Semi-Supervised Robust Dictionary Learning via Efficient $\ell_{2,0^+}$-Norms Minimization

Hua Wang[†], Feiping Nie[‡], Weidong Cai[♯], Heng Huang[‡*]
[†]Colorado School of Mines, Golden, Colorado 80401, USA
[‡]University of Texas at Arlington, Arlington, Texas 76019, USA
[♯]School of Information Technologies, University of Sydney, NSW 2006, Australia
huawangcs@gmail.com, feipingnie@gmail.com, tom.cai@sydney.edu.au, heng@uta.edu

Abstract

Representing the raw input of a data set by a set of relevant codes *is crucial to many computer vision applications. Due to the intrinsic sparse property of real-world data, dictionary learning, in which the linear decomposition of a data point uses a set of* learned *dictionary bases,* i.e.*, codes, has demonstrated state-of-the-art performance. However, traditional dictionary learning methods suffer from three weaknesses: sensitivity to noisy and outlier samples, difficulty to determine the optimal dictionary size, and incapability to incorporate supervision information. In this paper, we address these weaknesses by learning a Semi-Supervised Robust Dictionary (SSR-D). Specifically, we use the $\ell_{2,0^+}$-norm as the loss function to improve the robustness against outliers, and develop a new structured sparse regularization to incorporate the supervision information in dictionary learning, without incurring additional parameters. Moreover, the optimal dictionary size is automatically learned from the input data. Minimizing the derived objective function is challenging because it involves many non-smooth $\ell_{2,0^+}$-norm terms. We present an efficient algorithm to solve the problem with a rigorous proof of the convergence of the algorithm. Extensive experiments are presented to show the superior performance of the proposed method.*

1. Introduction

A crucial part of many computer vision problems is representing the raw input in terms of a set of *codes*, or refined features, which can capture the aspects of the input examples that are relevant to the tasks of interest, such as classification, ranking, tagging, *etc*. Compared to the raw features, these refined features are often more representative and discriminative with lower dimensionality, thus can potentially make the learning tasks easier to deal with and reduce the computational cost. For example, in image tagging, instead of using the raw pixel-wise features, semi-local or patch-based features, such as SIFT and geometric blur, are usually more desirable to achieve better performance. In practice, finding a set of compact features bases, also referred to as *dictionary*, with enhanced representative and discriminative power, plays a significant role in building a successful computer vision system. In this paper, we explore this important problem by proposing a novel formulation and its solution for learning Semi-Supervised Robust Dictionary (SSR-D), where we examine the challenges in dictionary learning, and seek opportunities to overcome them and improve the dictionary qualities.

1.1. Challenges in Dictionary Learning

Recent researches [1, 8] have shown that the linear decomposition of a signal using a few atoms of a *learned* dictionary, instead of a predefined one, usually leads to state-of-the-art results in a number of computer vision applications, such as image annotation [3], face recognition [18], texture classification [10, 9], and many other similar recognition tasks. Although a variety of aspects of dictionary learning have been studied by these prior studies, there still remain three following challenges that hinder the further use of dictionary learning to solve practical problems.

Robustness against outlier samples. Most, if not all, existing dictionary learning algorithms, *e.g.*, [8, 10, 9], routinely used the squared ℓ_2-norm as loss function to measure the reconstruction errors in their optimization objectives. Same as other least square minimization based algorithms in data mining and machine learning, such dictionary learning methods are sensitive to noisy and outlier training samples. Due to the recent explosion of digital media and insufficient human annotations, outliers are abundant in real-world image and video data sets. Therefore, a dictionary robust against noisy and outlier samples is important

*Corresponding author. This project was partially supported by U.S. NSF IIS-1117965, IIS-1302675, IIS- 1344152, and ARC grant.

for achieving good performance in contemporary real-world applications.

Optimal size of a compact dictionary. In traditional sparse learning, motivated by compressed sensing [6], dictionaries are always designed to be over-complete [5]. However, the number of the underlying patterns of most real world data are usually small. Consequently, from information theory perspective, many basis vectors in the dictionary are redundant, which are detrimental to the subsequent sparse solver. Moreover, existing methods usually pre-specify the dictionary size using either heuristics or prior knowledge before learning, whereas a principled way to determine the optimal dictionary size with respect to a specific input data set is seldom studied. Thus, learning a compact and efficient dictionary with automatically determined optimal dictionary size is highly desirable in practice.

Incorporating supervision information. Traditional sparse learning [5] and dictionary learning [8] are designed for a set of signals without human annotations, therefore supervision information are not used even when it is available. In order to make use of the prior labeling knowledge to improve the discriminativity of the learned dictionary, several recent studies [18, 10, 9] have made attempts to incorporate supervision information via additional regularization terms to the original dictionary learning objectives. Despite their successful empirical results, additional terms introduce additional parameters, which inevitably make the corresponding learning models less tractable and harder to fine tune. Therefore, taking advantage of supervision information contained in the training data without incurring extra parameters and keep the learning model compact is another important practical issue in designing an effective dictionary.

1.2. Our Contributions

Among the above three challenges in dictionary learning, the first two are rarely addressed in the literature. Although the third one has been studied in previous works, as pointed out, drawbacks exist which hinder their practical applicability. In this paper, we propose a novel Semi-Supervised Robust Dictionary (SSR-D) learning method to simultaneously address these three challenges, which is interesting from a number of aspects as following.

