Robust Dictionary Learning with Capped ℓ_1-Norm

Wenhao Jiang, Feiping Nie, Heng Huang*
University of Texas at Arlington
Arlington, Texas 76019, USA
cswhjiang@gmail.com, feipingnie@gmail.com, heng@uta.edu

Abstract

Expressing data vectors as sparse linear combinations of basis elements (dictionary) is widely used in machine learning, signal processing, and statistics. It has been found that dictionaries learned from data are more effective than off-the-shelf ones. Dictionary learning has become an important tool for computer vision. Traditional dictionary learning methods use quadratic loss function which is known sensitive to outliers. Hence they could not learn the good dictionaries when outliers exist. In this paper, aiming at learning dictionaries resistant to outliers, we proposed capped ℓ_1-norm based dictionary learning and an efficient iterative re-weighted algorithm to solve the problem. We provided theoretical analysis and carried out extensive experiments on real word datasets and synthetic datasets to show the effectiveness of our method.

1 Introduction

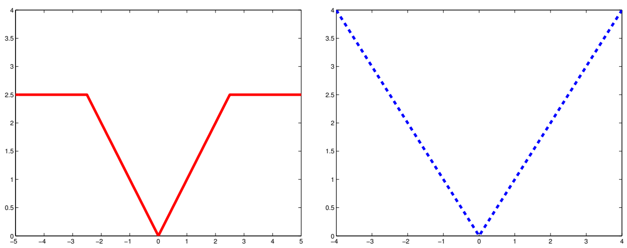
Dictionary learning [Tosic and Frossard, 2011] and sparse representation [Olshausen and Field, 1997] are important tools for computer vision. Dictionary learning seeks to learn an adaptive set of basis elements (dictionary) from data instead of predefined ones [Mallat, 1999], so that each data sample is represented by *sparse* linear combination of these basis vectors. It has achieved start-of-the-art performance for numerous image processing tasks such as classification [Raina *et al.*, 2007; Mairal *et al.*, 2009b], denoising [Elad and Aharon, 2006], self-taught learning [Wang *et al.*, 2013b], and audio processing [Grosse *et al.*, 2007]. Unlike principal component analysis, dictionary learning does not impose that the dictionary be orthogonal, hence allow more flexibility to represent data.

Usually, dictionary learning is formulated as:

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2, \quad s.t. \quad \|\mathbf{A}\|_0 \le \gamma\,. \tag{1}$$

In this formulation, $\mathbf{X} \in \mathbb{R}^{d\times n}$ is the data matrix whose columns represent samples, $\mathbf{A} \in \mathbb{R}^{K\times n}$ is the new representations of data and the matrix $\mathbf{D} \in \mathbb{R}^{d\times K}$ contains K basis vectors to learn. To prevent the ℓ_2 norm of $\mathbf{D}$ being arbitrarily large which would lead to arbitrarily small values in the columns of $\mathbf{A}$, it is common to constraint the columns have ℓ_2 norm less than or equal to 1. Hence, $\mathbf{D}$ is in the following convex set of matrices:

$$\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{d\times K} \mid \forall i \in \{1,\cdots,K\}, \|\mathbf{d}_i\|_2 \le 1\}\,. \tag{2}$$



(a) Capped ℓ_1-norm loss ($\varepsilon = 2.5$) (b) ℓ_1-norm loss

Figure 1: Capped ℓ_1-norm loss vs ℓ_1-norm loss

The dictionary learned is the foundation for sparse representations. Dictionary with good expressive ability is the key to achieve good performance. It is well known that the quadratic loss function is not robust to outliers. To train the dictionary, a large amount data will be used, and outliers will be included in the training data unavoidably. Hence it is necessary to learn dictionary resistant to outliers.

To be robust to outliers, the quadratic loss function could be replaced by an loss function that are not sensitive to outliers, e.g. ℓ_1 norm loss or Huber loss. In this paper, we use capped ℓ_1-norm based loss function $l_\varepsilon(u) = \min(|u|, \varepsilon)$, where ε is a parameter. It is illustrated in Fig. 1. This loss function treat u equally if $|u|$ is bigger than ε. Hence, it is more robust to outliers than ℓ_1 norm. Unfortunately, this loss function is not convex. In this paper, we propose an efficient algorithm to find the local optimal solutions. And we will also provide theoretical analysis of this method.

