Kernel Reconstruction ICA for Sparse Representation

Yanhui Xiao, Zhenfeng Zhu, Yao Zhao, *Senior Member, IEEE*, Yunchao Wei, and Shikui Wei

***Abstract*— Independent component analysis with soft reconstruction cost (RICA) has been recently proposed to linearly learn sparse representation with an overcomplete basis, and this technique exhibits promising performance even on unwhitened data. However, linear RICA may not be effective for the majority of real-world data because nonlinearly separable data structure pervasively exists in original data space. Meanwhile, RICA is essentially an unsupervised method and does not employ class information. Motivated by the success of the kernel trick that maps a nonlinearly separable data structure into a linearly separable case in a high-dimensional feature space, we propose a kernel RICA (kRICA) model to nonlinearly capture sparse representation in feature space. Furthermore, we extend the unsupervised kRICA to a supervised one by introducing a class-driven discrimination constraint, such that the data samples from the same class are well represented on the basis of the corresponding subset of basis vectors. This discrimination constraint minimizes inhomogeneous representation energy and maximizes homogeneous representation energy simultaneously, which is essentially equivalent to maximizing between-class scatter and minimizing within-class scatter at the same time in an implicit manner. Experimental results demonstrate that the proposed algorithm is more effective than other state-of-the-art methods on several datasets.**

***Index Terms*— Image classification, independent component analysis (ICA), nonlinear mapping, pattern recognition.**

I. Introduction

SPARSITY characterizes a great number of natural and man-made signals [1], and this attribute plays a vital role in many machine learning techniques, such as compressed sensing [2], sparse coding [3], dictionary learning [4], sparse auto-encoder [5], restricted Boltzmann machines (RBMs) [6], and independent component analysis (ICA) [7].

Among these techniques, ICA transforms an observed multidimensional random vector into components that are statistically as independent from each other as possible. Specifically, a general principle for estimating independent components is to maximize non-Gaussianity [7]. This principle is based on central limit theorem, which states that the summation of independent random variables is more similar to Gaussian than any of the original random variables, i.e., non-Gaussian is independent. Sparsity is one form of non-Gaussianity and is dominant in natural images [8]; and therefore, maximizing sparseness in natural images is equivalent to maximizing non-Gaussianity. Thus, maximizing sparseness is generally used to estimate independent components [8], [9]. ICA has been successfully applied to learning sparse representation in computer vision and machine learning tasks [7], [9].

Overcomplete basis, i.e., the number of basis vectors is substantially larger than their dimension, has been advocated because overcomplete representations are sparser, more flexible, and more robust to noise than complete or undercomplete ones. With an overcomplete basis, Coates *et al.* [10] showed that several approaches, such as sparse autoencoders [5], k-means [10], and RBMs [6], can improve classification. However, it is not trivial for standard ICA to learn an overcomplete basis, which puts the ICA at a disadvantage compared with these methods.

The above disadvantage is mainly due to the hard orthonormality constraint of standard ICA on basis matrix W, where rows of W are orthonormalized at each iteration by symmetric orthonormalization, i.e., $W \leftarrow (WW^T)^{-1/2}W$. However, this orthonormalization does not work for learning overcomplete basis. Although a number of alternative orthonormalization procedures [8] can be employed to learn overcomplete basis, these procedures would incur not only additional computational expense but also cumulative errors in alternative calculating processes. To solve these problems, Le *et al.* [9] replaced the orthonormality constraint with a robust soft reconstruction cost for ICA (RICA). RICA can learn an overcomplete basis even on unwhitened data. However, RICA is a linear technique and can therefore only explore sparse representation in the original data space. Moreover, RICA is an unsupervised method and does not make use of label information.

A linear discriminant may not be effective for the majority of real-world data because nonlinearly separable data structure generally exists in original data space [11]. To enhance the discrimination of data representation, kernel trick [11] can

Manuscript received May 13, 2013; revised March 25, 2014; accepted June 26, 2014. Date of publication July 23, 2014; date of current version May 15, 2015. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316400, in part by the National Natural Science Foundation of China under Grant 61025013, Grant 61172129, Grant 61202241, and Grant 61210006, in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT201206, in part by the Program for New Century Excellent Talents in University under Grant 13-0661, and in part by the Fundamental Research Funds for the Central Universities under Grant 2012JBZ012.

Y. Xiao, Z. Zhu, Y. Wei, and S. Wei are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100055, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: xiaoyanhui@gmail.com; zhfzhu@bjtu.edu.cn; 11112065@bjtu.edu.cn; shkwei@bjtu.edu.cn).

Y. Zhao is with the Institute of Information Science, Beijing Jiaotong University, and also with the State Key Laboratory of Rail Traffic Control and Safety, Beijing 100044, China (e-mail: yzhao@bjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2334711

be used to map a nonlinearly separable data structure into a linearly separable case in a high-dimensional feature space. Therefore, we develop a kernel extension of RICA (kRICA) to map a nonlinearly separable data structure into a linearly separable case in feature space. In addition, to exploit class information, we further extend unsupervised kRICA to a supervised one, i.e., d-kRICA, by introducing a class-driven discrimination constraint. This constraint minimizes inhomogeneous representation energy and maximizes homogeneous representation energy simultaneously, thereby resulting in a data sample that is well represented by overcomplete basis vectors from the corresponding class. In essence, it is equivalent to maximizing the between-class scatter and minimizing the within-class scatter simultaneously in an implicit way.

This paper is fundamentally based on our previous work, Discriminative Reconstruction Independent Component Analysis (DRICA) [12], which only minimized inhomogeneous representation energy to learn a discriminative sparse representation. DRICA failed to explicitly control homogeneous energy; thus, the resulting learned sparse representation may lack discrimination ability. Further improvements based on DRICA are as follows.

1) We introduce a class-driven discrimination constraint that is imposed on the sparse representation to explicitly optimize homogeneous and inhomogeneous representation energies simultaneously. Essentially, it is equivalent to maximizing the between-class scatter and minimizing the within-class scatter in an implicit way.
2) To make the discrimination constraint convex and stable, we propose an elastic term and then present a theoretical proof of its convexity. When discrimination constraint is incorporated into the ICA framework, the ICA problem remains an unconstrained convex optimization problem.
3) By utilizing the kernel trick, we replace the linear projection with a nonlinear one so that nonlinearly separable data structure existing in the original data space can be mapped into a linearly separable case in a high-dimensional feature space.

The rest of this paper is organized as follows. In Section II, we review related works on sparse coding and ICA as well as describe the connection between them. We then provide a brief review of reconstruction ICA in Section III. Section IV presents the details of our proposed kRICA, including its optimization and implementation. By incorporating discrimination constraint, kRICA is further extended to a supervised learning method in Section V. Section VI presents extensive experimental results on several datasets. Section VII concludes our study.

II. Related Work

In this section, we review related works in the following aspects: 1) sparse coding and its applications; 2) connection between ICA and sparse coding; and 3) other kernel sparse representation algorithms.