- We address the dictionary robustness problem by using a new $\ell_{2,0^+}$-norm loss function, which is a generalization of traditional $\ell_{2,1}$-norm loss function [11, 4], yet more robust against noisy and outlier training samples.
- We design a data adaptive dictionary by imposing structured sparsity on the data representation coefficients to automatically select prominent dictionary basis vectors, such that the optimal dictionary size is learned from input data in a principled way and no heuristic pre-specification is required. Mathematically, instead of using the traditional $\ell_{2,1}$-norm regularization to impose structured sparsity, we use the $\ell_{2,0^+}$-norm regularization, which can more closely approximate $\ell_{2,0}$ constraint to better select dictionary bases.
- By further developing the structured sparse regularization for data adaptation, the supervision information of a classification task is gracefully incorporated without incurring additional parameters. Moreover, our new formulation can naturally exploit both labeled and unlabeled data, which makes it a semi-supervised method to achieve better classification performance.
- Because we use multiple terms of $\ell_{2,0^+}$-norms in both the loss function and the regularization, the proposed objective is highly non-smooth therefore difficult to solve in general. We present an efficient algorithm with a rigorous proof of its convergence.
- We conduct extensive empirical evaluations and apply the proposed SSR-D method in several real world applications, where the promising results demonstrate the effectiveness of the proposed method.

2. Learning a Semi-Supervised Robust Dictionary

In this section, we gradually develop our objective to learn a semi-supervised robust dictionary, followed by an efficient algorithm to optimize the proposed objective with a rigorous proof of its convergence. Finally, we describe the classification rules using the learned dictionaries.

Notations and definitions. Throughout this paper, we write matrices as bold uppercase letters and vectors as bold lowercase letters. Given a matrix $\mathbf{M} = [m_{ij}]$, its i-th row and j-th column are denoted as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively.

Given $p > 0$, the ℓ_p-norm[1] of the vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}$. The ℓ_0-norm of the vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_0 = \sum_{i=1}^n |v_i|^0$, which counts the non-zero entries of $\mathbf{v}$.[2]

The Frobenius norm of the matrix $\mathbf{M}$ is denoted as $\|\mathbf{M}\|_{\mathrm{F}}$, and the $\ell_{2,1}$-norm (also called as $\ell_{1,2}$-norm in some research papers) of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i \left\|\mathbf{m}^i\right\|_2$. Given $r > 0$ and $p > 0$, the $\ell_{2,1}$-norm can be generalized to $\ell_{r,p}$-norm as $\|\mathbf{M}\|_{r,p} = \left(\sum_i \left(\sum_j |m_{ij}|^r\right)^{\frac{p}{r}}\right)^{\frac{1}{p}}$, which is a valid norm because it

[1] When $p \geq 1$, $\|\mathbf{v}\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}$ strictly defines a norm that satisfies the three norm conditions, while it defines a quasinorm when $0 < p < 1$. Because the mathematical formulations and derivations in this paper equally apply to both norm and quasinorm, we do not differentiate these two concepts for notation brevity.

[2] Strictly speaking, ℓ_0-norm is not valid norm, the term "norm" used here is for convenience.

satisfies the three norm conditions [11]. Particularly, in this paper, when $r = 2$ and $p \to 0$, we refer to the $\ell_{r,p}$-norm of an input matrix as its $\ell_{2,0^+}$-norm.

2.1. Sparse Coding via ℓ_{0^+}-norm Minimization

Traditional sparse coding tasks deal with the problem to represent an input vector (*e.g.*, the vector representation of an input image) approximately as a weighted linear combination of a small number of "basis vectors" (also called as "codewords" in some literature). Concretely, given an input signal $\mathbf{x} \in \mathbb{R}^d$ and a fixed dictionary consisting of r basis vectors $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_r] \in \mathbb{R}^{d \times r}$ (allowing $r > d$ to make the dictionary over-complete), the task is to learn a new representation $\mathbf{a} \in \mathbb{R}^r$ of the signal $\mathbf{x}$ by minimizing the following objective[14, 1]:

$$J_0(\mathbf{a}) = \|\mathbf{a}\|_0, \quad s.t. \quad \mathbf{x} = \mathbf{D}\mathbf{a} \ . \tag{1}$$

Because minimizing ℓ_0-norm is a combinatorial integer optimization problem, solving the problem in Eq. (1) is NP-hard in general. To tackle this, in practice $\mathbf{a}$ is often learned by minimizing the following objective [14, 8, 10]:

$$J_1(\mathbf{a}) = \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1 \ , \tag{2}$$

where $\lambda > 0$ is a tradeoff parameter to balance the relative importance of the reconstruction error and the sparsity of the learned coefficients. When the input data satisfy the restricted isometry property [5, 14], the ℓ_1-norm regularization in Eq. (2) approximates the ℓ_0-norm constraint in Eq. (1). As a result, the learned $\mathbf{a}$ is sparse with very few non-zero coefficients [5, 14].

Although we can obtain the results by solving Eq. (2), ideally a better approximation to the ℓ_0-norm constraint in Eq. (1) is to use the ℓ_p-norm regularization with a very small p close to 0. Thus we minimize the following objective:

$$J_p(\mathbf{a}) = \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_p^p \ . \tag{3}$$

Obviously, when $p = 1$ Eq. (3) reduces to Eq. (2). It can also be easily verified that, the smaller the value of p is, the closer $\|\mathbf{a}\|_p$ is to $\|\mathbf{a}\|_0$. When $p \to 0$, $\|\mathbf{a}\|_p \to \|\mathbf{a}\|_0$, therefore J_p is able to better approximate J_0 than J_1 in terms of objective value, and could lead to a more sparse $\mathbf{a}$ given the same λ. Note that, when $0 < p < 1$, J_p is quasi-convex, therefore seeking a global optimal solution is still feasible. In this work, we use Eq. (3) for sparse coding and refer to it as ℓ_{0^+}-norm minimization problem because we set the value of p as a small constant that is very close to 0.