2 Related Work

In this section, we present a brief review on dictionary learning methods as follows.

Since the original dictionary learning model (1) is NP-hard, the straightforward way to find dictionary is to adopt greedy

*Corresponding Author. This work was partially supported by NSF IIS-1117965, IIS-1302675, IIS-1344152, DBI-1356628.

strategy. K-SVD [Aharon *et al.*, 2006] tried to solve the following model:

$$\begin{aligned}&\min_{\mathbf{D},\mathbf{A}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_2^2\\ &s.t.\ \mathbf{d}_j = 1,\ \text{for } j = 1,2,\cdots,K\\ &\qquad \|\mathbf{a}_i\|_0 \le T_0,\ \text{for } i = 1,2,\cdots,n\,,\end{aligned} \tag{3}$$

where $\mathbf{d}_i$ is the ith column of matrix $\mathbf{D}$. The K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms. The columns of dictionary are updated with SVD sequentially, and the corresponding coefficients are updated with any pursuit algorithm. The K-SVD algorithm showed good performance on image denoising. As extensions of K-SVD, LC-KSVD [Jiang *et al.*, 2013] and discriminative K-SVD [Zhang and Li, 2010] introduced label information into the procedure of learning dictionaries. Hence the dictionaries learned are more discriminative.

Except for adopting greedy strategy to solve problems with l_0 constraints, the original problem can be relaxed into the following traditional dictionary learning problem with ℓ_1 regularization:

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1\,. \tag{4}$$

It is convex separately with respect to $\mathbf{D}$ and $\mathbf{A}$, a ℓ_1 regularized least squares problem and a ℓ_2 constrained least squares problem are needed to be solved iteratively to find the solutions. In [Lee *et al.*, 2007], an feature-sign search algorithm for learning coefficients, and a Lagrange dual method for learning the bases are proposed to compute the dictionary and corresponding representations. An online algorithm was proposed to solve it efficiently in [Mairal *et al.*, 2009a]. The method mentioned above all use quadratic loss functions to measure the reconstruction errors, hence these methods are not robust enough to outliers.

In order to learn dictionary robustly, Lu *et al.* proposed online robust dictionary learning (ORDL) [Lu *et al.*, 2013], which uses ℓ_1 loss function and ℓ_1 regularization term that could be expressed as

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_1^1 + \lambda\|\mathbf{a}\|_1. \tag{5}$$

It updates dictionary and representations alternately in an online way. For visual tracking, Wang proposed online robust non-negative dictionary learning [Wang *et al.*, 2013c], which uses the Huber loss function. In [Wang *et al.*, 2013a], the semi-supervised robust dictionary learning model was proposed by solving the ℓ_p-norm based objective.

3 Proposed New Algorithm

To facilitate the presentation of our method, we describe the notations in the following subsection.

3.1 Notations

Throughout this paper, the following definitions and notations are used. We use bold upper letters for matrices, bold lower letters for vectors and regular lower letters for elements. For vector $\mathbf{a} = (a_1, a_2, \cdots, a_m)^T \in \mathbb{R}^m$, let $\|\mathbf{a}\|_1 = \sum_{i=1}^m |a_i|$ be the ℓ_1-norm of $\mathbf{a}$. Similarly, we define $\|\mathbf{A}\|_0$ as the number of nonzero elements in matrix $\mathbf{A}$ and $\|\mathbf{A}\|_1 = \sum \|\mathbf{a}_i\|_1$ be the ℓ_1 norm of matrix $\mathbf{A}$.

3.2 Robust dictionary learning

Capped ℓ_1-norm has been used in [Zhang, 2010; 2013] as a better approximation of ℓ_0-norm regularization. An extension capped-$\ell_{1,1}$ regularization was proposed in the multi-task feature learning setting [Gong *et al.*, 2013]. In this paper, we use capped ℓ_1-norm as a robust loss function.