Sparse coding aims to learn an overcomplete basis set such that only a small fraction of the basis vectors would be necessary to reconstruct a given sample. Sparse coding is attracting increasing attention in the field of computer vision because of its plausible statistical theory [13]. Sparse coding has been successfully applied to image denoising [14], image restoration [15], image classification [16]–[18], and face recognition [19], and so on. The success of its application mainly resulted from two factors. First, sparsity ubiquitously exists in many computer vision applications. For example, in image classification, image components can be sparsely reconstructed by utilizing similar components of other images from the same class [16]. Second, images are often corrupted by noises because of sensor imperfection, poor illumination, or communication errors. Moreover, sparse coding can effectively select the related basis vectors to reconstruct the clean image and overcome noises by allowing a degree of reconstruction error.

Sparse coding is closely related to ICA because the estimation of ICA is roughly equivalent to sparse coding if the components are constrained to be uncorrelated [8]. In addition, both approaches can learn localized and oriented basis vectors (most of them look like edges) on natural images and can thus produce similar sparse representations. However, the goal of ICA is to learn a basis set such that the representation is as independent as possible, whereas sparse coding aims to make the distribution of obtained representation highly sparse [20].

The above mentioned studies only seek sparse representations of the input data in the original data space, and they do not represent data in a high-dimensional feature space. To solve this problem, Yang *et al.* [21] developed a two-phase kernel ICA algorithm, i.e., whitened kernel principal component analysis plus ICA. Different from [21], another solution [22] was proposed to use the contrast function based on canonical correlations in a reproducing kernel Hilbert space. However, both methods cannot learn overcomplete sparse representation in feature space because of the orthonormality constraint.

III. Reconstruction ICA

For sparse sources, maximizing independence implies maximizing sparsity [8]. Thus, given the unlabeled dataset $X = \{x_i\}_{i=1}^m$ where $x_i \in R^n$, the optimization problem of standard ICA for estimating independent components [7] is generally defined as

$$\begin{aligned}&\min_W \frac{1}{m}\sum_{i=1}^{m} g(Wx_i)\\ &\text{s.t.}\quad WW^T = I \end{aligned}\tag{1}$$

where $g(\cdot)$ is a nonlinear convex penalty function, $W \in R^{K\times n}$ is the basis matrix, K is the number of basis vectors and I is the identity matrix. In addition, the orthonormality constraint $WW^T = I$ is traditionally utilized to prevent the basis vectors in W from becoming degenerate. A widely used smooth penalty function is $g(\cdot) = \log(\cosh(\cdot))$ [8].

However, as previously mentioned, the orthonormality constraint hinders the standard ICA from learning an overcomplete basis. This drawback restricts ICA from scaling to high-dimensional data. Consequently, RICA [9] used a soft reconstruction cost to replace the orthonormality constraint in ICA. By applying this replacement, RICA can be formulated as the following unconstrained problem:

$$\min_{W} \frac{1}{m} \sum_{i=1}^{m} \left(||W^T W x_i - x_i||_2^2 + \lambda g(W x_i) \right) \tag{2}$$

where the parameter $\lambda > 0$ is the tradeoff between reconstruction error and sparsity. By swapping the orthonormality constraint with a reconstruction penalty, RICA could learn sparse representations even on unwhitened data when W is overcomplete.

Nevertheless, penalty g can only produce data representations that are sparse, but not invariant [8]. Thus, RICA [9], [23] replaced it by a L_2 pooling penalty, which encourages pooling features to group similar features together to achieve complex invariances, such as scale and rotational invariances. In addition, L_2 pooling also promotes sparsity for feature learning. In particular, L_2 pooling [24], [25] is a two-layered network with square nonlinearity in the first layer $(.)^2$ and square-root nonlinearity in the second layer $\sqrt{(.)}$

$$g(W x_i) = \sum_{j=1}^{K} \sqrt{\varepsilon + H_j \cdot \left((W x_i) \odot (W x_i) \right)} \tag{3}$$

where H_j is the row of spatial pooling matrix $H \in R^{K \times K}$ fixed to uniform weights, i.e., each element in H is 1, $\odot$ denotes the element-wise multiplication, and $\varepsilon > 0$ is a small constant.

However, RICA is a linear technique that explores sparse representation in the original data space only. In addition, this model just learns the overcomplete basis in an unsupervised manner and does not make use of the association between the training sample and its class, which may not work well for classification tasks. To address these problems, on one hand, we develop a kernel extension of RICA to find the sparse representation in the feature space. On the other hand, we aim to learn a more discriminative basis by introducing class information than unsupervised RICA, which leads better performance of sparse representation in the classification tasks.

IV. Kernel Extension for RICA

Motivated by the success of kernel trick, which maps a nonlinearly separable data structure into a linearly separable case in a high-dimensional feature space [11], we propose a kernel version of RICA, called kRICA, to discover sparse representation in feature space.

A. Model Formulation

Suppose there exists a kernel function $\kappa(\cdot, \cdot)$ induced by a high-dimensional feature mapping $\phi: R^n \rightarrow R^{\mathcal{N}}$, where $n \ll \mathcal{N}$. Given two data points x_i and x_j, $\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ represents a nonlinear similarity between them. Thus, the function maps the data and basis from the original data space to the feature space as

$$\begin{gathered} x \overset{\phi}{\rightarrow} \phi(x) \\ W = [w_1, \ldots, w_K]^T \overset{\phi}{\rightarrow} \mathcal{W} = [\phi(w_1), \ldots, \phi(w_K)]^T. \end{gathered} \tag{4}$$

Furthermore, by substituting the mapped data and basis into (2), we can obtain the following objective function of kRICA:

$$\min_{\mathcal{W}} \frac{1}{m} \sum_{i=1}^{m} (||\mathcal{W}^T \mathcal{W} \phi(x_i) - \phi(x_i)||_2^2 + \lambda g(\mathcal{W} \phi(x_i))). \tag{5}$$

In consideration of its excellent performance in many computer vision applications [11], [26], Gaussian kernel, i.e., $\kappa(x_i, x_j) = \exp(-\gamma ||x_i - x_j||_2^2)$ is used in this paper.[1] Thus, norm ball constraints on basis W in RICA [9] can be removed owing to $\phi(w_i)^T \phi(w_i) = \kappa(w_i, w_i) = 1$.

B. Implementation

Equation (5) is an unconstrained convex optimization problem. To solve this problem, we rewrite the objective as follows:

$$\begin{aligned} f(W) &= \frac{1}{m} \sum_{i=1}^{m} (||\mathcal{W}^T \mathcal{W} \phi(x_i) - \phi(x_i)||_2^2 + \lambda g(\mathcal{W} \phi(x_i))) \\ &= \frac{1}{m} \sum_{i=1}^{m} \Bigg(1 + \sum_{u=1}^{K} \sum_{v=1}^{K} \kappa(w_u, x_i) \kappa(w_u, w_v) \kappa(w_v, x_i) \\ &\qquad - 2 \sum_{u=1}^{K} (\kappa(w_u, x_i))^2 \\ &\qquad + \lambda \sum_{j=1}^{K} \sqrt{\varepsilon + \sum_{u=1}^{K} h_{ju} (\kappa(w_u, x_i))^2} \Bigg) \end{aligned} \tag{6}$$

where w_u and w_v are the rows of basis W, and h_{ju} is the element in the pooling matrix H. Row w_j of W is contained in the kernel $\kappa(w_j, \cdot)$; consequently, the optimization methods in RICA, such as L-BFGS and CG [28], are difficult to directly utilize to compute for the optimal basis. Thus, to solve this problem, we alternatively optimize each row of basis W instead. With respect to each updating row w_p of W, the derivative of $f(W)$ is

$$\begin{aligned} \frac{\partial f}{\partial w_p} = \frac{-\gamma}{m} \sum_{i=1}^{m} \Bigg(& \sum_{v=1}^{K} 4 \kappa(w_p, x_i) \kappa(w_p, w_v) \kappa(w_v, x_i) \\ & \times ((w_p - x_i) + (w_p - w_v)) \\ & - 8 \kappa(w_p, x_i)(w_p - x_i) \\ & + 2\lambda \sum_{j=1}^{K} \frac{h_{jp} \kappa(w_p, x_i)(w_p - x_i)}{\sqrt{\varepsilon + \sum_{v=1}^{K} h_{jv} (\kappa(w_v, x_i))^2}} \Bigg). \end{aligned} \tag{7}$$