Given a data set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ of n training samples $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$, and its sparse coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{r \times n}$ where $\mathbf{a}_i \in \mathbb{R}^r$ is the sparse representation of $\mathbf{x}_i$ with respect to a specific dictionary. We can learn $\mathbf{A}$ by minimizing the following objective:

$$\begin{aligned} J'_p(\mathbf{A}) &= \sum_{i=1}^n \left(\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_p^p \right) \\ &= \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda \|\mathbf{A}\|_{p,p}^p \ . \end{aligned} \tag{4}$$

2.2. Learning Robust Dictionary via $\ell_{2,0^+}$-norm Loss Function

In Eqs. (1–4), the dictionary $\mathbf{D}$ and its basis vectors in the learning objectives are assumed to be fixed. Recently, the advances in sparse coding has shown that linearly decomposing a signal using a few atoms of a *learned* dictionary instead of a predefined one usually leads to state-of-the-art performance for a number of practical computer vision applications [7, 13, 10, 15]. Specifically, we can jointly learn the dictionary $\mathbf{D}$ and the sparse representations $\mathbf{A}$ from the input signals by minimizing the following objective [8, 13, 10]:

$$\begin{aligned} &J'_{\mathrm{D}}(\mathbf{D}, \mathbf{A}) = \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda \|\mathbf{A}\|_{1,1}, \\ &s.t. \quad \|\mathbf{d}_j\|_2 \le 1, \ \forall\, 1 \le j \le r \ , \end{aligned} \tag{5}$$

where the constraints on the ℓ_2-norms of the basis vectors are used to avoid degenerate solutions, because the reconstruction errors in the first term of Eq. (5) are invariant to simultaneously scaling $\mathbf{D}$ by a scalar and $\mathbf{A}$ by its inverse. For notation brevity, we denote the feasible domain of $\|\mathbf{d}_j\|_2 \le 1, \ \forall\, 1 \le j \le r$ as $\mathcal{C}$ in the sequel of this paper.

Same as in Section 2.1, to achieve better sparsity, we learn the dictionary by minimizing the following objective:

$$\min_{\mathbf{D} \in \mathcal{C}} J_{\mathrm{D}}(\mathbf{D}, \mathbf{A}) = \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda \|\mathbf{A}\|_{p,p}^p \ . \tag{6}$$

Again, when $p = 1$ Eq. (6) is reduced to Eq. (5).

In the first term of Eq. (6) the objective J_{D} uses squared ℓ_2-norm to measure reconstruction errors, therefore same as other least square optimization objectives in machine learning and data mining, J_{D} is sensitive to noisy and outlier training samples. Because dictionary learning is typically performed on data sets with large sample sizes where outlier samples are inevitable by nature, the learned dictionary could be seriously biased. Therefore, robustness against outlier training samples needs to be taken into account in dictionary learning for its practical use.

A widely used remedy in statistical learning to address outliers is to use not-squared ℓ_2-norm reconstruction error [11, 16], which minimizes the following objective:

$$\min_{\mathbf{D} \in \mathcal{C}} J'_{\text{R-D}}(\mathbf{D}, \mathbf{A}) = \sum_{i=1}^n \left(\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2 + \lambda \|\mathbf{a}_i\|_p^p \right) \tag{7}$$

where the reconstruction error $\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2$ is not squared, thus outliers have less influence than those in Eq. (6). However, when the number of noisy data samples is big or some

outlier data samples are deviated very far from the true data distribution, a more robust loss function is desired. Motivated by the previous ℓ_p-norm regularization to impose sparsity, we consider to use the ℓ_p-norm to measure the reconstruction errors, by which we minimize the following the following objective:

$$\begin{aligned} J_{\text{R-D}}(\mathbf{D},\mathbf{A}) &= \sum_{i=1}^{n}\left(\|\mathbf{x}_i-\mathbf{D}\mathbf{a}_i\|_2^q+\lambda\|\mathbf{a}_i\|_p^p\right) \\ &= \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,q}^q+\lambda\|\mathbf{A}\|_{p,p}^p, \quad s.t. \quad \mathbf{D}\in\mathcal{C} \ . \end{aligned} \tag{8}$$

Obviously, the smaller the value of q is, the less impact the outlier samples have. Again, when $q=1$, $J_{\text{R-D}}$ degenerates to $J'_{\text{R-D}}$. Empirically, we select a small q that is close to 0, therefore we call the measurement of the reconstruction errors defined in the first term of Eq. (8) as $\ell_{2,0^+}$-norm loss function, which is more robust to outliers than both squared Frobenius norm loss function and $\ell_{2,1}$-norm loss function.

2.3. Learning Adaptive Dictionary

Compared to the standard dictionary learning objective in Eq. (5), our new objective $J_{\text{R-D}}$ in Eq. (8) has better sparsity and improved robustness against noisy and outlier training samples. However, same as J'_{D} in Eq. (5), $J_{\text{R-D}}$ in Eq. (8) suffers from a critical problem that can hinder its practical use. In standard dictionary learning settings, because we usually do not know the optimal dictionary size in a priori and the dictionary is typically designed to be over-complete, many of the basis vectors in the learned dictionary $\mathbf{D}$ are redundant, which makes the computation to obtain sparse representation for subsequent unseen data computationally inefficient. Therefore, selecting only the most relevant dictionary bases by pruning the redundant ones is of essential use to reduce the computational load for practical applications. To this end, we consider to learn a compact dictionary that is adaptive to input data.

