Trace Quotient Meets Sparsity: A Method for Learning Low Dimensional Image Representations

Xian Wei, Hao Shen, Martin Kleinsteuber
Department of Electrical Engineering and Information Technology
Technische Universität München, Arcisstr. 21, 80333 Munich, Germany
{xian.wei, hao.shen, kleinsteuber}@tum.de

Abstract

This paper presents an algorithm that allows to learn low dimensional representations of images in an unsupervised manner. The core idea is to combine two criteria that play important roles in unsupervised representation learning, namely sparsity and trace quotient. The former is known to be a convenient tool to identify underlying factors, and the latter is known as a disentanglement of underlying discriminative factors. In this work, we develop a generic cost function for learning jointly a sparsifying dictionary and a dimensionality reduction transformation. It leads to several counterparts of classic low dimensional representation methods, such as Principal Component Analysis, Local Linear Embedding, and Laplacian Eigenmap. Our proposed optimisation algorithm leverages the efficiency of geometric optimisation on Riemannian manifolds and a closed form solution to the elastic net problem.

1. Introduction

Finding appropriate low dimensional representations of data is a long-standing challenging problem in data processing and machine learning. Recent development in representation learning shows that appropriate data representations are the key to the success of machine learning algorithms, as different representations can entangle different explanatory information of the data, cf. [4]. The aim of this paper is to construct effective low dimensional representation learning approaches to disentangle various explanatory or discriminative information in the data for solving unsupervised learning problems.

Sparse representation (SR) was developed as an instrument to leverage the underlying sparse structure of data. It has led to a great success in signal reconstruction, denoising and image super-resolution, cf. [2, 9, 16, 31]. These methods can be considered as *data-driven sparse representation* approaches. On the other hand, sparse coefficients can also be interpreted as features that are suitable for the tasks of learning, such as face recognition [29], subspace clustering [11], and image classification [26].

Furthermore, it is also evidential that sparse representation of the data can be further processed by applying other disentangling instruments, in order to reveal task-related information. For example, the work in [32, 33] incorporates linear classifiers with sparse representation to jointly learn a sparsifying dictionary and a classifier. Similarly, adopting sparse representations in a classical expected risk minimisation formulation leads to the so-called *task-driven dictionary learning* approaches, specifically for supervised learning tasks, cf. [24]. In the unsupervised learning setting, directly applying a Principal Component Analysis (PCA) on sparse representations also results in promising performance in 3D visualisation and clustering, cf. [12]. From a perspective of representation learning, it is a logical conclusion that sparse representations contain rich distributed information of the data with respect to certain learning tasks, and it also necessitates application of further learning mechanisms to disentangle the underlying explanatory information.

In this work, we are interested in the problem of unsupervised learning. Among various unsupervised learning techniques, the trace quotient criterion is a simple but powerful instrument for data discrimination, in particular for Dimensionality Reduction (DR). This generic criterion is shared by various classic DR methods, such as PCA, Linear Local Embedding (LLE) [25], Orthogonal Neighbourhood Preserving Projection (ONPP) [22], Locality Preserving Projections (LPP) [17], and Orthogonal LPP (OLPP) [7]. Our main construction in this work is to apply the trace quotient criterion to further disentangle sparse representations for unsupervised learning tasks.

The paper is organised as follows. Section 2 provides a brief review on both sparse representations and the trace quotient criterion. In Section 3, we construct a generic cost function for learning both the sparsifying dictionary and the orthogonal DR transformation, and develop a geomet-

ric conjugate gradient algorithm on the underlying smooth manifold. Three applications of the proposed generic model are discussed in Section 4, together with their experimental evaluations presented in Section 5. Finally, conclusions and outlooks are given in Section 6.

2. The Elastic Net Solution and the Trace Quotient Maximisation

In this section, we recall some facts about both sparse representations and trace quotient optimisation based dimensionality reduction.

2.1. Sparse Coding

Given a set of data samples $x_i \in \mathbb{R}^m$, the aim of sparse coding is to find a collection of atoms $d_j \in \mathbb{R}^m$ such that each data sample can be approximated by a linear combination of only a few of the atoms $\{d_j\}$. According to [4], the atoms can be interpreted as underlying factors that are responsible for explaining the discrepancy in the data set. In other words, sparse coding generates sparsely distributed representations of data with respect to the specific atoms.

The collection of atoms (often as columns in a matrix) is called a dictionary $D \in \mathbb{R}^{m\times r}$, leading to the model

$$x_i = D\phi_i + \epsilon_i, \tag{1}$$

where $\phi_i \in \mathbb{R}^r$ is the corresponding sparse representation, and $\epsilon_i \in \mathbb{R}^m$ is additive noise. In this work, we restrict each column $d_i \in \mathbb{R}^m$ of D to have unit norm, i.e.

$$\mathcal{S}(m,r) := \left\{ D \in \mathbb{R}^{m\times r} \middle| \operatorname{rk}(D) = m, \|d_i\|_2 = 1 \right\}, \tag{2}$$

which is a product manifold of $(m-1)$-dimensional unit spheres. In the literature, there are several well-established methods for sparse coding, which all depend on a certain sparsity measure, cf. [2, 10].

Once a dictionary is given, there are several ways of finding the sparse representation. If sparsity is measured by employing a combination of the ℓ_1- and the ℓ_2-norm, a solution to the elastic net problem [34], i.e.

$$\phi^* := \operatorname*{argmin}_{\phi\in\mathbb{R}^r} \tfrac{1}{2}\|x - D\phi\|_2^2 + \lambda_1\|\phi\|_1 + \tfrac{\lambda_2}{2}\|\phi\|_2^2 \tag{3}$$

yields a convenient way to obtain the sparse representation. The two regularisation parameters $\lambda_1 \in \mathbb{R}^+$ and $\lambda_2 \in \mathbb{R}^+$ are chosen to ensure stability and uniqueness of the solution. Solutions to the elastic net problem (3) share a convenient fact that, under certain assumptions, there exists a closed from expression.