Moreover, to compute for the optimal w_p, we set $\partial f / \partial w_p = 0$. Equation (7) is challenging to solve because w_p is contained in kernel function $\kappa(w_p, \cdot)$. Thus, we seek

[1] Kernel trick often brings time consuming steps. To alleviate the memory and computational burdens of kernel trick, Nyström low-rank approximation method [27] can be used.

the approximate solution instead of the exact solution. Inspired by the fixed point algorithm [26], to update w_p in the (q)th iteration, we utilize the result of w_p in the $(q-1)$th iteration to calculate for the part in the kernel function. In addition, we initialize the basis as done before [9]. Denote $w_{p,(q)}$ as w_p in the (q)th iteration and (7) with respect to $w_{p,(q)}$ becomes

$$\begin{aligned}\frac{\partial f}{\partial w_{p,(q)}} \cong \frac{-\gamma}{m}\sum_{i=1}^{m}\Bigg(&\sum_{v=1}^{K} 4\kappa(w_{p,(q-1)}, x_i)\kappa(w_{p,(q-1)}, w_v)\\ &\times \kappa(w_v, x_i)((w_{p,(q)}-x_i)+(w_{p,(q)}-w_v))\\ &- 8\kappa(w_{p,(q-1)}, x_i) \times (w_{p,(q)} - x_i)\\ &+ 2\lambda\sum_{j=1}^{K}\frac{h_{jp}\kappa(w_{p,(q-1)}, x_i)(w_{p,(q)} - x_i)}{\sqrt{\varepsilon + \sum_{v=1}^{K} h_{jv}(\kappa(w_v, x_i))^2}}\Bigg)\\ = 0.&\end{aligned}$$

When all of the remaining rows are fixed, the problem becomes a linear equation of $w_{p,(q)}$, which can be solved straightforwardly.

C. Connection Between kRICA and KSR

Gao *et al.* [26], [29] presented a kernel sparse representation (KSR) coding method, which is similar to kRICA, in a high-dimensional feature space, and its optimization problem is

$$\min_{\mathcal{W}, s_i} \frac{1}{m}\sum_{i=1}^{m}(||\mathcal{W}^T s_i - \phi(x_i)||_2^2 + \lambda||s_i||_1) \tag{8}$$

where $s_i \in R^K$ is the sparse representation of sample x_i. Although the proposed kRICA and KSR are very similar, they have two major differences.

1) The objective of (8) in KSR is not convex, and the basis $\mathcal{W}$ and sparse codes s_i should be optimized, alternatively. By contrast, kRICA only focuses on solving basis $\mathcal{W}$ because sparse representation s_i can be regarded as $\mathcal{W}\phi(x_i)$ [9].
2) The L_1 penalty, $g(s_i) = ||s_i||_1$, is utilized by KSR to optimize sparsity, whereas kRICA uses L_2 pooling instead, thus forcing the pooling feature to group similar features together to achieve invariance while promoting sparsity.

V. Supervised kRICA

Given the labeled training data, our goal is to utilize class information to learn a structured basis set that consists of basis vectors from different basis subsets corresponding to different class labels. Each subset should sparsely represent well for its own class but not for the others. Thus, to learn such basis, we further extend the unsupervised kRICA to a supervised one by introducing a class-driven discrimination constraint, namely d-kRICA.

Mathematically, when sample x_i is labeled as $y_i \in \{1, \ldots, c\}$, where c is the total number of classes, we can further utilize class information to learn a structured basis set $W = [W^{(1)}, W^{(2)}, \ldots, W^{(c)}]^T \in R^{K\times n}$, where $W^{(y_i)} \in R^{k\times n}$ is the basis subset that can well represent the sample x_i belonging to y_ith class rather than to others, k is the number of basis vectors for each subset, and $K = k \times c$. Considering that Wx_i can be regarded as the sparse representation of sample x_i [9], let s_i denote $s_i = Wx_i$.

A. Class-Driven Discrimination Constraint

We aim to utilize class information to learn a structured basis; and thus, we intend that the sample x_i labeled as y_i would only be reconstructed by the basis subset $W^{(y_i)}$ with coefficient s_i. To achieve this goal, an inhomogeneous representation energy constraint [12] was introduced as follows:

$$P_- = ||\rho_{y_i}^- s_i||_2^2 \tag{9}$$

where $\rho_{y_i}^- \in R^K$ is a row vector, and selects the inhomogeneous representation coefficients of s_i, i.e., coefficients corresponding to basis vectors other than those belonging to $W^{(y_i)}$. Minimizing inhomogeneous energy leads to learning a structured basis set whose basis vectors have specific class labels. However, the constraint in (9) fails to explicitly optimize homogeneous energy and may thus cause the learned sparse representation to lack discrimination ability.

To solve this problem, we propose a new class-driven discrimination constraint imposed on both inhomogeneous and homogeneous parts of sparse representation. Specifically, similar to inhomogeneous representation energy, we can define homogeneous representation energy as

$$P_+ = ||\rho_{y_i}^+ s_i||_2^2 \tag{10}$$

where $\rho_{y_i}^+ \in R^K$ selects the homogeneous representation coefficients of s_i. For example, assuming $W = [W^{(1)}, W^{(2)}, W^{(3)}]^T$, $W^{(y_i)} \in R^{2\times n}(y_i \in \{1, 2, 3\})$, and $y_i = 2$, $\rho_{y_i}^+$ and $\rho_{y_i}^-$ can be, respectively, defined as follows:

$$\begin{aligned}\rho_2^+ &= [0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0]\\ \rho_2^- &= [1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1].\end{aligned}$$

Intuitively, we can define the class-driven discrimination constraint function $d(s_i)$ as $P_- - P_+$, indicating that the sparse representation s_i, in terms of basis matrix W, would only concentrate on the basis subset $W^{(y_i)}$. However, this constraint is nonconvex and unstable. To address the problem, we propose to incorporate an elastic term $||s_i||_2^2$ into $d(s_i)$. Thus, $d(s_i)$ is defined as

$$d(s_i) = ||\rho_{y_i}^- s_i||_2^2 - ||\rho_{y_i}^+ s_i||_2^2 + \eta||s_i||_2^2. \tag{11}$$

If $\eta \geq k + 1$, $d(s_i)$ is proven to be strictly convex with respect to s_i. Please see the Appendix for a detailed proof. By minimizing the objective (11), it maximizes homogeneous representation energy and minimizes inhomogeneous representation energy simultaneously. In essence, this is equivalent to maximizing the between-class scatter and minimizing the within-class scatter simultaneously in an implicit way.