Due to the flat nature of the $\ell_{1,1}$-norm regularization in Eq. (5) and the $\ell_{p,p}$-norm regularization in Eq. (8), all the basis vectors in the learned dictionary $\mathbf{D}$ are evenly treated and used in subsequent signal representations. However, because the underlying high-level patterns of input signals are not known beforehand, the dictionary may contain redundancy. Moreover, the dictionary size has to be specified before learning, whereas how to determine the optimal dictionary size in a principled way is rarely studied in literature. To address the both issues, we propose to enforce structured sparsity on $\mathbf{A}$ using $\ell_{2,1}$-norm regularization [12, 2]. As a result, the dictionary size is automatically determined by the learned $\mathbf{A}$ and irrelevant basis vectors are pruned. Specifically, we learn $\mathbf{D}$ and $\mathbf{A}$ from $\mathbf{X}$ by minimizing the following objective:

$$J_{\underset{\mathbf{D}\in\mathcal{C}}{\text{RA-D}}}(\mathbf{D},\mathbf{A}) = \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,q}^q+\lambda\|\mathbf{A}\|_{2,1} \ . \tag{9}$$

Following the same analysis before, when $0<p<1$, $\ell_{2,p}$-norm of a given input matrix is closer to its $\ell_{2,0}$-norm than $\ell_{2,1}$-norm. As a result, $\ell_{2,p}$-norm regularization could better approximate $\ell_{2,0}$-norm constraint to select dictionary bases, by which we minimize the following objective:

$$J_{\underset{\mathbf{D}\in\mathcal{C}}{\text{RA-D}}}(\mathbf{D},\mathbf{A}) = \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,q}^q+\lambda\|\mathbf{A}\|_{2,p}^p \ . \tag{10}$$

In Eq. (10), the $\ell_{2,p}$-norm regularization term $\|\mathbf{A}\|_{2,p}$ penalizes all n representation coefficients (*i.e.*, all entries in $\mathbf{a}^i$) corresponding to one single basis vector of $\mathbf{D}$ as a whole, and compute the ℓ_p-norm of $\mathbf{a} = \left[\left\|\mathbf{a}^1\right\|_2,\ldots,\left\|\mathbf{a}^r\right\|_2\right]^T$. As a result, when $p<2$, sparsity is conferred on $\mathbf{a}$, and the basis vectors in $\mathbf{D}$ corresponding to the non-zero entries of resulted $\mathbf{a}$ are automatically selected for succeeding data representation. To be more precise, let

$$\mathcal{D}_{\mathbf{X}} = \left\{\mathbf{d}_i \;\middle|\; \left\|\mathbf{a}^i\right\|_2 > 0\right\} \ , \tag{11}$$

we construct $\mathbf{D}_{\mathbf{X}} \in \mathbb{R}^{d\times|\mathcal{D}_{\mathbf{X}}|}$ by using all $\mathbf{d}_i \in \mathcal{D}_{\mathbf{X}}$ as its columns. Apparently, the dictionary size $|\mathcal{D}_{\mathbf{X}}|$ is learned form the input data $\mathbf{X}$, but not by pre-specification as in previous works. Because $\mathcal{D}_{\mathbf{X}}$ thereby $\mathbf{D}_{\mathbf{X}}$ is specific to input data $\mathbf{X}$, together with its robustness, we call $\mathbf{D}_{\mathbf{X}}$ as the learned *Robust and Adaptive Dictionary (RA-D)*.

Note that, because $\mathbf{D}_{\mathbf{X}}$ is a subset of $\mathbf{D}$, we call $\mathbf{D}$ as the *super-dictionary*, whose size has to be specified beforehand, same as prior studies. However, as shown later in Section 3.2, the performance of data representation and classification using $\mathbf{D}_{\mathbf{X}}$ is considerably stable in a very large range of the size of $\mathbf{D}$.

2.4. Learning Semi-Supervised Dictionary

Because incorporating the supervision information to learn a discriminative dictionary usually improves the performance of subsequent classifications [10, 9, 3, 18], we further develop the $J_{\text{RA-D}}$ in Eq. (9) to take advantage of label information of an input data set. Different from existing works that incorporate label information by an additional term, we make use of the structural sparsity on the representation coefficient matrix $\mathbf{A}$, such that no extra parameter is required and our model is easier to fine tune.

Given a classification task with K classes, besides the input data $\{\mathbf{x}_i\}_{i=1}^n$, we also have their associated class labels. Let the binary vector $\mathbf{y}_i \in \{0,1\}^K$ represent the labels attached to $\mathbf{x}_i$, we write $\mathbf{Y} = [\mathbf{y}_1,\ldots,\mathbf{y}_n]^T$, such that $y_{ik}=1$ if $\mathbf{x}_i$ belongs to the k-th class, and 0 otherwise. The goal of the classification task is to learn from $\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^n$ a function that is able to predict labels for unseen data points.