Let us define the set of indices of the non-zero entries of the solution $\phi^* = [\varphi_1^*, \ldots, \varphi_r^*]^\top \in \mathbb{R}^r$ by $\Lambda := \{i \in \{1,\ldots,r\} | \varphi_i^* \neq 0\}$ and $k := |\Lambda|$. Then the solution of the elastic net (3) has a closed-form expression as

$$\phi_D^*(x) := \left(D_\Lambda^\top D_\Lambda + \lambda_2 I_k\right)^{-1} \left(D_\Lambda^\top x - \lambda_1 s_\Lambda\right), \tag{4}$$

where I_k is the $k \times k$ identity matrix, $s_\Lambda \in \{\pm 1\}^k$ carries the signs of ϕ_Λ^*, and $D_\Lambda \in \mathbb{R}^{m\times k}$ is the subset of D in which the index of atoms (columns) fall into the support Λ. With a reasonable assumption that the dictionary D is suitably incoherent, the solution $\phi_D^*(x)$ shares an algorithmically convenient property of being locally twice differentiable with respect to both D and x, cf. [24]. Such a prominent property leads to the framework of task-driven dictionary learning, which is specifically dedicated to supervised learning problems.

2.2. Low dimensional representations based on the trace quotient criterion

The goal of low dimensional representation learning is to find representations $y_i \in \mathbb{R}^l$ of given data samples $x_i \in \mathbb{R}^m$ with $l < m$, via a mapping

$$\mu\colon \mathbb{R}^m \to \mathbb{R}^l, \qquad x \mapsto \mu(x), \tag{5}$$

which captures certain desired properties of the data to facilitate the specific applications. Many classic DR methods restrict the mapping μ to be an orthogonal transformation, i.e. $\mu(x) := U^\top x$ with $U \in St(l,m)$. Here $St(l,m)$ denotes the Stiefel manifold

$$St(l,m) := \left\{ U \in \mathbb{R}^{m\times l} | U^\top U = I_l \right\}. \tag{6}$$

In the category of unsupervised learning, this model includes various classic DR methods, such as PCA, Orthogonal Locality Preserving Projection (OLPP) [7], Orthogonal Linear Graph Embedding (OLGE) [30], and Orthogonal Neighbourhood Preserving Projections (ONPP) [22].

Often, the aforementioned DR methods find the orthogonal transformation $U \in St(l,m)$ via maximising the so-called trace quotient or trace ratio, i.e.

$$g\colon St(l,m) \to \mathbb{R}, \qquad g(U) := \frac{\operatorname{tr}(U^\top A U)}{\operatorname{tr}(U^\top B U)}, \tag{7}$$

where $A \in \mathbb{R}^{m\times m}$ is assumed to be symmetric positive semi-definite and $B \in \mathbb{R}^{m\times m}$ is often assumed to be symmetric positive definite. Both matrices are constructed according to the specific problems, cf. [8, 21], and examples will be given and discussed in Section 4.

Due to the rotation invariance of the function g, i.e. $g(U\Theta) = g(U)$ for $\Theta \in \mathbb{R}^{l\times l}$ being orthogonal, we can redefine the trace quotient function as

$$f\colon Gr(l,m) \to \mathbb{R}, \qquad f(P) := \frac{\operatorname{tr}(AP)}{\operatorname{tr}(BP)}, \tag{8}$$

where $Gr(l,m)$ denotes the Graßmann manifold as the set of all m-dimensional rank-l orthogonal projectors, i.e.

$$Gr(l,m) := \left\{ UU^\top | U \in St(l,m) \right\}. \tag{9}$$

Various efficient optimisation algorithms over Riemannian manifolds have been developed to maximise the function f, cf. [8, 14, 18, 19].

3. The Proposed Joint Learning Framework

In this section, we firstly present a generic cost function, which adopts the sparsifying dictionary learning in the framework of trace quotient maximisation in Section 3.1. Then a geometric conjugate gradient algorithm is presented in Section 3.2.

3.1. A generic cost function

As suggested by the work of [12, 24], further processing on the sparse representation is capable of unveiling task-related underlying factors, potentially for both supervised and unsupervised learning tasks. In what follows, we construct a cost function, which allows to jointly learn both the sparsifying dictionary and the orthogonal transformation in the framework of trace quotient maximisation.

Let us denote by $\Phi(D, X) := [\phi_{x_1}(D), \ldots, \phi_{x_n}(D)] \in \mathbb{R}^{r\times n}$ the sparse representation of the data $X = [x_1, \ldots, x_n]$ for a given dictionary D. We confine ourselves to the solutions of the elastic net minimisation. Let $\mathcal{A}\colon \mathbb{R}^{r\times n} \to \mathbb{R}^{r\times r}$ and $\mathcal{B}\colon \mathbb{R}^{r\times n} \to \mathbb{R}^{r\times r}$ be two smooth functions that serve as generating functions for the matrices A and B in the trace quotient (7). The specific constructions of the mappings $\mathcal{A}$ and $\mathcal{B}$ are given in Section 4. Then we define a generic trace quotient function in sparse representations as

$$\begin{aligned} &\widetilde{f}\colon \mathcal{S}(m,r) \times Gr(l,r) \to \mathbb{R}, \\ &\qquad \widetilde{f}(D,P) := \frac{\mathrm{tr}(\mathcal{A}(\Phi(D,X))P)}{\mathrm{tr}(\mathcal{B}(\Phi(D,X))P) + \sigma}, \end{aligned} \tag{10}$$

with $\sigma \in \mathbb{R}^+$ being chosen to guarantee the denominator of $\widetilde{f}$ to be positive. manifold $\mathcal{S}(m,r) \times Gr(l,r)$. In the rest of the paper, we refer to our proposed model as SPARse LOW dimensional representation learning (*SparLow*).

