3 and Fig. 4 illustrate the two-dimensional mapped images and exemplars of the two databases, respectively. From these results, we make the following remarks:

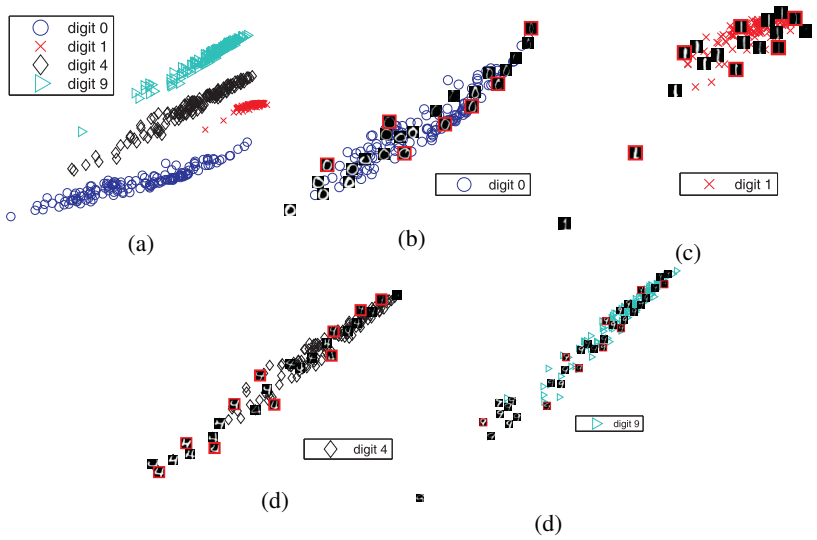


Fig. 3. Four digits from USPS are projected into the 2D latent space shown in (a). (b), (c) and (d) are three digits zoomed in with the selected exemplars highlighted in red box (best seen in color).

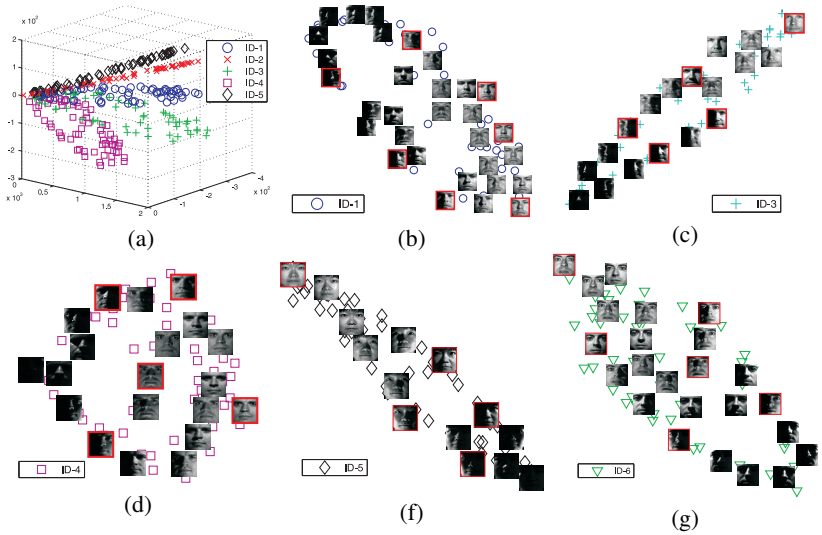


Fig. 4. (a) Six individuals from Extended YaleB database are projected in the latent space. (b), (c) and (d) present facial images of three persons, and the images in red box are the selected exemplars (best seen in color).

1. rSRC, pSRC and lSRC reduce the dimension in the feature space and then apply SRC for classification in separate stages. As seen in the tables, these methods do not generate favorable results compared to (K-)EMLSC. In contrast, (K-)EMLSC jointly learns the latent space for reducing the feature-space dimension and selects the exemplars to represent class-specific manifolds. Therefore, with lower feature-space dimension and less training samples (exemplars) for classification, it can still produce very promising results. When kernel tricks are adopted, K-EMLSC achieves better results than EMLSC.
2. Even if LASERC jointly learns projection and dictionaries, the learned dictionary atoms cannot represent the data points but act as bases for reconstructing the data as argued previously. Moreover, LASERC learns the dictionary and projection for each class individually, therefore, classification performance cannot be guaranteed. On the contrary, (K-)EMLSC jointly considers the discrimination power and the representation power of the exemplars in the class-specific manifold viewpoint. Fig.3 and Fig. 4 clearly illustrate the merit of EMLSC on this respect.
3. Both SMRS and (K-)EMLSC select exemplars, but (K-)EMLSC obtains higher accuracy, even with much lower feature-space dimension. This is because (K-)EMLSC considers the discrimination power of the mapped exemplars in the latent space, and these exemplars can better represent the class-specific manifolds in a discriminative way.
4. SRC, which performs in the original space over all training data, produces decent results in the two databases. But (K-)EMLSC, with much fewer and much lower dimensional exemplars as bases, achieves comparable performances to SRC on USPS database, and even outperforms it in Extended-YaleB database. This observation further demonstrates the merit of (K-)EMLSC.
5. From Fig.3 and Fig. 4, we can see the selected exemplars by EMLSC mainly reside on the contour of the manifolds. This is a good result, because it is reasonable to anticipate data which locate within the manifold can be well approximated by linearly combining the exemplars. This phenomenon validates the effectiveness of using exemplars to represent the data manifold structure.