For convenience, we write $\mathbf{X}_k \in \mathbb{R}^{d\times n_k}$ as the data matrix of the k-th class consisting all of its n_k training data points. We also write $\mathbf{X}_0$ as the unlabeled data matrix, whose columns are the unlabeled data points. We

denote $\widetilde{\mathbf{X}} = [\mathbf{X}_0, \ldots, \mathbf{X}_K]$. We similarly denote $\widetilde{\mathbf{A}} = [\mathbf{A}_0, \ldots, \mathbf{A}_K]$. Thus, $\widetilde{\mathbf{A}}$ is the sparse coefficient matrix corresponding to $\widetilde{\mathbf{X}}$, and $\mathbf{A}_k$ is the coefficient matrix for the data points belonging to the k-th class and $\mathbf{A}_0$ is that for unlabeled data. We learn $\mathbf{D}$ and $\widetilde{\mathbf{A}}$ from $\widetilde{\mathbf{X}}$ by minimizing the following objective:

$$J_{\text{SSR-D}}\left(\underset{\mathbf{D}\in\mathcal{C}}{\mathbf{D}}, \widetilde{\mathbf{A}}\right) = \left\| \left(\widetilde{\mathbf{X}} - \mathbf{D}\widetilde{\mathbf{A}}\right)^T \right\|_{2,q}^q + \lambda \sum_{k=0}^{K} \|\mathbf{A}_k\|_{2,p}^p \quad , \quad (12)$$

Upon solution, let $\mathcal{D}_k = \left\{ \mathbf{d}_i \mid \left\|\mathbf{a}_k^i\right\|_2 > 0 \right\}$ where $\mathbf{a}_k^i$ is the i-th row of $\mathbf{A}_k$, we construct the k-th class specific dictionary $\mathbf{D}_k \in \mathbb{R}^{d\times|\mathcal{D}_k|}$ using all $\mathbf{d}_i \in \mathcal{D}_k$ as its columns. Obviously, the resulted $\mathbf{D}_k$ is adaptive to both input data and class supervision information of the k-th class. Again, the dictionary size of $\mathbf{D}_k$ is automatically determined by $|\mathcal{D}_k|$. Because we learn the dictionaries from both the labeled and unlabeled data, we call $\mathbf{D}_k\ (1 \le k \le K)$ learned by Eq. (12) as the proposed *Semi-Supervised Robust Dictionary (SSR-D)*. Exploiting both unlabeled and labeled data in a unified framework without incurring extra parameters is an important advantage of the proposed method.

2.5. Optimization Algorithm

Because the $\ell_{2,p}$-norm function is non-smooth, the objective $J_{\text{SSR-D}}$ in Eq. (12) is highly non-smooth due to involving $K+2$ $\ell_{2,p}$-norm terms. Thus, minimizing $J_{\text{SSR-D}}$ is difficult in general by existing algorithms. To solve the problem, we derive an efficient algorithm as summarized in Algorithm 1 and its convergence is guaranteed by the following theorem (the proof is skipped due to space limit and will be provided in the extended version of the paper).

Theorem 1 *The algorithm decreases the objective value in Eq.* (12) *in each iteration.*

Because $J_{\text{SSR-D}}$ in Eq. (12) is obviously lower bounded by 0, Theorem 1 guarantees the convergence of Algorithm 1.

2.6. Classification Using Learned Dictionaries

Given an unlabeled data point $\mathbf{x}$, and the learned dictionaries $\mathbf{D}_k\ (1 \le k \le K)$, we may compute the sparse representation of $\mathbf{x}$ with respect to the k-th class, $\mathbf{a}^{(k)}$, by solving the following problem:

$$\min_{\mathbf{a}^{(k)}} \left\|\mathbf{x} - \mathbf{D}_k\mathbf{a}^{(k)}\right\|_2^2 + \lambda \left\|\mathbf{a}^{(k)}\right\|_1 \quad . \quad (13)$$

Thus the reconstruction error of $\mathbf{x}$ with respect to the k-th class is computed as:

$$\mathbf{e}^{(k)} = \left\|\mathbf{x} - \mathbf{D}_k\mathbf{a}^{(k)}\right\|_2 \quad . \quad (14)$$

Sorting $\mathbf{e}^{(k)}$, we can easily assign labels to $\mathbf{x}$ to the class with minimum reconstruction error:

$$l(\mathbf{x}) = \arg\min_k \ \mathbf{e}^{(k)} \quad . \quad (15)$$

Algorithm 1: An efficient iterative algorithm to minimize the objective value of Eq.(12).

Input: $\widetilde{\mathbf{X}} = [\mathbf{X}_1, \ldots, \mathbf{X}_K] \in \mathbb{R}^{d\times\tilde{n}}$.

1. Initialize diagonal matrices $\mathbf{U}^{(t)} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ and $\mathbf{V}_k^{(t)} (1 \le k \le K) \in \mathbb{R}^{r\times r}$. Initialize $\widetilde{\mathbf{A}}^{(t)} \in \mathbb{R}^{r\times\tilde{n}}$.

while *not converge* **do**

2. Calculate $\hat{\mathbf{X}} = \widetilde{\mathbf{X}}(\mathbf{U}^{(t)})^{\frac{1}{2}}$ and $\hat{\mathbf{A}} = \widetilde{\mathbf{A}}^{(t)}(\mathbf{U}^{(t)})^{\frac{1}{2}}$, and compute:

$$\mathbf{D}^{(t+1)} = \arg\min_{\mathbf{D}\in\mathcal{C}} \left\|(\hat{\mathbf{X}} - \mathbf{D}\hat{\mathbf{A}})^T\right\|_F^2 \quad . \quad (16)$$

3. For each $k (1 \le k \le K)$, calculate the i-th column of $\mathbf{A}_k^{(t+1)}$ by

$$(\mathbf{U}_k)_{ii}^{(t)} \left[(\mathbf{U}_k)_{ii}^{(t)} \left(\mathbf{D}^{(t+1)}\right)^T \mathbf{D}^{(t+1)} + \lambda\mathbf{V}_k^{(t)}\right]^{-1} (\mathbf{D}^{(t+1)})^T (\mathbf{X}_k)_i \ , \quad (17)$$

and construct $\widetilde{\mathbf{A}}^{(t+1)}$ by $\mathbf{A}_k^{(t+1)} (1 \le k \le K)$.