In order to prevent solution dictionaries from being highly coherent, which is critical for guaranteeing the local smoothness of the elastic net solutions, we employ a log-barrier function on the scalar product of all dictionary columns to control the mutual coherence of the learned dictionary D, cf. [16], i.e.

$$r(D) := -\sum_{1\le i<j\le k} \log(1 - (d_i^\top d_j)^2). \tag{11}$$

Furthermore, the authors in [28] argue that an appropriate dictionary of choice in sparse representation can reveal the semantics of the data. We propose the following regulariser on the dictionary to be learned

$$h(D) = \|D - D^*\|_F^2, \tag{12}$$

where D^* is the optimal data-driven dictionary learned from sparse coding of the data X, and $\|\cdot\|_F$ denotes the Frobenius norm. Practically, we use a dictionary $\widehat{D}$ produced by state of the art methods, such as K-SVD. Our experiments have verified that the heuristic regulariser h ensures solutions of the generic cost function J defined in Eq. (13) to be self explanatory to the data, and guarantees stable performance towards the task of learning.

Algorithm 1: A CG-SparLow Algorithm.

Input : $X \in \mathbb{R}^{m\times n}$ and functions $\mathcal{A}\colon \mathbb{R}^{r\times n} \to \mathbb{R}^{r\times r}$ and $\mathcal{B}\colon \mathbb{R}^{r\times n} \to \mathbb{R}^{r\times r}$ as specified in Section 4 ;

Output: Accumulation point $D^* \in \mathcal{S}(m,r)$ and $P^* \in Gr(l,r)$;

Step 1: Generate initialization $D^{(0)} \in \mathcal{S}(m,r)$ and $P^{(0)} \in Gr(l,r)$, and set $j = 1$;

Step 2: Compute $G^{(1)} = H^{(1)} \leftarrow (\nabla_J(D^{(0)}), \nabla_J(P^{(0)}))$;

Step 3: Set $j = j + 1$;

Step 4: Update $(D^{(j+1)}, P^{(j+1)}) \leftarrow (D^{(j+1)}, P^{(j+1)}) + \lambda H^{(j)}$, where λ is computed by employing a backtracking line search along geodesics;

Step 5: Update $H^{(j+1)} \leftarrow G^{(j+1)} + \gamma H^{(j)}$, where $G^{(j+1)} = (\nabla_J(D^{(j+1)}), \nabla_J(P^{(j+1)}))$, and γ is chosen according to Eq. (14) ;

Step 6: If $j \bmod (r(m-1) + l(r-l) - 1) = 0$, set $H^{(j+1)} \leftarrow G^{(j+1)}$;

Step 7: If $\left\|G^{(j+1)}\right\|$ is small enough, stop. Otherwise, go to Step 3;

By combining the two regularisers discussed above, we construct the following cost function to jointly learn both the sparsifying dictionary and the orthogonal transformation, i.e.

$$\begin{aligned} &J\colon \mathcal{S}(m,r) \times Gr(l,r) \to \mathbb{R}, \\ &\qquad J(D,P) := \widetilde{f}(D,P) + \mu_1 r(D) + \mu_2 h(D), \end{aligned} \tag{13}$$

where the two weighting factors $\mu_1, \mu_2 \in \mathbb{R}^+$ control the influence of the two constraints on the final solution.

3.2. A geometric conjugate gradient algorithm

In this subsection, we present a geometric CG algorithm on the product manifold $\mathcal{S}(m,r) \times Gr(l,r)$ to maximise the generic cost function J, defined in (13). It is well known that CG algorithms offer prominent properties, such as a superlinear rate of convergence and the applicability to large scale optimisation problems with low computational complexity, e.g. in sparse recovery [15]. We refer to [1] for further technical details for these computations.

We denote by $T_{(D,P)}\mathcal{S}(m,r) \times Gr(l,r)$ the tangent space of $\mathcal{S}(m,r) \times Gr(l,r)$ at (D,P), the Riemannian

gradient of J at (D, P) by $G := (\nabla_J(D), \nabla_J(P)) \in T_{(D,P)}\mathcal{S}(m,r) \times Gr(l,r)$, and by $H \in T_{(D,P)}\mathcal{S}(m,r) \times Gr(l,r)$ the conjugate gradient direction. Given $\dim \mathcal{S}(m,r) = r(m-1)$ and $\dim Gr(l,r) = l(r-l)$, we summarise a conjugate gradient algorithm for maximising the function J as defined in (13), cf. Algorithm 1.

For updating the direction parameter γ in Step 5, we confine ourselves to a formula, which is proposed in [23], as

$$\gamma = \frac{\left\langle G^{(j+1)}, G^{(j+1)} - G^{(j)} \right\rangle}{\left\langle H^{(j)}, G^{(j)} \right\rangle}, \tag{14}$$

with the inner product $\langle \cdot, \cdot \rangle$ being defined as

$$\langle (P_1, Q_1), (P_2, Q_2) \rangle = \mathrm{tr}(P_1^\top P_2) + \mathrm{tr}(Q_1^\top Q_2). \tag{15}$$

Finally in Step 6, the search direction is periodically reset to the gradient, in order to achieve fast convergence.

4. Applications of the SparLow Model

In the previous section, we propose a generic regularised cost function, and develop a geometric conjugate gradient algorithm to maximise the generic cost function J. In what follows, we present counterparts of three classic unsupervised learning methods, namely, PCA, LLE, and Laplacian Eigenmap. Experimental evaluations are conducted in Section 5, to illustrate the performance of our proposed framework, in comparison to several direct competitors.