For dictionary learning, our goal is to find a set of high quality atoms. An straightforward way is to use ℓ_1 loss function. To provide better robustness, we go further to use capped ℓ_1-norm loss function. As illustrated in Fig 1, the objective values of capped ℓ_1-norm loss does not increase any more if $|u|$ is larger than ε. Therefore, capped ℓ_1-norm loss is more robust than ℓ_1-norm loss. We use the same constraints for dictionary and the objective function of our roust dictionary learning with capped ℓ_1-norm is expressed as:

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1} \sum_{i=1}^{n} \min(\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2, \varepsilon) + \lambda\|\mathbf{A}\|_1\,. \tag{6}$$

From this objective function, we can see that if ε is set as ∞, the above objective function becomes $\ell_{2,1}$-norm with the same constraints.

We define $f(\mathbf{A}) = \|\mathbf{A}\|_1$, which is a convex function. And define a concave function $\bar{L}(u) = \min(\sqrt{u}, \varepsilon)$ where $u > 0$ and $g(\mathbf{D}, \mathbf{a}) = \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2$. We have:

$$\bar{L}'(u) = \begin{cases} \frac{1}{2\sqrt{u}}, & \text{if } 0 < u < \varepsilon^2 \\ 0, & \text{if } u > \varepsilon^2 \end{cases} \tag{7}$$

Our objective function is sum of a convex function and a concave function. According to the re-weighted method proposed in [Nie *et al.*, 2010; 2014], it can be solved by updating $\mathbf{D}, \mathbf{A}$ and auxiliary variables s_i with following updating rules:

$$[\mathbf{D},\mathbf{A}] = \arg\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1} \sum_i s_i\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\|\mathbf{A}\|_1\,, \tag{8}$$

$$s_i = \begin{cases} \frac{1}{2\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2}, & \text{if } \|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2 \le \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The whole iterative re-weighted method is listed in Alg. 1.

The subproblem (8) is similar to the traditional dictionary learning. The only difference lies in the fact that samples are not treated equally. It is weighted dictionary learning. From the updating rule for s_i, we can see that the samples with lower reconstruction errors have higher weights, which is very intuitive for learning dictionary robustly.

The subproblem (8) is convex separately with respect to $\mathbf{D}$ and $\mathbf{A}$. We adopt the same strategy as solving traditional dictionary learning. We solve it by updating $\mathbf{D}$ and $\mathbf{A}$ alternately. With $\mathbf{A}$ fixed, the objective function becomes:

$$\min_{\mathbf{D},\|\mathbf{d}_i\|\le 1} \sum_i s_i\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2\,, \tag{10}$$

which is equivalent to:

$$\min_{\mathbf{D},\|\mathbf{d}_i\|\leq 1} \sum_i \|\mathbf{z}_i - \mathbf{D}\mathbf{c}_i\|_2^2 \,, \tag{11}$$

where $\mathbf{z}_i = \sqrt{s_i}\mathbf{x}_i$ and $\mathbf{c}_i = \sqrt{s_i}\mathbf{a}_i$. To solve the above problem, we update $\mathbf{D}$ column by column which is similar to the traditional dictionary learning. The algorithm we use is adopted from [Mairal *et al.*, 2009a], which is described in Alg. 3.

With $\mathbf{D}$ fixed, the objective of (8) becomes:

$$\mathbf{A} = \operatorname*{argmin}_{\mathbf{A}} \sum_i s_i\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\|\mathbf{A}\|_1 \,, \tag{12}$$

which can be decomposed into n independent problems as follows:

$$\mathbf{a}_i = \operatorname*{argmin}_{\mathbf{a}} s_i\|\mathbf{x}_i - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \,. \tag{13}$$

If $s_i \neq 0$, this problem can be transformed into the following LASSO problem:

$$\mathbf{a}_i = \operatorname*{argmin}_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}\|_2^2 + \frac{\lambda}{s_i}\|\mathbf{a}\|_1 \,, \tag{14}$$

which could be solved efficiently. The algorithm to solve weighted dictionary learning is summarized in Alg. 2.