4. Calculate the diagonal matrix $\mathbf{U}^{(t+1)}$, where the i-th diagonal element is $\frac{q}{2}\left\|\tilde{\mathbf{x}}_i - \mathbf{D}^{(t+1)}\tilde{\mathbf{a}}_i^{(t+1)}\right\|_2^{q-2}$.

5. For each $k (1 \le k \le K)$, calculate the diagonal matrix $\mathbf{V}_k^{(t+1)}$, where the i-th diagonal element is $\frac{p}{2}\left\|\left(\mathbf{A}_k^{(t+1)}\right)^i\right\|_2^{p-2}$.

Output: $\mathbf{D} \in \mathbb{R}^{p\times r}$ and $\widetilde{\mathbf{A}} = [\mathbf{A}_1, \ldots, \mathbf{A}_K] \in \mathbb{R}^{r\times\tilde{n}}$.

3. Experimental Results

In this section, we experimentally evaluate a variety of aspects of the proposed methods, where we experiment with following six benchmark data sets: **AT&T** data set, **USPS** data set, **BinAlpha** data set, **Reuters** data set, **TDT** data set, **TRECVID 2005** data set, among which the first five data sets are single-label data set whilst the last one is a multi-label data set.

3.1. Improved Data Representation Capability via $\ell_{2,0^+}$-norm Loss Function and Regularization

Because an important contribution of this paper is to use the $\ell_{2,0^+}$-norm loss function and regularization in the dictionary learning objectives to obtain a dictionary with better sparsity and improved robustness against outlier data samples, we first evaluate its usefulness.

Experimental setups. We experiment with the AT&T face data set. Our goal is to cluster the face images, by which we examine the data representation capability of the learned dictionary when p and q in the proposed dictionary learning objectives vary in the range of 0.1 to 2. For simplicity, we set $p = q$ in our experiments. We set the size of the super-dictionary $\mathbf{D}$ as 400, which is the total number of the images in the data set. In order to evaluate the data representation capability of the learned dictionaries, we conduct the experiment in an unsupervised way. Specifically, we do not assign labels to the data points and conduct K-means clustering on the learned data representations. We learn dictionaries $\mathbf{D}_\mathbf{X}$ and the corresponding data representations $\mathbf{A}$

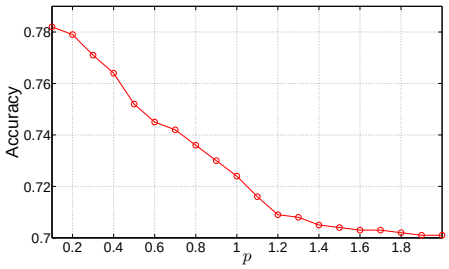

Figure 1. Clustering accuracy using the learned dictionary by the proposed method *vs*. p and q (where $p = q$) on the AT&T data set.

from an input data set by solving $J_{\text{RA-D}}$ in Eq. (9). For K-means clustering, we set K to be the true class numbers. Through our preliminary studies, we fine tune the parameter λ of $J_{\text{RA-D}}$ method in Eq. (9) by searching the grid of $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$, and report the best results. We vary the value of p and q, and report the K-means clustering accuracy using the learned representations $\mathbf{A}$. We repeat the experiment at each parameter setting for 50 times and report the average clustering accuracy in Figure 1.

Experimental results. From Figure 1 we can see that, smaller p and q lead to better clustering accuracy, *i.e*., the dictionary learned with smaller p and q has better data representation capability, which clearly confirms the correctness to use $\ell_{2,0^+}$-norm loss function and regularization in dictionary learning. Upon the results in Figure 1, we set $p = q = 0.1$ in all our following experiments.

3.2. Improved Data Representation Capability of the Learned Adaptive Dictionary

Now we further evaluate the data representation capability of the learned adaptive dictionaries by the proposed method on the three single-label image data sets: the AT&T data set, the USPS data set and the BinAlpha data set.

Experimental setups. We use the same experimental settings as in the previous subsection. We vary the size of the super-dictionary, denoted as $|\mathbf{D}|$, and examine the learned data adaptive dictionary size, denoted as $|\mathbf{D}_{\mathbf{X}}|$, and report the K-means clustering accuracy using the learned representations $\mathbf{A}$. We also report the K-means clustering accuracy on the learned representations by the efficient sparse coding (ESC) method [8], a baseline sparse learning method, in which the dictionary size is specified as that of $|\mathbf{D}|$. The ESC method also has a parameter β that acts same as λ in our method, we thus fine tune it in the same range as that for λ, and report the best performance. We choose the ESC method for comparison, because it is an unsupervised dictionary learning method by directly solving its optimization objective, similar to our $J_{\text{RA-D}}$ method but not robustifying the reconstruction errors. The experimental results are reported in Table 1, in which the accuracies of K-means clustering on the original data are also listed.

Table 1. K-means clustering accuracy using the sparse representations learned by the proposed $J_{\text{RA-D}}$ method and ECS method on the three benchmark data sets.

	$\lvert\mathbf{D}\rvert$	$\lvert\mathbf{D}_{\mathbf{X}}\rvert$	Clustering Accuracy		
			Our method	ESC	K-means
AT&T (40 classes)	400	77	**0.782**	0.776	0.691
	300	76	**0.771**	0.737	–
	100	73	**0.768**	0.705	–
	50	47	**0.715**	0.658	–
USPS (10 classes)	1000	38	**0.656**	0.631	0.605
	500	37	**0.651**	0.622	–
	100	33	**0.643**	0.591	–
	50	30	**0.630**	0.526	–
BinAlpha (26 classes)	500	67	**0.496**	0.477	0.421
	200	63	**0.488**	0.442	–
	100	61	**0.472**	0.401	–
	50	44	**0.431**	0.367	–

Experimental results. A first glance at the results in Table 1 show that, the clustering accuracies using the sparse representations learned by our method is consistently better than those by ESC method, which validate the effectiveness of the proposed method in terms of data representation.