4.1. PCA-like SparLow

The standard PCA method computes an orthogonal transformation $U \in St(l,m)$ such that the variance of the low dimensional representations is maximised, i.e. U is the maximiser of the problem

$$\max_{U \in St(l,m)} \mathrm{tr}\left(U^\top X H_n X^\top U\right), \tag{16}$$

where $H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centring matrix with $\mathbf{1}_n = [1, \cdots, 1]^\top \in \mathbb{R}^n$. In the framework of trace quotient, the denominator can be trivially considered to be $\mathrm{tr}(U^\top B_{pca} U)$ with $B_{pca} = \mathrm{tr}(X H_n X^\top) I_n$, which is a constant. By adopting the sparse representations $\Phi(D, X)$, we construct straightforwardly

$$\mathcal{A}_{pca}(\Phi(D,X)) := \Phi(D,X) H_n (\Phi(D,X))^\top, \tag{17}$$

and

$$\mathcal{B}_{pca}(\Phi(D,X)) := \mathrm{tr}(\Phi(D,X) H_n (\Phi(D,X))^\top) I_r. \tag{18}$$

4.2. LLE-like SparLow

The original LLE method aims to find low dimensional representations of the data via fitting directly the barycentric coordinates of a point based on its neighbours constructed in the original data space, cf. [25]. It is well known that the low dimensional representations in the LLE method can only be computed implicitly. In order to overcome this drawback, the so-called Orthogonal Neighbourhood Preserving Projections (ONPP) is developed in [22], by introducing an explicit orthogonal transformation between the original data and its low dimensional representation.

Specifically, the ONPP method solves the problem

$$\max_{U \in St(l,m)} \mathrm{tr}\left(U^\top X M X^\top U\right), \tag{19}$$

where $M = (I_n - W)^\top (I_n - W)$ with $W \in \mathbb{R}^{n \times n}$ being the matrix of barycentric coordinates of the data. Similar to the previous subsection, we construct the following functions for an LLE-like SparLow approach, i.e.

$$\mathcal{A}_{lle}(\Phi(D,X)) := \Phi(D,X) M (\Phi(D,X))^\top, \tag{20}$$

and

$$\mathcal{B}_{lle}(\Phi(D,X)) := \mathrm{tr}(\Phi(D,X) M (\Phi(D,X))^\top) I_r. \tag{21}$$

4.3. Laplacian SparLow

Another category of DR methods are the ones involving a Laplacian matrix of the data. It includes, for example, Locality Preserving Projection (LPP) [17], Orthogonal LPP (OLPP) [7], Linear Graph Embedding [30]. Similar to the approaches applied in the previous two subsections, we adapt a simple formulation by setting

$$\mathcal{A}_{lap}(\Phi(D,X)) := \Phi(D,X) M (\Phi(D,X))^\top, \tag{22}$$

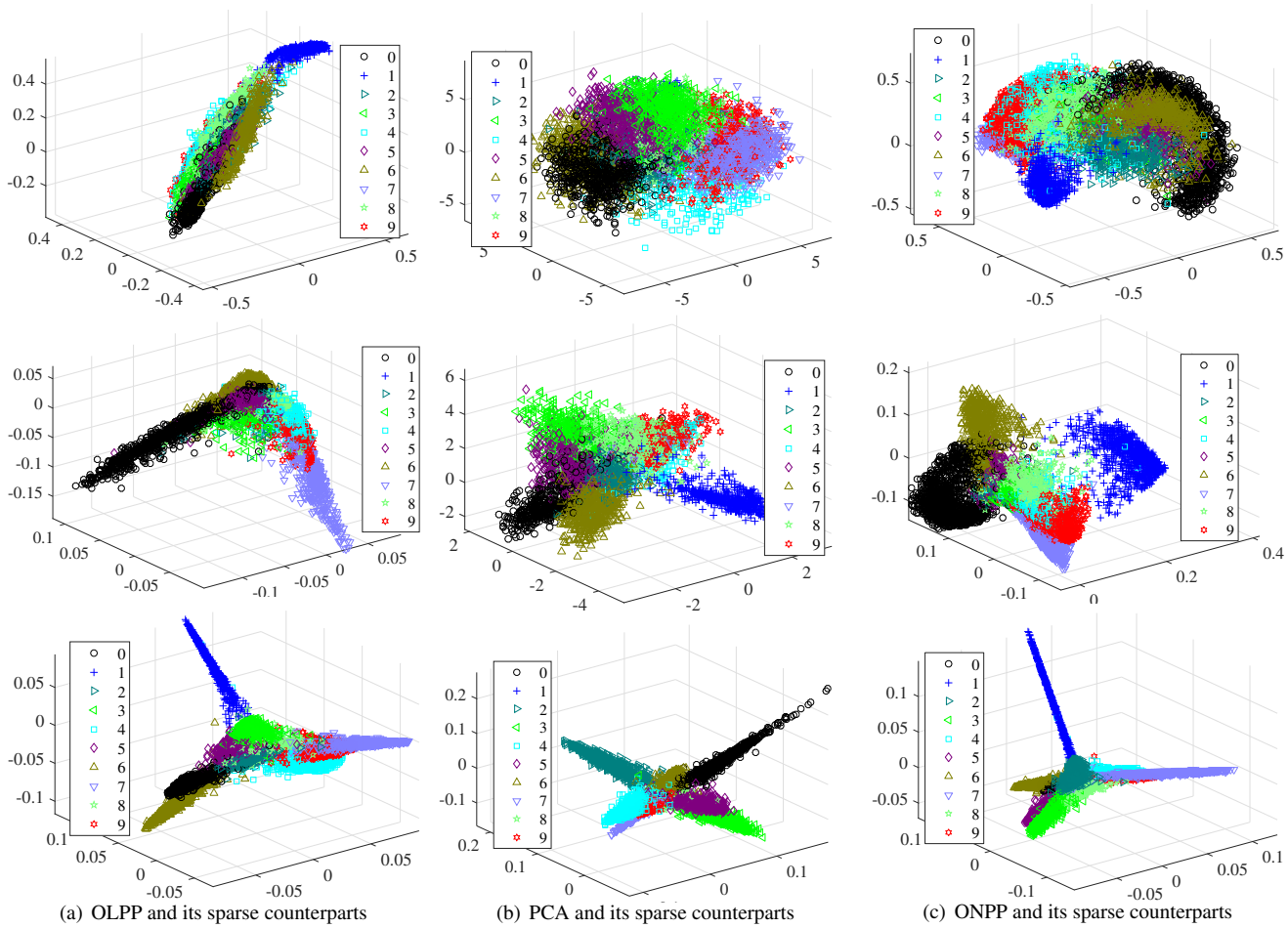
with $M := \{m_{ij}\} \in \mathbb{R}^{n \times n}$ being a real symmetric matrix measuring the similarity between data pairs (x_i, x_j), and