During the stage of training dictionary $\mathbf{D}$, if s_i is 0, the corresponding representation $\mathbf{a}_i$ will be a zero vector. Hence, when training $\mathbf{D}$, our method does not find codings for outliers. However, the outliers in training $\mathbf{D}$ stage might not be outliers for classification tasks. Hence, we let the classification algorithm to decide how to use the codings of the outliers. With the learned $\mathbf{D}$, in order to get the codings for outliers, we just run Alg. 1 with setting ε as ∞ and omit the $\mathbf{D}$ updating step (step 1 in Alg. 2) to learn the representations for both training data and testing data with the same λ. In fact, we are solving a dictionary learning problem with $\ell_{2,1}$-norm loss function.

Algorithm 1 Robust dictionary learning with capped ℓ_1-norm

input: Data matrix $\mathbf{X}$, dictionary size K, λ and ε.

Initialize $\mathbf{D}$, $\mathbf{A}$ and $s_i = 1$ for $i = 1, 2, \cdots, n$.
repeat
1. Solve subproblem (8) with algorithm 2
2. Update s_i for $i = 1, \cdots, n$ as (9)

until Converge

output: $\mathbf{D}$ and $\mathbf{A}$

3.3 Convergence analysis

Theorem 3.1. *The algorithm described in Alg. 1 will decrease the objective value of* (6) *in each iteration until it converges.*

Proof. We define $h(w) = w^2$, and denote $u = \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2$ and $\bar{L}(u) = \min(\sqrt{u}, \varepsilon)$. The capped ℓ_1-norm loss function $\min(\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2, \varepsilon)$ could be re-written as:

$$\begin{aligned} &\min(\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2, \varepsilon) && (15)\\ =& \inf_{s\geq 0} \left[sh(\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2) - L^*(s) \right] \,, && (16) \end{aligned}$$

Algorithm 2 Weighted dictionary learning

input: Data matrix $\mathbf{X}$, initial dictionary matrix $\mathbf{D}_0$ and representation matrix $\mathbf{A}_0$, λ and weight vector $\mathbf{s}$.

Initialize $\mathbf{D} = \mathbf{D}_0$ and $\mathbf{A} = \mathbf{A}_0$
repeat
1. Update $\mathbf{D}$ with algorithm 3.
2. Update $\mathbf{a_i}$ for $i = 1, \cdots, n$ as

$$\mathbf{a}_i = \begin{cases} \operatorname{argmin}_{\mathbf{a}} \|\mathbf{x}_i - \mathbf{D}\mathbf{a}\|_2^2 + \frac{\lambda}{s_i}\|\mathbf{a}\|_1 & \text{if } s_i > 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$$

until Converge

output: $\mathbf{D}, \mathbf{A}$

Algorithm 3 Dictionary update

input: Data matrix $\mathbf{X}$, dictionary matrix $\mathbf{D}$, representation matrix $\mathbf{A}$ and weight vector $\mathbf{s}$.

Compute matrix $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_K] \in \mathbb{R}^{K\times K}$ and $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K] \in \mathbb{R}^{d\times K}$ as

$$\mathbf{G} = \sum_{i=1}^n s_i\mathbf{a}_i\mathbf{a}_i^T, \quad \mathbf{H} = \sum_{i=1}^n s_i\mathbf{x}_i\mathbf{a}_i^T.$$

repeat
for $j = 1$ **to** K **do**
Update the j-th column of $\mathbf{D}$:

$$\mathbf{u}_j = \frac{1}{\mathbf{G}_{jj}}(\mathbf{h}_j - \mathbf{D}\mathbf{g}_j) + \mathbf{d}_j$$

$$\mathbf{d}_j = \frac{1}{\max(\|\mathbf{u}_j\|_2, 1)}\mathbf{u}_j$$

end for
until Converge

output: $\mathbf{D}$

where $L^*(s)$ is the concave dual of $\bar{L}(u)$ defined as:

$$L^*(s) = \inf_u \left[su - \bar{L}(u) \right] \,. \tag{17}$$

Plug in $\bar{L}(u)$ and it is easy to find that:

$$L^*(s) = \begin{cases} -\frac{1}{4s}, & \text{if } \sqrt{u} < \varepsilon \\ s\varepsilon^2 - \varepsilon, & \text{if } \sqrt{u} \geq \varepsilon \end{cases} . \tag{18}$$