7 Conclusion

In this paper, we propose an algorithm to simultaneously learn a latent space and find exemplars from the mapped dataset for classification. The selected exemplars can naturally represent the data manifolds, and our method analyze the manifold structure in a discriminative way. Therefore, the exemplar-based class-specific manifolds of the classes are driven to be discriminative in the mapped latent space. We further extend this method to nonlinear version with kernel tricks, therefore, the kernelized method can deal with highly nonlinear cases. Through experiments, we demonstrate the merit of our proposed method. In face of big-data era, as a future work, it is worth extending our method to online learning framework to deal with large-scale situations.

References

1. Jolliffe, I.: Principal Component Analysis. Springer (2002)

2. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (2000) 2323–2326
3. Tenenbaum, J.B.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500) (2000) 2319–2323
4. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: NIPS. (2001)
5. Cai, D., He, X., Zhou, K., Han, J., Bao, H.: Locality sensitive discriminant analysis. In: IJCAI. (2007)
6. He, X., Niyogi, P.: Locality preserving projections. In: NIPS. (2003)
7. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In: CVPR. (2012)
8. Aharon, M., Elad, M., Bruckstein, A.M.: The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11) (2006) 4311–4322
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised dictionary learning. In: NIPS. (2008)
10. Elad, M.: Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing. Springer (2010)
11. Kong, S., Wang, D.: A dictionary learning approach for classification: Separating the particularity and the commonality. In: ECCV. (2012)
12. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: CVPR. (2008)
13. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR. (2009)
14. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: ICML. (2010)
15. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2) (2009) 210–227
16. Nguyen, H.V., Patel, V.M., Nasrabadi, N.M., Chellapa, R.: Sparse embedding: A framework for sparsity promoting dimensionality reduction. In: ECCV. (2012)
17. Elhamifar, E., Sapiro, G., Vidal, R.: Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In: NIPS. (2012)
18. Kong, S., Wang, X., Wang, D., Wu, F.: Multiple feature fusion for face recognition. In: FG. (2013)
19. Kong, S., Wang, D.: Learning individual-specific dictionaries with fused multiple features for face recognition. In: FG. (2013)
20. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comp. Math. Appl. **2**(1) (1976) 17–40
21. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. **16**(5) (1994) 550–554
22. Lee, K.C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5) (2005) 684–698

Appendix A: Proof of Proposition 1

Denote the optimal solution of $\mathbf{W}$ by $\mathbf{W}^*$. we show that $\mathbf{W}^*$ must have the form $\mathbf{W}^* = \mathbf{XP}$, for some $\mathbf{P} \in \mathbb{R}^{N \times m}$.

In advance, we rewrite the objective function into a compact form:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{W}} &\|\mathbf{W}^T\mathbf{X}-\mathbf{W}^T\mathbf{X}\mathbf{A}\|_F^2+\alpha\|\mathbf{W}^T\mathbf{X}-\mathbf{W}^T\mathbf{X}\hat{\mathbf{A}}\|_F^2\\ &+\beta\|\mathbf{W}^T\mathbf{X}\tilde{\mathbf{A}}\|_F^2+\lambda\|\mathbf{A}\|_{1,2}\\ \text{s.t. }&\mathbf{W}^T\mathbf{W}=\mathbf{I}.\end{aligned} \tag{22}$$

where $\mathbf{A}=[\mathbf{A}_1,\ldots,\mathbf{A}_c,\ldots,\mathbf{A}_C]$, $\hat{\mathbf{A}}=[\mathbf{Q}_1\mathbf{Q}_1^T\mathbf{A}_1,\ldots,\mathbf{Q}_c\mathbf{Q}_c^T\mathbf{A}_c,\ldots,\mathbf{Q}_C\mathbf{Q}_C^T\mathbf{A}_C]$, and $\tilde{\mathbf{A}}=[\tilde{\mathbf{Q}}_1\tilde{\mathbf{Q}}_1^T\mathbf{A}_1,\ldots,\tilde{\mathbf{Q}}_c\tilde{\mathbf{Q}}_c^T\mathbf{A}_c,\ldots,\tilde{\mathbf{Q}}_C\tilde{\mathbf{Q}}_C^T\mathbf{A}_C]$.