$$\mathcal{B}_{lap}(\Phi(D,X)) := \Phi(D,X) W (\Phi(D,X))^\top, \tag{23}$$

with $W = \{w_{ij}\} \in \mathbb{R}^{n \times n}$ being a diagonal matrix having $w_{ii} := \sum_{j \neq i} m_{ij}$ for all $i = 1, \ldots, n$. Specifically, the similarity matrix M can be computed by applying a Gaussian kernel function on the distance between two data samples, i.e. $m_{ij} = \exp(-\|x_i - x_j\|_2^2 / t)$ or constant weights ($m_{ij} = 1$ if ϕ_i and ϕ_j are adjacent, $m_{ij} = 0$ otherwise).

5. Experimental Evaluations

In this section, we investigate the performance of our proposed *SparLow* methods on real image data. We apply the *SparLow* methods to firstly learn low dimensional representations of real images, and then evaluate their performance in two unsupervised learning scenarios, namely, (1) the 1NN classification using known class labels, cf. [13]; and (2) 3D visualisation of disentangling factors learned by applying the SparLow.



(a) OLPP and its sparse counterparts (b) PCA and its sparse counterparts (c) ONPP and its sparse counterparts

Figure 1. 3D visualization using OLPP, PCA and ONPP on USPS handwritten digits. From top to bottom: Applying OLPP/PCA/ONPP in original space, in sparse space with respect to initial dictionary $\widehat{D}$, and in sparse space with respect to learned dictionary via *SparLow*, respectively.

5.1. Experimental Settings

In the following of the paper, we refer to the three *SparLow* methods proposed in Section 4.1, 4.2, and 4.3 as *PCA-SparLow*, *LLE-SparLow* and *Lap-SparLow*, respectively. Similarly, we refer to direct applications of the classic DR methods on sparse representations that are generated with respect to a fixed dictionary as *SparLDR*. Three members of the *SparLDR* family are investigated in our experiments, namely, *SparPCA*, *SparOLPP*, and *SparONPP*, as the three corresponding counterparts of the *SparLow*.

In our experiments, dictionaries are initialised as a column-wise normalised Gaussian matrix and then improved by employing the K-SVD algorithm [2]. The learned dictionaries $\widehat{D}$'s are used in the regulariser h, as defined in (12). Once an initial dictionary $\widehat{D}$ is given, the orthogonal projection $P \in Gr(l, r)$ can be obtained by applying classical DR methods on the sparse representations. However, when the size of the training dataset is huge, directly performing classical DR methods is often prohibitive. In order to overcome this difficulty, we propose to randomly select a relatively small number of samples, and then to employ the classical DR methods on their sparse representations to obtain an estimation of the initial orthogonal projection $P_0 \in Gr(l, r)$.

Throughout all experiments, we consistently set $\sigma = 10^{-3}$ in Eq. (10). We treat each image as an m-dimensional vector, and normalise it to one. Let n be the number of all signals which contain c classes, we use n_{train}, n_{test} to denote the number of total training samples and the number of total testing samples, respectively. Usually, we set $n = n_{\text{train}} + n_{\text{test}}$.

5.2. Handwritten digit images

Our first experiment is performed on the handwritten digits from the MNIST database [1] and the USPS [20]. The

[1] http://yann.lecun.com/exdb/mnist/

MNIST database consists of $60,000$ handwritten digits images for training and $10,000$ digits images for testing. All images are grayscale between 0 and 1 and have a uniform size of 28×28 pixels. The USPS database has $7,291$ training images and $2,007$ testing images of size (16×16). By vectorising the pixel intensity values of the images, each image is represented as a vector of dimension $m = 784$ or $m = 256$ for the MNIST database and the USPS database, respectively.

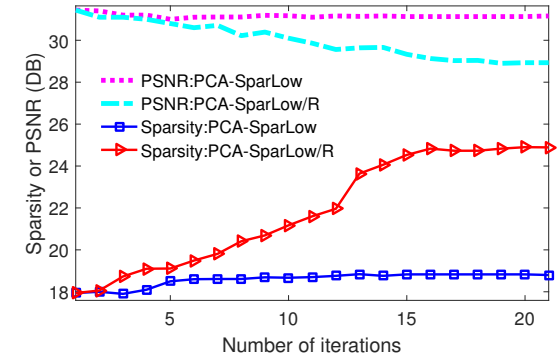
In this experiment, the parameters for elastic net are set to be $\lambda_1 = 0.2$, $\lambda_2 = 2 \times 10^{-5}$, and $\mu_1 = 5 \times 10^{-3}$, $\mu_2 = 2.5 \times 10^{-4}$, for both experiments on the MNIST and USPS datasets. The size of dictionary is chosen to be $r = 1000$. For CS-PCA [12], we employ one common strategy of randomly choosing a certain number of data points as dictionary in a given set of training data, cf. [29].