By incorporating the discrimination constraint into the kRICA framework (d-kRICA), we can obtain the following objective function:

$$\min_{\mathcal{W}} \frac{1}{m} \sum_{i=1}^{m} \big(||\mathcal{W}^T \mathcal{W} \phi(x_i) - \phi(x_i)||_2^2 + \lambda g(\mathcal{W}\phi(x_i)) + \alpha d(\mathcal{W}\phi(x_i)) \big) \tag{12}$$

where $\lambda > 0$ and $\alpha > 0$ are scalars controlling the relative contribution of the corresponding terms. Given a test sample, (12) means that the learned basis set can sparsely represent it while demanding its homogeneous representations to be as large as possible and its inhomogeneous representations to be as small as possible. Following kRICA, the optimization problem (12) can be easily solved by the proposed fixed point algorithm.

VI. Experiments

First, we introduce the feature extraction for image classification. Then, we evaluate the performances of our kRICA and d-kRICA for image classification on four public datasets: Caltech 101 [30], CIFAR-10 [10], STL-10 [10], and Caltech 256 [31]. To evaluate the performances of kRICA and d-kRICA for image clustering, we further conduct an experiment on Cambridge ORL dataset.[2] Moreover, we study the selections of tuning parameters and kernel functions for our method. Finally, we present the similarity matrix to further illustrate the performances of kRICA and d-kRICA.

TABLE I

Image Classification AC on Caltech 101 Dataset

	Training Size	15	30
Non-kernel Methods	ScSPM [16]	67.0%	73.2%
	D-KSVD [4]	65.1%	73.0%
	LC-KSVD [18]	67.7%	73.6%
	RICA [9]	67.1%	73.7%
	DRICA [12]	67.8%	74.4%
	d-RICA	68.7%	75.6%
Kernel Methods	KICA [21]	65.2%	72.8%
	KSR [26]	67.9%	75.1%
	d-KSR	70.0%	76.5%
	kRICA	68.2%	75.4%
	d-kRICA	**71.3%**	**77.1%**

TABLE II

Image Classification AC on CIFAR-10 Dataset

	Model	Accuracy
Non-kernel Methods	ILCC [17]	74.5%
	CDBN (2 layers) [33]	78.9%
	Sparse auto-encoder [10]	73.4%
	Sparse RBM [10]	72.4%
	K-means (Hard) [10]	68.6%
	K-means (Triangle) [10]	77.9%
	K-means (Triangle, 4000 features) [10]	79.6%
	RICA [9]	81.4%
	DRICA [12]	82.1%
	d-RICA	82.9%
Kernel Methods	KICA [21]	78.3%
	KSR [26]	82.6%
	d-KSR	83.5%
	kRICA	83.4%
	d-kRICA	**84.5%**

A. kRICA for Image Classification

1) Image Patch Feature Extraction: Given a $p \times p$ input image patch (with d channels) $x \in R^n$ ($n = p \times p \times d$), kRICA can transform the image patch to a new representation $s = \mathcal{W}\phi(x_i) \in R^K$ in the feature space, where p is termed as the receptive field size. For an image with $N \times M$ pixels (with d channels), we can obtain a $(N-p+1) \times (M-p+1)$ (with K channels) feature following the same setting in [9] by estimating the representation for each $p \times p$ subpatch of the input image. To reduce the dimensionality of the image representation, we utilize the same pooling method in [9] to form a reduced $4K$-dimensional pooled representation for image classification. Given the pooled feature for each image, we utilize linear support vector machine for classification. To facilitate fair comparison, we only use the image patch features for Caltech 101, CIFAR-10, and STL-10 datasets. For Caltech 256 dataset, the SIFT feature [32] is utilized.

2) Classification on Caltech 101: Caltech 101 dataset consists of 9144 images which are divided into 101 object classes and one background class, including animals, vehicles, and so on. Following the common experiment setup [16], we implement our algorithm on 15 and 30 training images per category with basis size $K = 1020$ and 10×10 receptive fields, respectively. Table I shows the results of the comparison. We compare our classification accuracy (AC) with ScSPM [16], D-KSVD [4], LC-KSVD [18], RICA [9], KICA [21], KSR [26], and DRICA [12].

DRICA focused only on minimizing inhomogeneous representation energy, whereas the proposed discrimination constraint (11) optimizes both inhomogeneous and homogeneous representation energies, simultaneously. To compare with DRICA and KSR, we incorporate the constraint (11) into their frameworks, i.e., d-RICA and d-KSR. Table I shows that kRICA and d-kRICA outperform other corresponding approaches.

3) Classification on CIFAR-10: The CIFAR-10 dataset includes 10 categories, such as airplane, automobile, truck, horse, and so on, and a total of 60 000 32×32 color images with 6000 images per category. In addition, there are 50 000 training images and 10 000 testing images. Specifically, 1 000 images from each class are randomly selected as test images and the other 5000 images from each class are selected as training images. In this experiment, we fix the size of basis set to 4000 with 6×6 receptive fields followed by [10]. We compare our approach with ILCC [17], CDBN [33], sparse auto-encoder [10], sparse RBM [10], k-means [10], RICA, DRICA, d-RICA, KICA, KSR, and d-KSR.

Table II shows the effectiveness of our proposed kRICA and d-kRICA.

4) Classification on STL-10: STL-10 has 10 categories, (e.g., airplane, dog, monkey, ship, and so on); each category contains 500 training images and 800 test images, and each image is a color image with 96×96 pixels. Meanwhile, this

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

TABLE III
IMAGE CLASSIFICATION AC ON STL-10 DATASET

	Model	Accuracy
Non-kernel Methods	Raw pixels [10]	31.8%
	K-means(Triangle 1600 features) [10]	51.5%
	RICA(8x8 receptive field) [9]	51.4%
	RICA(10x10 receptive field) [9]	52.9%
	DRICA [12]	54.2%
	d-RICA	54.8%
Kernel Methods	KICA [21]	51.1%
	KSR [26]	54.4%
	d-KSR	55.7%
	kRICA	55.2%
	d-kRICA	**56.9%**

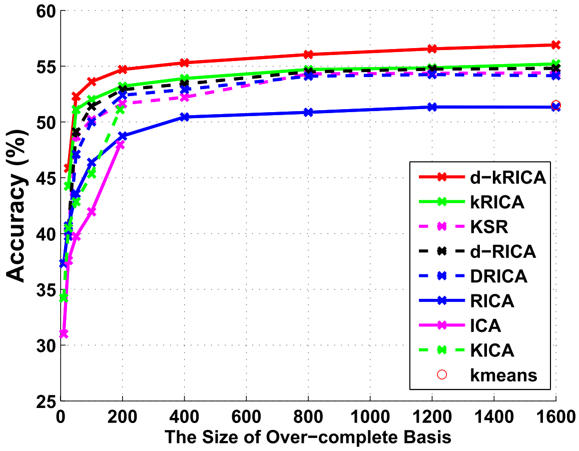


Fig. 1. Classification performance on the STL-10 dataset with varying basis size and 8 × 8 receptive fields.

dataset includes 100 000 extra unlabeled images for unsupervised learning. In our experiments, we set the size of basis set as $K = 1600$ and 8×8 receptive fields in the same manner described in [9]. In accordance with the recommended STL-10 testing protocol [10], d-kRICA performs supervised training on each of the 10 supervised training folds and reports the mean AC on the full test set across the 10 predefined training folds.