Using the orthogonal decomposition of $\mathbf{W}^*$, we have:

$$\mathbf{W}^*=\mathbf{W}_\parallel+\mathbf{W}_\perp,\quad\text{where }\mathbf{W}_\parallel=\mathbf{X}\mathbf{P},\text{ and }\quad\mathbf{W}_\perp\mathbf{X}=\mathbf{0} \tag{23}$$

for some appropriate $\mathbf{P}\in\mathbb{R}^{N\times m}$. Columns of $\mathbf{W}_\parallel$ and $\mathbf{W}_\perp$ are in and orthogonal to the column subspace of $\mathbf{X}$, respectively. Substituting Eq. 23 back into objective function Eq. 8, we can rewrite the first term in the following:

$$\begin{aligned}\|\mathbf{W}^T\mathbf{X}-\mathbf{W}^T\mathbf{X}\mathbf{A}\|_F^2=&\|(\mathbf{W}_\parallel+\mathbf{W}_\perp)^T(\mathbf{X}-\mathbf{X}\mathbf{A})\|_F^2=\|\mathbf{W}_\parallel^T(\mathbf{X}-\mathbf{X}\mathbf{A})\|_F^2\\ =&\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}(\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T\mathbf{X}^T\mathbf{W}_\parallel\big\},\end{aligned} \tag{24}$$

the second term as:

$$\begin{aligned}\|\mathbf{W}^T\mathbf{X}-\mathbf{W}^T\mathbf{X}\hat{\mathbf{A}}\|_F^2=&\|(\mathbf{W}_\parallel+\mathbf{W}_\perp)^T(\mathbf{X}-\mathbf{X}\hat{\mathbf{A}})\|_F^2=\|\mathbf{W}_\parallel^T(\mathbf{X}-\mathbf{X}\hat{\mathbf{A}})\|_F^2\\ =&\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T\mathbf{X}^T\mathbf{W}_\parallel\big\},\end{aligned} \tag{25}$$

and the third term as below:

$$\begin{aligned}\|\mathbf{W}^T\mathbf{X}\tilde{\mathbf{A}}\|_F^2=&\|(\mathbf{W}_\parallel+\mathbf{W}_\perp)^T\mathbf{X}\tilde{\mathbf{A}}\|_F^2=\|\mathbf{W}_\parallel^T\mathbf{X}\tilde{\mathbf{A}}\|_F^2\\ =&\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\mathbf{X}^T\mathbf{W}_\parallel\big\}.\end{aligned} \tag{26}$$

Let $\mathbf{X}^T\mathbf{X}=\mathbf{K}$, by putting Eq. 23, Eq. 24 , Eq. 25 and Eq. 26 together, and omitting the unrelated term to $\mathbf{W}$, we have:

$$\begin{aligned}\min_{\mathbf{W}}&\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}(\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T\mathbf{X}^T\mathbf{W}_\parallel\big\}\\ &+\alpha\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T\mathbf{X}^T\mathbf{W}_\parallel\big\}+\beta\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\mathbf{X}^T\mathbf{W}_\parallel\big\},\\ \text{s.t. }&\mathbf{W}^T\mathbf{W}=\mathbf{I}.\end{aligned} \tag{27}$$

Let the singular value decomposition (SVD) of $\mathbf{K}=\mathbf{U}\mathbf{S}\mathbf{U}^T=\mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T$, and $\mathbf{H}=(\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T+\alpha(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T+\beta\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$, we have:

$$\begin{aligned}\mathsf{tr}\big\{\mathbf{W}_\parallel^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W}_\parallel\big\}=&\mathsf{tr}\big\{\mathbf{P}^T\mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P}\big\},\\ =&\mathsf{tr}\big\{\mathbf{G}^T\tilde{\mathbf{H}}\mathbf{G}\big\},\end{aligned} \tag{28}$$

where $\tilde{\mathbf{H}}=\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}$ and $\mathbf{G}=\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P}$. Therefore, we have:

$$\mathsf{tr}\big\{\mathbf{G}^T\tilde{\mathbf{H}}\mathbf{G}\big\}\geq\sum_{i=1}^{m}\sigma_i, \tag{29}$$

where σ_i is the i^{th} smallest eigenvalue of $\tilde{\mathbf{H}}$. In order for the objective function to achieve its minimum, columns of $\mathbf{G}$ have to be the same with eigenvectors corresponding to the smallest eigenvalues of $\tilde{\mathbf{H}}$. Therefore we have $\mathbf{G}^T\mathbf{G} = \mathbf{I}$. Equivalently, we have the constraint:

$$\begin{aligned}\mathbf{W}^T\mathbf{W} = \mathbf{I} =&(\mathbf{W}_{\parallel} + \mathbf{W}_{\perp})^T(\mathbf{W}_{\parallel} + \mathbf{W}_{\perp})\\ =&\mathbf{W}_{\parallel}^T\mathbf{W}_{\parallel} + \mathbf{W}_{\perp}^T\mathbf{W}_{\perp}\\ =&\mathbf{P}^T\mathbf{K}\mathbf{P} + \mathbf{W}_{\perp}^T\mathbf{W}_{\perp}\\ =&\mathbf{G}^T\mathbf{G} + \mathbf{W}_{\perp}^T\mathbf{W}_{\perp},\end{aligned} \tag{30}$$

which means $\mathbf{W}_{\perp} = \mathbf{0}$. In short, the optimal solution of $\mathbf{W}$ has the form:

$$\mathbf{W}^* = \mathbf{W}_{\parallel} = \mathbf{X}\mathbf{P}. \tag{31}$$

This completes the proof.

Appendix B: Proof of Proposition 2

When fixing $\mathbf{A}$, by omitting unrelated terms, we derive from objective function Eq. 13 as below:

$$\begin{aligned}&\|\mathbf{P}^T\mathbf{K}(\mathbf{I}-\mathbf{A})\|_F^2 + \alpha\|\mathbf{P}^T\mathbf{K}(\mathbf{I}-\hat{\mathbf{A}})\|_F^2 + \beta\|\mathbf{P}^T\mathbf{K}\tilde{\mathbf{A}}\|_F^2\\ =&\mathrm{tr}\Big\{\mathbf{P}^T\mathbf{K}\big((\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T + \alpha(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T + \beta\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T\big)\mathbf{K}\mathbf{P}\Big\}\\ =&\mathrm{tr}\Big\{\mathbf{P}^T\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{P}\Big\},\\ &\text{s.t. } \mathbf{P}^T\mathbf{K}\mathbf{P} = \mathbf{I},\end{aligned} \tag{32}$$

where $\mathbf{H} = (\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^T + \alpha(\mathbf{I}-\hat{\mathbf{A}})(\mathbf{I}-\hat{\mathbf{A}})^T + \beta\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$. Let SVD of $\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{U}^T = \mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T$, then we have:

$$\begin{aligned}\mathrm{tr}\Big\{\mathbf{P}^T\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{P}\Big\} =&\mathrm{tr}\Big\{\mathbf{P}^T\mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P}\Big\}\\ =&\mathrm{tr}\Big\{(\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P})^T\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}(\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P})\Big\}\\ =&\mathrm{tr}\Big\{\mathbf{G}^T\tilde{\mathbf{H}}\mathbf{G}\Big\}\\ &\text{s.t. } (\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P})^T(\mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P}) = \mathbf{I},\end{aligned} \tag{33}$$

where $\mathbf{G} = \mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{P}$ and $\tilde{\mathbf{H}} = \mathbf{S}^{\frac{1}{2}}\mathbf{U}^T\mathbf{H}\mathbf{U}\mathbf{S}^{\frac{1}{2}}$. And the constraint can also be simplified as $\mathbf{G}^T\mathbf{G} = \mathbf{I}$.

Now, we can see the equivalence of optimization in Eq. 13 and Eq. 33. And the optimal solution $\mathbf{P}^*$ can be recovered as in Eq. 14, *i.e.* $\mathbf{P}^* = \mathbf{U}\mathbf{S}^{-\frac{1}{2}}\mathbf{G}^*$.

Note that since $\mathbf{K}$ is a positive semidefinite matrix, the diagonal matrix $\mathbf{S}$ has non-negative entries. $\mathbf{S}^{-\frac{1}{2}}$ is obtained by setting non-zero entries along the diagonal of $\mathbf{S}$ to the inverse of their square root and keeping zero elements the same.

This completes the proof.