To demonstrate the effectiveness of the proposed algorithms, experiments of 3D visualisation were conducted on the USPS dataset, compared to the classic DR methods and the SparLDR methods, see Fig. 1. It is easily seen that the low dimensional representations captured in the original data space, shown in the first row in Fig. 1, are very hard to cluster or group. In particular, the boundary between each pair of digits are completely entangled. Direct applications on the sparse representations for a given dictionary, i.e. the second row in Fig. 1, show a significant improvement in disentangling the class information. Furthermore, it is evidentially clear that visualisation powered by the *SparLow*, i.e. the third row in Fig. 1, leads to direct clustering of the handwritten digits.

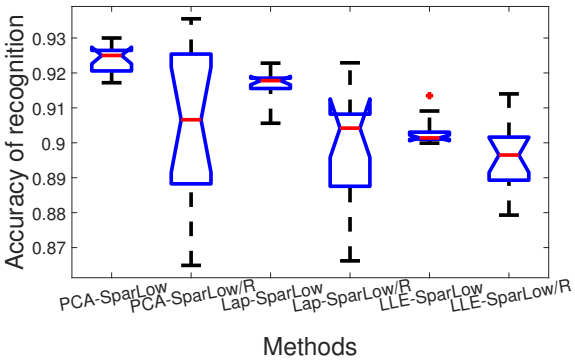
Table 1. Classification Performance for the MNIST & USPS datasets of the Proposed *SparLow* methods, with comparisons to some classical unsupervised DR approaches.

Methods	Accuracy: USPS	Accuracy: MNIST
PCA [12]	$86.36\%, l = 50$	$83.43\%, l = 50$
OLPP [7]	84.10%	82.50%
ONPP [22]	87.38%	83.08%
KPCA [21]	$89.15\%, l = 50$	–
LLE [25]	68.80%	66.09%
LE [21]	71.88%	68.16%
$t-$SNE [27]	77.12%	76.59%
ISOMAP [21]	64.80%	60.51%
GTM [5]	56.92%	60.63%
PCA-SparLow	$92.18\%, l = 50$	$89.42\%, l = 50$
Lap-SparLow	91.80%	87.19%
LLE-SparLow	90.12%	86.55%

Let us denote by δ_i the i^{th} largest eigenvalue of $\Phi(D,X)H_n(\Phi(D,X))^\top$, and further define "Ratio of eigenvalues" in Fig. 4 as $t_l = \sum_{i=1}^{l} \delta_i / \sum_{j=1}^{r} \delta_j$. Fig. (4) shows that our proposed *PCA-SparLow* significantly increase the ratio t_l. It also can be seen, our t_l are consistently

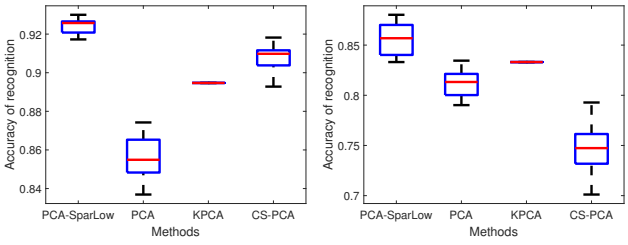


(a) PSNR and Sparsity per image in learning *PCA-SparLow*



(b) Box plot of 1NN classification with or without Regularizations

Figure 2. Performing *SparLow* with or without developed regularizations on USPS database. *PCA-SparLow/R* denotes *PCA-SparLow* without regularizations, and in same way to *Lap-SparLow/R* and *LLE-SparLow/R*.



(a) USPS (b) MNIST

Figure 3. Comparison of 1NN classification using *PCA-SparLow*, PCA, KPCA, CS-PCA on MNIST & USPS database. Dictionary size is 1000.

larger than those of CS-PCA and KPCA, which indicates that the *PCA-SparLow* method captures more structure information, which preserves power in the l dimensional subspace, cf. [7].

One obvious benefit of the proposed *SparLow* model is that the learned low dimensional representations share both reconstructive and discriminative capacities. In this experiment, after applying the *SparLow* methods on the images from the USPS database, we employ the 1NN method

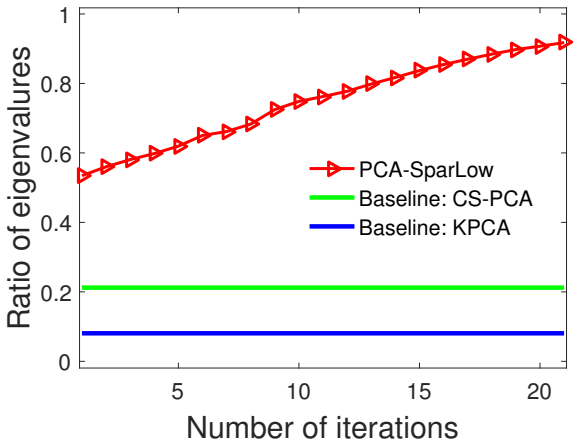


Figure 4. Ratio of top l largest eigenvalues against all eigenvalues in learning process of *PCA-SparLow*.

to classify the reduced features. Reconstruction errors in terms of Peak Signal-to-Noise Ratio (PSNR) are presented in Fig. 2(a). Fig. 2(b) shows the box plot of results of applying the 1NN classification ten times on the USPS database with random initialisations. It is clear that the regulariser h, defined in Eq. (12) has the capability of ensuring good reconstruction, and achieving stable discriminations.

Finally, we compare the *SparLow* methods to several state of the art methods, on the task of 1NN classification. For PCA, KPCA and *PCA-SparLow*, we set $l = 50$, for other methods, we set $l = 20$. For USPS, we use the full training and testing database. For MNIST, we randomly choose $20,000$ images for training, and use standard $10,000$ testing database. According to Fig. 3 and Table 1, it is obvious that the *SparLow* methods consistently outperform the state of the arts.