Table III shows the classification results of the raw pixels [10], k-means, RICA, KSR, DRICA, d-RICA, kRICA, and d-kRICA.

As can be seen, d-RICA achieves better performance than DRICA on all of the datasets. It is because DRICA only minimizes inhomogeneous representation energy for structured basis learning, whereas d-RICA simultaneously maximizes homogeneous representation energy and minimizes inhomogeneous representation energy, which makes the learned sparse representation much more discriminative. Although both DRICA and d-RICA introduce class information, unsupervised kRICA still performs better than both these algorithms in most cases. kRICA implies much discriminative power for classification by representing the data in the feature space. In addition, since kRICA and d-kRICA utilize the L_2 pooling instead of L_1 penalty to achieve feature invariance; and thus, it achieves better performance than KSR and d-KSR. Furthermore, the d-kRICA achieves better performance than kRICA in all cases because of the integration of class information.

TABLE IV
IMAGE CLASSIFICATION AC ON CALTECH 256 DATASET

	Training Size	15	30	60
Non-kernel Methods	ScSPM [16]	27.7%	34.1%	40.1%
	RICA [9]	28.3%	34.7%	40.6%
	DRICA [12]	28.9%	35.3%	41.1%
	d-RICA	29.7%	36.1%	41.5%
Kernel Methods	KSR(Gaussian) [26]	29.8%	35.7%	40.3%
	KSR(HIK) [29]	33.6%	40.6%	47.0%
	d-KSR(Gaussian)	31.4%	37.9%	42.6%
	d-KSR(HIK)	35.1%	41.4%	47.7%
	kRICA (Gaussian)	30.2%	36.4%	41.1%
	kRICA(HIK)	33.9%	41.1%	47.2%
	d-kRICA(Gaussian)	32.5%	38.7%	43.1%
	d-kRICA (HIK)	**36.5%**	**42.2%**	**48.1%**

We also investigate the effect of basis size for our proposed kRICA and d-kRICA on STL-10 dataset. In our experiments, we try seven sizes: 50, 100, 200, 400, 800, 1200, and 1600. As shown in Fig. 1, the classification accuracies of d-kRICA and kRICA continue to increase when the basis size increases up to 1600 and the performances augment slightly from the basis size of 800. In particular, d-kRICA outperforms all other algorithms.

5) Classification on Caltech 256: Compared with Caltech 101, Caltech 256 is a more challenging dataset in both image content and dataset scale. Caltech 256 contains 29 780 images, which are divided into 256 categories. More categories inevitably decrease both the ratio of within-class similarity to between-class similarity, and classification performance. Following the work of KSR [29], we use the SIFT feature [32] with dense grid sampling strategy for local patch characterization to fix the step size and patch size to 8 and 16, respectively. In addition, maximum pooling is performed on sparse codes.

Given that histogram intersection kernel (HIK), i.e., $\kappa(x, y) = \sum_i \min(x_i, y_i)$, reports the best result on SIFT features in [29]. We also implement our algorithm with HIK and Gaussian kernel for Caltech 256 data. In addition, we test our algorithm on 15, 30, and 60 training images per category with basis size of $K = 4096$. Table IV shows the performances of all the compared methods. Specifically, HIK achieves better performance on SIFT features than Gaussian kernel. Moreover, our proposed d-kRICA with HIK outperforms all of the other algorithms.

B. kRICA for Image Clustering

To evaluate the performance of the proposed method for image clustering, we further conduct an experiment on Cambridge ORL dataset.

1) Evaluation Metrics: Clustering results are usually evaluated by comparing the cluster label of each sample with its ground truth. Similar to [34], two standard clustering metrics, AC, and normalized mutual information (NMI) metric, are utilized to measure the clustering performance. Given a dataset with n images, for each image x_i,

TABLE V
CLUSTERING RESULTS ON CAMBRIDGE ORL DATASET

		AC		NMI	
	b	5	10	5	10
Non-kernel Methods	K-means	78.8%	69.9%	75.4%	73.1%
	CNMF [34]	80.2%	77.0%	76.7%	81.8%
	RICA [9]	87.4%	76.5%	82.3%	80.4%
	DRICA [12]	88.5%	77.9%	83.1%	82.3%
	d-RICA	88.7%	78.7%	83.4%	82.8%
Kernel Methods	KICA [21]	83.5%	74.4%	80.7%	77.9%
	KSR [26]	88.6%	78.5%	83.1%	82.7%
	d-KSR	89.1%	79.8%	84.3%	83.4%
	kRICA	88.9%	79.2%	83.5%	83.0%
	d-kRICA	**89.8%**	**80.3%**	**84.5%**	**83.8%**

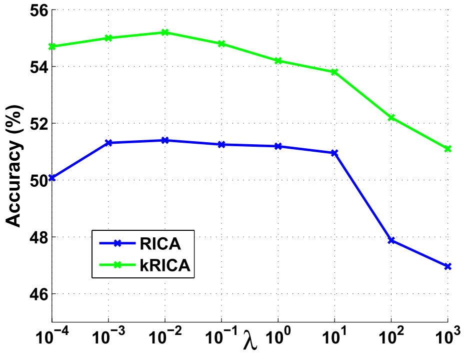


Fig. 2. Relationship between the weight of sparsity term (λ) and classification AC on STL-10 dataset.

let e_i and r_i denote the cluster label and the label provided by the database, respectively. The metric AC is defined as follows:

$$\mathrm{AC} = \frac{\sum_{i=1}^{n} \delta(r_i, \mathrm{map}(e_i))}{n} \tag{13}$$

where $\delta(x, y)$ is the delta function, which is equal to 1 if $x = y$, and is equal to 0, otherwise. $\mathrm{map}(e_i)$ is the mapping function that maps each cluster label e_i to the best label from the database. The best mapping can be found by employing Kuhn–Munkres algorithm [35].

Let C denote the set of clusters obtained from the ground truth and $\tilde{C}$ obtained from our algorithm. Their mutual information metric $MI(C, \tilde{C})$ is defined as follows:

$$\mathrm{MI}(C, \tilde{C}) = \sum_{c_i \in C, \tilde{c}_j \in \tilde{C}} p(c_i, \tilde{c}_j) \cdot \log \frac{p(c_i, \tilde{c}_j)}{p(c_i) \cdot p(\tilde{c}_j)} \tag{14}$$

where $p(c_i)$ and $p(\tilde{c}_j)$ are the probabilities that an image arbitrarily selected from the dataset belongs to clusters c_i and $\tilde{c}_j$, respectively, and $p(c_i, \tilde{c}_j)$ is the joint probability that the arbitrarily selected image belongs to clusters c_i as well as $\tilde{c}_j$ at the same time. In our experiments, we use the NMI as follows:

$$\mathrm{NMI}(C, \tilde{C}) = \frac{MI(C, \tilde{C})}{\max(H(C), H(\tilde{C}))} \tag{15}$$

where $H(C)$ and $H(\tilde{C})$ are the entropies of C and $\tilde{C}$, respectively. Note that $\mathrm{NMI}(C, \tilde{C})$ ranges from 0 to 1. NMI = 1 if the two sets of clusters are identical, and NMI = 0 if the two sets are independent.