Therefore, the capped ℓ_1-norm loss could be expressed as:

$$\min(\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2, \varepsilon) = \inf_{s\geq 0} L_{\mathbf{x}}(s, \mathbf{D}, \mathbf{a}) \,, \tag{19}$$

where

$$\begin{aligned} &L_{\mathbf{x}}(s, \mathbf{D}, \mathbf{a}) \\ =& \begin{cases} s\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \frac{1}{4s}, & \text{if } \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2 < \varepsilon \\ s\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 - s\varepsilon^2 + \varepsilon, & \text{if } \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2 \geq \varepsilon \end{cases} . \end{aligned} \tag{20}$$

Hence, the objective function (6) is equivalent to the following objective:

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\leq 1} \sum_{i=1}^n \inf_{s_i\geq 0} L_{\mathbf{x}_i}(s_i, \mathbf{D}, \mathbf{a}_i) + \lambda\|\mathbf{a}_i\|_1 \,, \tag{21}$$

which can be re-written as:

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1, s_i\ge 0} o(\mathbf{D},\mathbf{A},\mathbf{s})\,, \tag{22}$$

where we denote:

$$o(\mathbf{D},\mathbf{A},\mathbf{s}) = \sum_{i=1}^{n} L_{\mathbf{x}_i}(s_i,\mathbf{D},\mathbf{a}_i) + \lambda\|\mathbf{a}_i\|_1\,. \tag{23}$$

Therefore, the algorithm described in Alg. 1 can be seen as a two stage optimization method:

Updating $\mathbf{D}$ and $\mathbf{A}$ stage: $\mathbf{D}$ and $\mathbf{A}$ are updated by solving:

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1} o(\mathbf{D},\mathbf{A},\mathbf{s})\,, \tag{24}$$

which is equivalent to:

$$\min_{\mathbf{D},\mathbf{A},\|\mathbf{d}_i\|\le 1} \sum_{i=1}^{n} s_i\|\mathbf{x}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda\|\mathbf{a}_i\|_1\,. \tag{25}$$

The updated $\mathbf{D}$ and $\mathbf{A}$ decrease the objective values of (25). Hence, they also decrease the objective values of (24).

Updating s_i stage: The auxiliary variable $\mathbf{s}$ are updated by solving:

$$\min_{s_i\ge 0} o(\mathbf{D},\mathbf{A},\mathbf{s})\,, \tag{26}$$

which is equivalent to n independent problems

$$\min_{s_i\ge 0} L_{\mathbf{x}_i}(s_i,\mathbf{D},\mathbf{a}_i)\,. \tag{27}$$

Both stages will decrease the value of the objective function (21), hence our algorithm also decrease the value of the original objective function (6) and is guaranteed to converge. $\square$

4 Experimental Results

In this section, we analyze and illustrate the performance of our method on real word datasets.

4.1 Preliminary study on natural image patches

First, we present a preliminary analysis of our method on natural image patches. In our method, λ mainly controls the sparsity of representations and ε mainly controls the number of outliers identified by the algorithm. ε is related to the residuals of representations. If the residual of a sample is larger than ε, it is not used to train the dictionary, since the corresponding s is zero. We set ε as infinity first to get the distribution of the residuals. We use this distribution as an estimation of the distribution of residuals with optimal $\mathbf{D}$ and $\mathbf{A}$. We set ε such that a fraction of samples are seen as outliers. In this experiment, we set the fraction of outliers as 0.05 and $\lambda = 0.1$ empirically. The values of objective function during iterations are presented in Fig. 2. We can see that our method converged in only 22 iterations. The bases learned on natural image patches by our method is shown in Fig. 3.