5.3. CMU PIE faces analysis

In this subsection, we test the *SparLow* methods on the CMU PIE face database [6]. The CMU PIE face database contains 68 human subjects with $41,368$ face images. As suggested in [6], a subset containing $11,554$ PIE faces are chosen, all of which are manually aligned and cropped, thus we nearly get 170 images for each individual, with the scale 32×32 and 256 grey levels per pixel. All experiments are repeated ten times with different randomly selected training and test images, and the average of per-class recognition rates is recorded for each run. In our experiments, we set $\lambda_1 = 10^{-2}, \lambda_2 = 10^{-5}, \mu_1 = 2.5 \times 10^{-4}, \mu_2 = 5 \times 10^{-3}$.

First of all, similar to the experiments conducted on the handwritten digits, Fig. 5 gives the 3D visualisation of low dimensional representations learned by the *SparLow* methods and their classical counterparts. It unveils a same message that the SparLow methods can disentangle the class information very clearly. Fig. 6 illustrates the performance of *LDR*, *SparLDR* and *SparLow* in terms of recognition accuracy. It is easily seen that the *SparLow* methods outperform the state of the art algorithms, such as PCA, OLPP and ONPP.

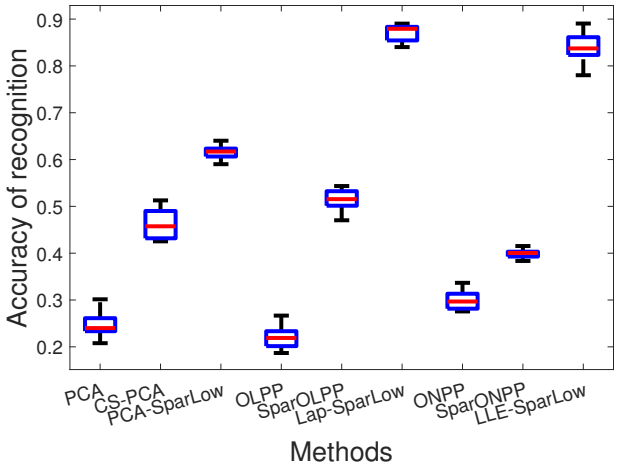


Figure 6. Face recognition on 68 class PIE faces. The classifier is 1NN. Randomly choose $n_{\text{train}} = 8160$ and $n_{\text{test}} = 3394$.

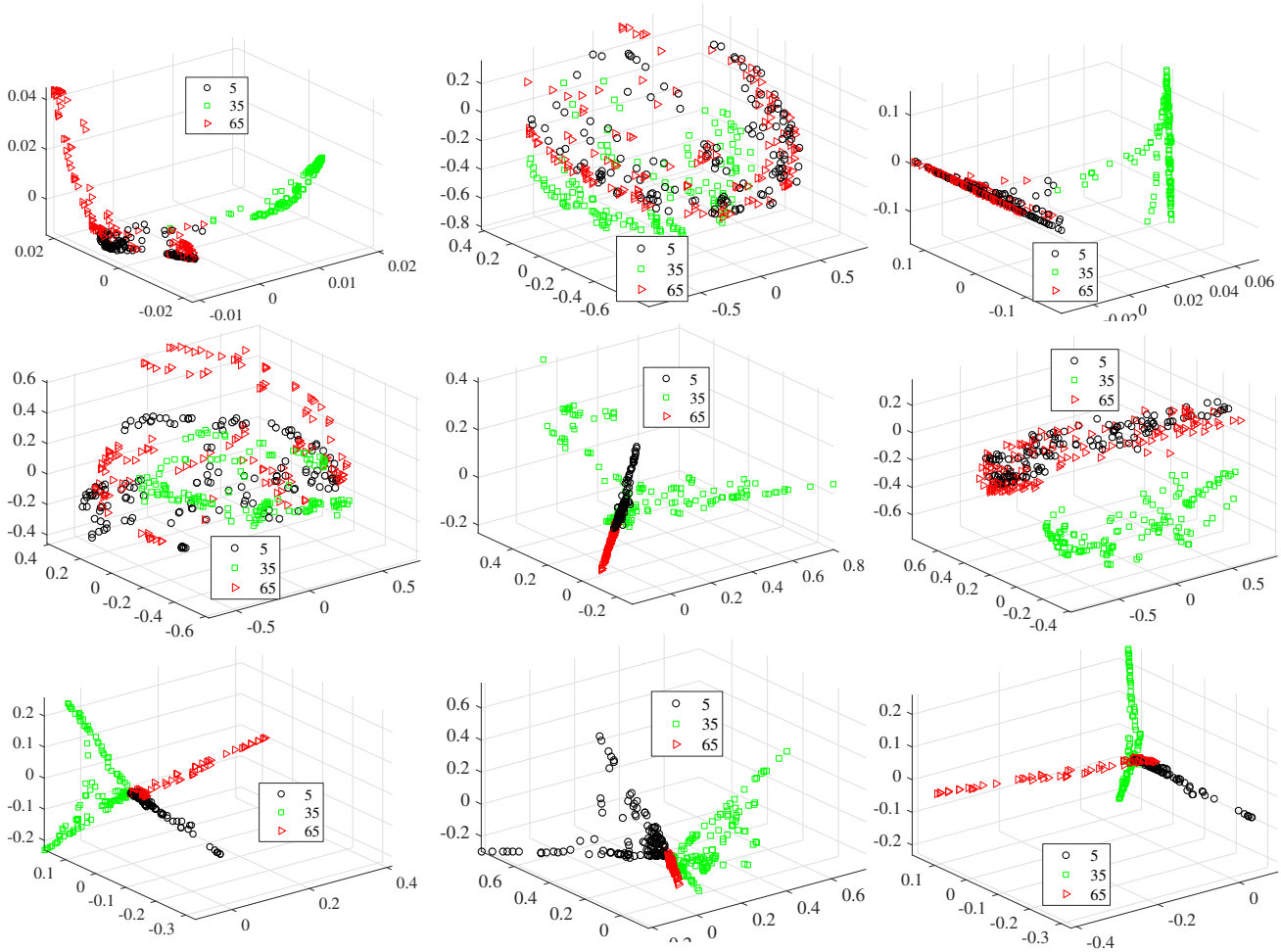
Moreover, visualising the facial features is a common approach to assess the performance of DR methods. In order to facilitate this task, we define the j^{th} disentangling factor v_j as