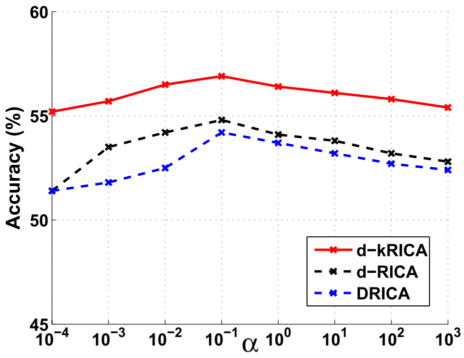


Fig. 3. Relationship between the weight of discrimination constraint term (α) and classification AC on STL-10 dataset.

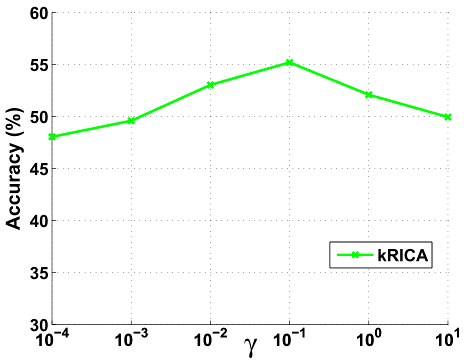


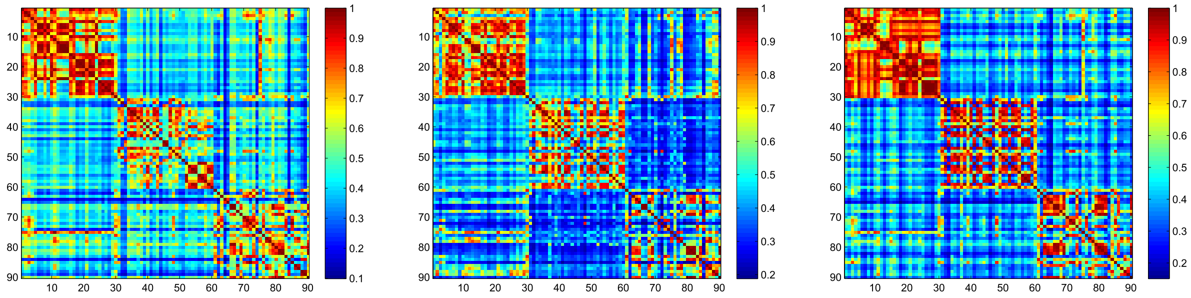
Fig. 4. Classification performance on STL-10 dataset with varying kernel parameter (γ) in Gaussian kernel.

TABLE VI
CLASSIFICATION PERFORMANCES OF DIFFERENT KERNELS ON STL-10 DATASET

Kernel	Accuracy
polynomial kernel	54.2%
inverse distance kernel	38.3%
inverse square distance kernel	47.6%
exponential histogram intersection kernel	36.5%
Gaussian kernel	**56.9%**

2) Clustering on Cambridge ORL: Cambridge ORL dataset contains 10 different images of each 40 distinct subjects, thus, 400 images in total. For some subjects, the images were taken at different times, with varying lighting, facial expressions (open/closed eyes and smiling/not smiling), and facial details (glasses/no glasses).

Following the same setting in CNMF [34], b categories will be randomly picked from the dataset by fixing cluster number b. In addition, all of these images are mixed as the collection X for clustering, and the dimensionality of the new space is set to be the same as the number of clusters b. Afterward, k-means will be used for image clustering on the new data representation. For d-KSR and our d-kRICA, two images are randomly selected from each category in X as training data. The above process will be repeated 10 times, and the average clustering performance is given as the final



(a) (b) (c)

Fig. 5. Similarity matrices for sparse representations of (a) RICA, (b) kRICA, and (c) d-kRICA. The red point means that two images are similar, and the blue point means that two images are dissimilar.

result. Table V shows the effectiveness of the proposed method on ORL dataset. In particular, kRICA and d-kRICA demonstrate better performance than KSR and d-KSR because both utilize L_2 pooling instead of L_1 penalty to achieve feature invariance. Furthermore, d-kRICA achieves better performance than kRICA in all cases by bringing in class information.

C. Tuning Parameters and Kernel Selection

In the experiments, the parameters in kRICA and d-kRICA, i.e., λ, α, and γ in the objective function, are difficult to set. We examine the effect of these parameters in this subsection.

1) Effect of λ: Parameter λ is the weight of sparsity term, which is an important factor in kRICA. To facilitate parameter selection, we experimentally investigate how the performance of kRICA varies with the parameter λ on STL-10 dataset in Fig. 2 (where $\gamma = 10^{-1}$). Fig. 2 shows that kRICA achieves the best performance when λ is fixed at 10^{-2}. Thus, we set $\lambda = 10^{-2}$ for STL-10 data. In addition, we test the AC of RICA under the same sparsity weight. Our proposed nonlinear RICA (kRICA) can consistently outperform linear RICA with respect to λ. Similarly, we experimentally set $\lambda = 10^{-1}$ for Caltech 101 data, $\lambda = 10^{-2}$ for CIFAR-10 data, $\lambda = 10^{-1}$ for ORL data, and $\lambda = 10^{-2}$ for Caltech 256 data.

2) Effect of α: Parameter α controls the weight of the discrimination constraint term. When $\alpha = 0$, the supervised d-kRICA optimization problem becomes the unsupervised kRICA problem. Fig. 3 shows the relationship between the weight of discrimination constraint term α and classification AC on STL-10. d-kRICA evidently achieves the best performance when $\alpha = 10^{-1}$. Hence, we set $\alpha = 10^{-1}$ for STL-10 data. In particular, d-RICA achieves better performance than DRICA in a wide range of α values because DRICA only minimizes inhomogeneous representation energy, whereas d-RICA optimizes both homogeneous and inhomogeneous representation energies for basis learning. Thus, d-RICA makes learned sparse representations more discriminative. Furthermore, by representing the data in feature space, d-kRICA implies more discriminative power for classification and outperforms both these algorithms. Similarly, we set $\alpha = 1$ for Caltech 101 data, $\alpha = 10^{-1}$ for CIFAR-10 data, $\alpha = 10^{-1}$ for ORL data, and $\alpha = 10^{-1}$ for Caltech 256 data.

3) Effect of γ: When utilizing Gaussian kernel in kRICA, it is vital to select the kernel parameter γ, which affects image classification AC. Fig. 4 shows the relationship between γ and classification AC on STL-10 dataset. Therefore, we set $\gamma = 10^{-1}$ for STL-10 data. Similarly, we experimentally set $\gamma = 10^{-2}$ for Caltech 101 data, $\gamma = 10^{-1}$ for CIFAR-10 data, $\gamma = 1$ for ORL data, and $\gamma = 10^{-2}$ for Caltech 256 data.