Upon a more careful examination on the results, we can see that, although the sizes of the pre-specified super-dictionary $\mathbf{D}$ vary in a very large range, the sizes of the learned data adaptive dictionaries $\mathbf{D}_{\mathbf{X}}$ remain considerably stable. From this observation, we can draw a number of interesting conclusions in the following.

First, when the size of the super-dictionary varies in a rather big range, the clustering performance on the learned representations of the data by our method does not fluctuate too much. This confirms that the data representation power of the learned dictionary is not heavily dependent on the pre-specified size of the super-dictionary. In other words, our method is able to automatically determine the optimal compact dictionary bases.

Second, although the sample size of a real world data set is large, the number of its underlying patterns may be small, as revealed by $|\mathbf{D}_{\mathbf{X}}|$. This is consistent with the basic statistical assumption and provides another evidence to support the correctness of our model.

Third, our $J_{\text{RA-D}}$ method is able to capture the essential patterns of an input data set. As can be seen, the sizes of the learned data adaptive dictionaries $|\mathbf{D}_{\mathbf{X}}|$ are comparable to the ground truth class numbers of all the three data sets.

Fourth, because our method automatically picks up the most representative basis vectors from the learned super-dictionary $\mathbf{D}$ and uses them to represent input data, as long as the pre-specified super-dictionary size $|\mathbf{D}|$ is not very small, clustering on the learned representations using the data adaptive dictionary $\mathbf{D}_{\mathbf{X}}$ by our method can always achieve satisfactory accuracy. However, the clustering accuracy on the data representations learned by the ESC method degrades very quickly when the dictionary

size decreases. This is because the representation power of its learned dictionary generally depends on the number of available basis vectors, which is also the reason why dictionaries are usually designed to be over-complete in traditional sparse learning.

Finally, when the pre-specified size of the super-dictionary is too small, *e.g.*, 50 in either the AT&T data set or the BinAlpha data set, the representation capability of our method is also degraded. This again confirms that real world data has certain inherent patterns and in sparse learning the dictionary size should be greater than this inherent pattern number.

In summary, our method has demonstrated superior data representation capability through data adaptation, which is generally satisfactory in a variety of data conditions. Empirically, when $|\mathbf{D}| \geq 2K$, the subsequent classification accuracy is generally satisfactory. Thus, in all our following experiments, we set $|\mathbf{D}| = \min\{1000, n\}$.

3.3. Improved Classification Performance

Because making use of supervision information via enhanced data adaptation is an important advantage of the proposed method, we evaluate it in classification tasks.

Experimental setups. We compare our methods against the following most recent dictionary learning methods. For unsupervised dictionary learning methods, we compare to the two baseline methods including the K-SVD [1] method and the efficient sparse coding (ESC) method [8]. For supervised dictionary methods, we compare to the discriminative K-SVD (D-K-SVD) method [9], the supervised dictionary learning (SDC) [10] method, and the group sparse coding (GSC) [3] method. We implement these methods following the details in their original papers and reference the algorithms published by the authors. The parameters are fined tuned according to their original papers. Once the sparse representations of the input data are learned by these methods, support vector machine (SVM) is used for classification. We use the LIBSVM[3] package. Gaussian kernel is employed, *i.e.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$, where the parameters γ and C are fine tuned by searching the grid of $\left\{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\right\}$.

In order to evaluate the different components of the proposed method, we implement two versions of our method, *i.e.*, unsupervised RA-D method in Eq. (9) and semi-supervised SSR-D method in Eq. (12). The parameter λ are again fine tuned in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\right\}$. Note that, our methods naturally provides classification rules via reconstruction errors as introduced in Section 2.6.

We conduct standard five-fold cross-validation on each data set by compared methods, and report the average classification accuracy of the five single-label data sets in the top half of Table 2 and the average performances of the multi-label TRECVID data set in the top half of Table 3.

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 2. Comparison of the classification accuracies of the compared methods on the five data sets. The results in the top part are for classifications on the original data, while those in the bottom part are for classification on noisy data (20% training samples are incorrectly labeled to emulate noise).

Method	AT&T	USPS	BinAlpha	Reuters	TDT2
K-SVD	0.748	0.636	0.481	0.701	0.715
ESC	0.779	0.658	0.491	0.746	0.726
D-K-SVD	0.773	0.679	0.502	0.751	0.764
SDL	0.784	0.706	0.511	0.880	0.861
GSC	0.797	0.715	0.505	0.878	0.864
RA-D	0.807	0.720	0.511	0.879	0.868
SSR-D	0.815	0.734	0.519	0.883	0.879
K-SVD	0.706	0.587	0.412	0.635	0.671
ESC	0.716	0.603	0.442	0.703	0.684
D-K-SVD	0.731	0.642	0.453	0.715	0.718
SDL	0.740	0.691	0.477	0.824	0.807
GSC	0.743	0.694	0.476	0.837	0.813
RA-D	0.793	0.707	0.503	0.861	0.854
SSR-D	0.803	0.715	0.507	0.861	0.856

In order to verify the robustness of our methods, for each of the 5 trials we randomly select 20% training samples and assign them with incorrect labels to emulate outliers. The classification results on the noisy data are reported in the bottom halves of Table 2 and Table 3.