$$v_j = Du_j \in \mathbb{R}^m, \tag{24}$$

with u_j being the j^{th} column vector of projection matrix U. This construction is similar to the concept of eigenfaces in [3], laplacianfaces in [17], orthogonal laplacianfaces in [7], and orthogonal LLEfaces in [25]. Fig. 7(b) gives the first 10 basis vectors of learned disentangling factors for *PCA-SparLow*, *Lap-SparLow* and *LLE-SparLow*. As for comparison, Fig. 7(a) shows the first 10 eigenfaces, laplacianfaces, and LLEfaces. It shows that (i) our learned facial features are more prominent, especially for laplacianfaces and LLE faces, (ii) our learned facial features captures richer information, such as varying pose and expression (e.g. smile).

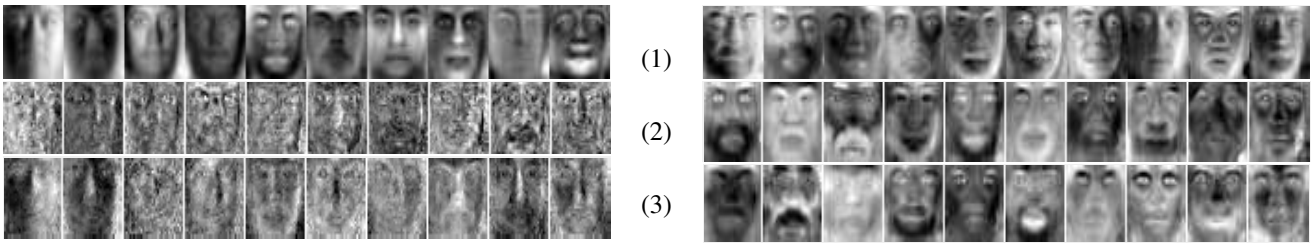
6. Conclusions and Outlooks

In this work, we present an unsupervised low dimensional representation learning approach, coined here as *SparLow*, which leverages both the sparse representation and the trace quotient criterion. It can be considered as a two-step disentangling mechanism, which applies the trace quotient criterion on the sparse representations. Our proposed generic cost function is defined on a sparsifying dictionary and an orthogonal transformation, which form a product Riemannian manifold. A geometric CG algorithm is developed for optimizing the cost function. Our experimental results depict that in comparison with the state of the art unsupervised representation learnings methods, our proposed *SparLow* method possesses promising performance



(a) OLPP and its sparse counterparts (b) PCA and its sparse counterparts (c) ONPP and its sparse counterparts

Figure 5. 3D visualization using OLPP, PCA and ONPP on PIE faces. From top to bottom: Applying OLPP/PCA/ONPP in original space, in sparse space with respect to initial dictionary, and in sparse space with respect to learned dictionary, respectively.



(a) Features extracted from original data (b) Features extracted from sparse representations

Figure 7. Visualisation of facial features. The presented features are generated via Eq.(24). From top to bottom: (1) PCA eigenfaces; (2) Laplacianfaces; (3) LLEfaces.

in data visualisation and 1NN classification. The proposed *SparLow* is flexible and can be extended to more general cases of low dimensional representation learning models with orthogonal constraints.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) through grant number KL 2189/9-1.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[3] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[5] C. M. Bishop, M. Svensén, and C. K. Williams. Gtm: The generative topographic mapping. *Neural computation*, 10(1):215–234, 1998.

[6] D. Cai. *Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning*. PhD thesis, University of Illinois at Urbana-Champaign, 2009.

[7] D. Cai, X. He, J. Han, and H.-J. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 15(11):3608–3614, 2006.

[8] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 2015.

[9] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[12] J. Gao, Q. Shi, and T. S. Caetano. Dimensionality reduction via compressive sensing. *Pattern Recognition Letters*, 33(9):1163–1170, 2012.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[14] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *European Conference on Computer Vision, ECCV*, pages 17–32. Springer, 2014.

[15] S. Hawe, M. Kleinsteuber, and K. Diepold. Cartoon-like image reconstruction via constrained ℓ_p-minimization. In *Proceedings of the 37^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 717–720, 2012.

[16] S. Hawe, M. Seibert, and M. Kleinsteuber. Separable dictionary learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 438–445. IEEE, June 2013.

[17] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[18] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 140–149, 2015.

[19] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 720–729, 2015.

[20] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[21] E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011.

[22] E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2143–2156, 2007.

[23] Y. Liu and C. Storey. Efficient generalized conjugate gradient algorithms, part 1: Theory,. *Journal of Optimization Theory and Applications*, 69(1):129–137, 1991.

[24] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[25] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[26] U. Srinivas, Y. Suo, M. Dao, V. Monga, and T. D. Tran. Structured sparse priors for image classification. *IEEE Transactions on Image Processing*, 24(6):1763–1776, 2015.

[27] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[28] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[29] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[30] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[31] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, 2012.

[32] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3517–3524. IEEE, 2010.

[33] M. Yang, L. Zhang, X. Feng, and D. Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232, 2014.

[34] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.