We also investigate the effect of different kernels for image patch features, i.e., polynomial kernel: $(1+x^T y)^b$, inverse distance kernel: $1/1 + b||x - y||$, inverse square distance kernel: $1/1 + b||x - y||^2$, and exponential HIK: $\sum_i \min(e^{bx_i}, e^{by_i})$.[3] Table VI demonstrates the classification performances of different kernels on STL-10 dataset, and it also shows that Gaussian kernel outperforms the other kernels. Thus, we employ Gaussian kernel for image patch features in our studies.

D. Similarity Analysis

In the previous sections, we have shown the effectiveness of kRICA and d-kRICA for image classification. To further illustrate their performances, we choose 90 images from three classes in Caltech 101, and 30 images for each class. We then compute for the similarity between sparse representations of these images for RICA, kRICA, and d-kRICA. Fig. 5 shows the similarity matrices corresponding to sparse representations of RICA, kRICA, and d-kRICA, respectively. Each element (i, j) in the similarity matrix is the sparse representation similarity (Cosine correlation)[4] between image i and j. A good sparse representation method can make the new representations belonging to the same class more similar; and thus, their similarity matrix should also be block-wise. Fig. 5 shows that nonlinear kRICA takes more discriminative power than linear RICA, and d-kRICA achieves the best performance because of its utilization of class information.

VII. Conclusion

In this paper, we propose a kernel ICA model with a reconstruction constraint (kRICA) to capture the sparse representation in feature space. To exploit the class information, we further extend the unsupervised kRICA to a supervised one by introducing a class-driven discrimination constraint. This constraint leads to learning a structured basis, whose

[3] Following the work in [29], we set $b = 3$ for polynomial kernel and $b = 1$ for the others.

[4] http://en.wikipedia.org/wiki/Cosine_similarity

basis vectors have specific class labels. Each basis vector represents well for its own class but not for the others, which essentially maximizes between-class scatter and minimizes within-class scatter in an implicit manner. Thus, data samples belonging to the same class would have similar sparse representations, thus causing the obtained representation to have more discriminative power. The experiments conducted on standardized datasets demonstrated the effectiveness of the proposed method.

Appendix

We rewrite (11) as

$$\begin{aligned} d(s_i) &= ||\rho_{y_i}^- s_i||_2^2 - ||\rho_{y_i}^+ s_i||_2^2 + \eta||s_i||_2^2 \\ &= Tr[s_i^T \rho_{y_i}^{-T} \rho_{y_i}^- s_i - s_i^T \rho_{y_i}^{+T} \rho_{y_i}^+ s_i + \eta s_i^T s_i]. \end{aligned} \tag{16}$$

Then, we can obtain its Hessian matrix $\nabla^2 d$ with respect to s_i

$$\nabla^2 d = 2\rho_{y_i}^{-T} \rho_{y_i}^- - 2\rho_{y_i}^{+T} \rho_{y_i}^+ + 2\eta I. \tag{17}$$

Without loss of generality, we assume

$$\begin{aligned} \rho_{y_i}^+ &= [0 \ \cdots \ 0 \ \overbrace{1 \ \cdots \ 1}^{k} \ 0 \ \cdots \ 0 \ ] \in R^K \\ \rho_{y_i}^- &= [1 \ \cdots \ 1 \ \overbrace{0 \ \cdots \ 0}^{k} \ 1 \ \cdots \ 1 \ ] \in R^K. \end{aligned}$$

After some derivations, we have $\nabla^2 d = 2A$, where

$$A = \begin{bmatrix} \eta+1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & \eta+1 & 0 & \cdots & 0 & 1 & \cdots & 1 \\ 0 & \cdots & 0 & \eta-1 & \cdots & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & \cdots & \eta-1 & 0 & \cdots & 0 \\ 1 & \cdots & 1 & 0 & \cdots & 0 & \eta+1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & \eta+1 \end{bmatrix}.$$

The convexity of $d(s_i)$ depends on whether its Hessian matrix $\nabla^2 d$, i.e., matrix A, is positive definite or not [36]. Meanwhile, the $K \times K$ matrix A is positive definite if and only if $z^T Az > 0$ for all nonzero vectors $z \in R^K$ [37].

Let the size of upper left matrix in A be $t \times t$, and suppose $z = [z_1, \ldots, z_t, z_{t+1}, \ldots, z_{t+k}, z_{t+k+1}, \ldots, z_K]^T$. Then, we have

$$Az = \begin{bmatrix} (\eta+1)z_1 + z_2 + \cdots + z_t + z_{t+k+1} + \cdots + z_K \\ \vdots \\ z_1 + z_2 + \cdots + (\eta+1)z_t + z_{t+k+1} + \cdots + z_K \\ (\eta-1)z_{t+1} - z_{t+2} - \cdots - z_{t+k} \\ \vdots \\ -z_{t+1} - z_{t+2} - \cdots + (\eta-1)z_{t+k} \\ z_1 + z_2 + \cdots + z_t + (\eta+1)z_{t+k+1} + \cdots + z_K \\ \vdots \\ z_1 + z_2 + \cdots + z_t + z_{t+k+1} + \cdots + (\eta+1)z_K \end{bmatrix}.$$

Furthermore, we can obtain

$$\begin{aligned} z^T Az = {} & (\eta+1)\sum_{i=1}^{t} z_i^2 + (\eta-1)\sum_{i=t+1}^{t+k} z_i^2 + (\eta+1)\sum_{i=t+k+1}^{K} z_i^2 \\ & + 2 \sum_{\substack{1 \le i \le t-1 \\ 2 \le j \le t \\ i < j}} z_i z_j + 2 \sum_{\substack{1 \le i \le t \\ t+k+1 \le j \le K}} z_i z_j + 2\sum_{\substack{t+1 \le i \le t+k-1 \\ t+2 \le j \le t+k \\ i<j}} z_i z_j \\ & - 2 \sum_{\substack{t+1 \le i \le t+k-1 \\ t+2 \le j \le t+k \\ i < j}} z_i z_j = \eta\left(\sum_{i=1}^{t} z_i^2 + \sum_{i=t+k+1}^{K} z_i^2\right) \\ & + (z_1 + z_2 + \cdots + z_t + z_{t+k+1} + \cdots + z_K)^2 \\ & + (\eta-1)\sum_{i=t+1}^{t+k} z_i^2 - 2 \sum_{\substack{t+1 \le i \le t+k-1 \\ t+2 \le j \le t+k \\ i < j}} z_i z_j. \end{aligned}$$

Define function $h(\eta) = z^T Az$. When $\eta \ge k+1$, it is easy to verify that

$$\begin{aligned} h(\eta) \ge h(k+1) = {} & (k+1)\left(\sum_{i=1}^{t} z_i^2 + \sum_{i=t+k+1}^{K} z_i^2\right) \\ & + (z_1 + \cdots + z_t + z_{t+k+1} + \cdots + z_K)^2 + k \sum_{i=t+1}^{t+k} z_i^2 \\ & - 2 \sum_{\substack{t+1 \le i \le t+k-1 \\ t+2 \le j \le t+k \\ i < j}} z_i z_j = k\left(\sum_{i=1}^{t} z_i^2 + \sum_{i=t+k+1}^{K} z_i^2\right) \\ & + (z_1 + \cdots + z_t + z_{t+k+1} + \cdots + z_K)^2 \\ & + \sum_{i=1}^{K} z_i^2 + \sum_{\substack{t+1 \le i \le t+k-1 \\ t+2 \le j \le t+k \\ i < j}} (z_i - z_j)^2. \end{aligned}$$

Since $\sum_{i=1}^{K} z_i^2 > 0$, we have $h(\eta) \ge h(k+1) > 0$. Thus, the Hessian matrix $\nabla^2 d$ is positive definite for $\eta \ge k+1$, which guarantees that $d(s_i)$ is convex with respect to s_i.