Experimental results. From the results in Table 2 and Table 3, our methods generally outperform other compared methods, sometimes by a significant margin, providing concrete evidence of the effectiveness of our methods in classification. Moreover, our SSR-D method is consistently better than its degenerated version of RA-D method, which is consistent with their mathematical formulations, *i.e.* the former is unsupervised without incorporating label information, while the latter is supervised and exploits the prior training knowledge. This also confirms that our enhanced data adaptation can improve the classification performance by taking advantage of supervision information. Finally, the classification performance of all the compared methods on the noisy data are decreased compared to those on the original clean data. However, the performance degradations of our methods are rather small, or even nonexistent, firmly supporting the usefulness of using the $\ell_{2,0^+}$-norm as the loss function as in our optimization objectives. That is, our methods are robust against noisy and outlier samples.

4. Conclusions

In this paper, we presented a novel dictionary learning method to address two important seldom studied issues in conventional sparse learning, *i.e.*, dictionary robustness and data adaptation. Different from existing dictionary learning

Table 3. Classification performance (definitions of the performance metrics can be found in [17]) comparison on TRECVID 2005 data. Top: on original clean data, bottom: on data with 20% noise added to training samples.

Methods	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
K-SVD	0.183 ± 0.020	0.346 ± 0.034	1.034 ± 0.075	0.189 ± 0.016	0.472 ± 0.023
ESC	0.180 ± 0.018	0.349 ± 0.029	1.064 ± 0.084	0.181 ± 0.014	0.479 ± 0.026
D-K-SVD	0.146 ± 0.012	0.307 ± 0.024	0.942 ± 0.064	0.167 ± 0.013	0.501 ± 0.031
SDL	0.139 ± 0.011	0.308 ± 0.022	0.951 ± 0.058	0.162 ± 0.011	0.504 ± 0.029
GSC	0.145 ± 0.010	0.301 ± 0.019	0.966 ± 0.035	0.160 ± 0.017	0.509 ± 0.016
RA-D	0.137 ± 0.011	0.305 ± 0.018	0.971 ± 0.036	0.167 ± 0.016	0.511 ± 0.015
SSR-D	$\mathbf{0.119 \pm 0.009}$	$\mathbf{0.275 \pm 0.018}$	$\mathbf{0.843 \pm 0.013}$	$\mathbf{0.141 \pm 0.010}$	$\mathbf{0.548 \pm 0.032}$
K-SVD	0.201 ± 0.021	0.379 ± 0.031	1.312 ± 0.075	0.206 ± 0.019	0.437 ± 0.021
ESC	0.196 ± 0.019	0.368 ± 0.023	1.217 ± 0.091	0.206 ± 0.017	0.441 ± 0.022
D-K-SVD	0.181 ± 0.014	0.349 ± 0.028	1.173 ± 0.044	0.198 ± 0.017	0.462 ± 0.033
SDL	0.164 ± 0.012	0.334 ± 0.023	1.096 ± 0.061	0.187 ± 0.014	0.474 ± 0.022
GSC	0.168 ± 0.011	0.335 ± 0.020	1.048 ± 0.032	0.191 ± 0.018	0.481 ± 0.019
RA-D	0.148 ± 0.012	0.316 ± 0.017	1.001 ± 0.035	0.171 ± 0.015	0.503 ± 0.012
SSR-D	$\mathbf{0.123 \pm 0.008}$	$\mathbf{0.281 \pm 0.017}$	$\mathbf{0.851 \pm 0.013}$	$\mathbf{0.169 \pm 0.012}$	$\mathbf{0.537 \pm 0.030}$

methods that use squared ℓ_2 loss function, we employed a new $\ell_{2,0^+}$-norm loss function to measure the reconstruction errors in our objectives, such that outlier samples have less importance and our objectives are more robust. In addition, instead of using additional terms to incorporate supervision information, we exploited such information by data adaptation via structural sparse regularization. This method does not incur extra parameters, such that our learning model is more stable and easier to fine tune. Due to the data adaptation nature, the dictionaries learned by our methods are adaptive to not only the input data but also their class labels, which improves the discriminativity of the learned dictionaries and makes them more suitable for classification tasks. Because we use $\ell_{2,0^+}$ regularization to adaptively select prominent basis vectors from a super-dictionary, the optimal dictionary size is automatically learned from the input data. An efficient algorithm to solve the objective was described, together with the rigorous proof of its convergence. We have evaluated several important aspects of the proposed methods. Significantly improved experimental results in extensive empirical studies demonstrated the usefulness of the proposed methods.

References

[1] 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.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *NIPS*, pages 41–48, 2007.

[3] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *NIPS*, 2009.

[4] X. Cai, F. Nie, H. Huang, and C. Ding. Multi-class l2,1-norm support vector machine. In *ICDM*, pages 91–100, 2011.

[5] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[6] E. Candès and M. WAKIN. An introduction to compressive sensing. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[7] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[8] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, pages 1–8, 2008.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2009.

[11] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization. In *NIPS*, 2010.

[12] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Technical report, Department of Statistics, University of California, Berkeley*, 2006.

[13] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, pages 759–766, 2007.

[14] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[15] H. Wang, F. Nie, and H. Huang. Robust and discriminative self-taught learning. In *ICML*, pages 298–306, 2013.

[16] H. Wang, F. Nie, H. Huang, S. L. Risacher, C. Ding, A. J. Saykin, L. Shen, and ADNI. A new sparse multi-task regression and feature selection method to identify brain imaging predictors for memory performance. *IEEE Conference on Computer Vision (ICCV)*, pages 557–562, 2011.

[17] M. Zhang and Z. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[18] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, pages 2691–2698, 2010.