Acknowledgment

The authors would like to thank W. Dong and C. Lu for helpful discussions, and acknowledge Q. V. Le for providing the RICA code.

References

[1] E. J. Candes and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.

[4] Q. Zhang and B. Li, “Discriminative K-SVD for dictionary learning in face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, Jun. 2010, pp. 2691–2698.

[5] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in *Proc. Conf. Neural Inform. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 153–160.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley, 2001.

[8] A. Hyvärinen, J. Hurri, and P. Hoyer, *Natural Image Statistics*. Berlin, Germany: Springer-Verlag, 2009.

[9] Q. Le, A. Karpenko, J. Ngiam, and A. Ng, “ICA with reconstruction cost for efficient overcomplete feature learning,” in *Proc. Adv. Neural Inform. Process. Syst.*, 2011, pp. 1017–1025.

[10] A. Coates, H. Lee, and A. Ng, “An analysis of single-layer networks in unsupervised feature learning,” in *Proc. 14th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Sardinia, Italy, 2010, pp. 215–223.

[11] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[12] Y. Xiao, Z. Zhu, S. Wei, and Y. Zhao, “Discriminative ICA model with reconstruction constraint for image classification,” in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, 2012, pp. 929–932.

[13] D. L. Donoho, “For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution,” *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, Jul. 2006.

[14] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Berlin, Germany: Springer-Verlag, 2010.

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Non-local sparse models for image restoration,” in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 2272–2279.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 1794–1801.

[17] K. Yu and T. Zhang, “Improved local coordinate coding using local tangents,” in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 1215–1222.

[18] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent K-SVD,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2011, pp. 1697–1704.

[19] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[20] D. B. Grimes and R. P. Rao, “Bilinear sparse coding for invariant vision,” *Neural Comput.*, vol. 17, no. 1, pp. 47–73, 2005.

[21] J. Yang, X. Gao, D. Zhang, and J. Yang, “Kernel ICA: An alternative formulation and its application to face recognition,” *Pattern Recognit.*, vol. 38, no. 10, pp. 1784–1787, Oct. 2005.

[22] F. R. Bach and M. I. Jordan, “Kernel independent component analysis,” *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, Mar. 2003.

[23] Q. Le *et al.*, “Building high-level features using large-scale unsupervised learning,” in *Proc. 29th Int. Conf. Mach. Learn*, Jul. 2012, pp. 81–88.

[24] Y. LeCun, “Learning invariant feature hierarchies,” in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 496–505.

[25] Y.-L. Boureau, J. Ponce, and Y. LeCun, “A theoretical analysis of feature pooling in visual recognition,” in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 111–118.

[26] S. Gao, I. W.-H. Tsang, and L.-T. Chia, “Kernel sparse representation for image classification and face recognition,” in *Proc. 11th Eur. Conf. Comput. Vis.*, Crete, Greece, 2010, pp. 1–14.

[27] P. Drineas and M. W. Mahoney, “On the Nyström method for approximating a gram matrix for improved kernel-based learning,” *J. Mach. Learn. Res.*, vol. 6, no. 6, pp. 2153–2175, Dec. 2005.

[28] M. Schmidt. (2005). *MinFunc* [Online]. Available: http://www.di.ens.fr/∼mschmidt/Software/minFunc.html

[29] S. Gao, I. W.-H. Tsang, and L.-T. Chia, “Sparse representation with kernels,” *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 423–434, Feb. 2013.

[30] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Apr. 2007.

[31] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” Caltech Technical Report, Pasadena, CA, USA, Tech. Rep. CNS-TR-2007-001, Apr. 2007.

[32] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[33] A. Krizhevsky, “Convolutional deep belief networks on CIFAR-10,” [Online]. Available: http://www.cs.toronto.edu/~kriz/index.html

[34] H. Liu, Z. Wu, X. Li, D. Cai, and T. Huang, “Constrained nonnegative matrix factorization for image representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012.

[35] D. Plummer and L. Lovász, *Matching Theory*. Budapest, Hungary: Akadémiai Kiadó, 1986.

[36] S. Boyd and L. Vandenberghe, “Convex optimization,” *IEEE Trans. Autom. Control*, vol. 51, no. 11, pp. 1859–1864, Nov. 2006.

[37] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 1996.

Yanhui Xiao received the B.S. degree from the Beijing Jiaotong University, Beijing, China, in 2007, where he is currently pursuing the Ph.D. degree.

His current research interests include sparse representation, independent component analysis, matrix factorization, computer vision, and machine learning.

Zhenfeng Zhu received the Ph.D. degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005.

He was a Visiting Scholar with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, USA, in 2010. He is currently an Associate Professor with the Institute of Information Science Beijing Jiaotong University, Beijing. His current research interests include image and video understanding, computer vision, and machine learning.

Yao Zhao (M’06–SM’12) received the B.S. degree in radio engineering from Fuzhou University, Fuzhou, China, in 1989, the M.E. degree in radio engineering from Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He was an Associate Professor at BJTU in 1998, where he became a Professor in 2001. He was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, from 2001 to 2002. He is currently the Director with the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. He serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, Associate Editor of the IEEE SIGNAL PROCESSING LETTERS, and Area Editor of *Signal Processing: Image Communication*.

Dr. Zhao was a recipient of the Distinguished Young Scholar by the National Science Foundation of China in 2010 and Chang Jiang Scholar of the Ministry of Education of China in 2013.

Yunchao Wei received the B.S. and M.S. degrees from the Hebei University of Economics and Business, in 2009, and Beijing Jiaotong University, in 2011, respectively. He is currently pursuing the Ph.D degree from Beijing Jiaotong University. His research interests include machine learning and its application to computer vision and multimedia analysis, e.g. image annotation and cross-media retrieval, etc.

Shikui Wei received the Ph.D. degree in signal and information processing from the Beijing Jiaotong University (BJTU), Beijing, China, in 2010.

He was with the Multimedia Library, Nanyang Technological University (NTU), Singapore, and the China-Singapore Institute of Digital Media, Singapore, from 2008 to 2010, during the Ph.D. degree. He was a Post-Doctoral Researcher with the School of Computer Engineering, NTU, from 2010 to 2011. He is currently an Associate Professor at BJTU. His current research interests include content-based multimedia retrieval, multimedia content analysis, pattern recognition, and video copy detection. He has authored more than 20 research papers in academic journals and conferences, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE International Conference on Multimedia.

He was a recipient of the Excellent Doctoral Dissertation Award by the China Computer Federation in 2011.