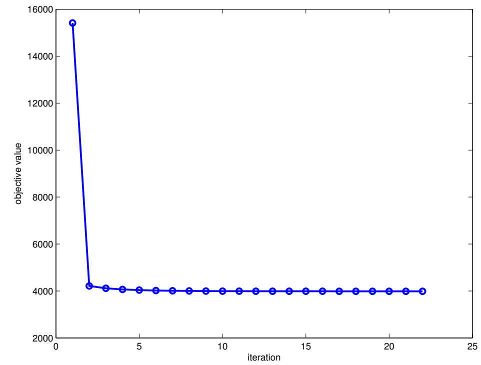


Figure 2: The values of objective function during iterations when learning dictionaries on 10,000 natural images patches.

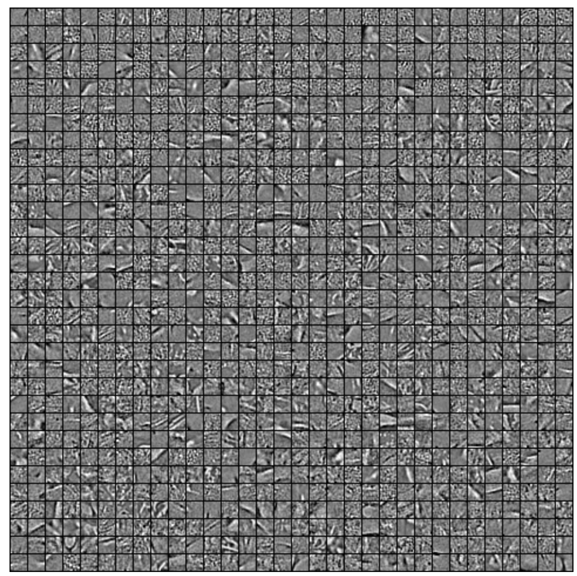


Figure 3: Learned 1,024 bases images bases (each 14×14 pixels) on 10,000 natural images patches.

4.2 Face recognition without occlusion

In this subsection, we provide the experimental results on face recognition tasks without synthetic occlusion on extended Yale B dataset [Georghiades *et al.*, 2001] and AR face dataset [Martinez, 1998].

Extended Yale B dataset The extended Yale B database [Georghiades *et al.*, 2001] contains 2,414 images of 38 human frontal faces under 64 illumination conditions and expressions. There are about 64 images for each person. The original images were cropped to 192×168 pixels. Some samples are shown in Fig. 4(a). We do not perform any preprocess on the images. Following [Wright *et al.*, 2009], we project each face image into a 504-dimensional feature vector using a random matrix. We split the database randomly into two halves. One half which contains about 32 images for each person was used for training the dictionary. The other half was used for testing.

(a) Sample images from extended Yale B dataset

(b) Sample images from AR face dataset

Figure 4: Sample images

In the practical implementation of our method, we first set ε as infinity and choose the best λ with cross validation. Then we fix λ and choose ε.

For face recognition, we compared our method with a few start-of-the-art methods, including traditional dictionary learning [Mairal *et al.*, 2009a], K-SVD [Aharon *et al.*, 2006], LC-KSVD1 and LC-KSVD2 [Jiang *et al.*, 2013], D-KSVD [Zhang and Li, 2010] and ORDL [Lu *et al.*, 2013]. The dictionary size is 570 for all methods, which means 15 items for person on average. The parameters for these methods were selected by cross validation. We ran all methods on 10 different splits of training and testing set. The results are summarized in Table 1. We can see that our method achieved the best performance. The reason that ORDL did not perform well might be the dictionary and representations trained with ℓ_1-norm loss function are not suitable for classification tasks.

We show that our method is robust when training dictionary. Recall that, in our method $s_i = 0$ means that the ith sample is recognized as outlier and not used in training the dictionary. In running on one random split, our method found 20 outliers, which are all shown in Fig. 5. We can see that most of the outliers found by our method are with extreme illumination, which will effect the quality of bases.

AR face dataset The AR face dataset [Martinez, 1998] consists of over 4,000 color images from 126 individuals. For each individual, 26 pictures were taken in two separate sessions. Compared to the extended Yale B dataset, as shown

Table 1: Average classification accuracies(%) on extended Yale B dataset.

Method	Accuracy (%)
Capped Norm	**96.91**
Traditional DL	95.70
KSVD	95.54
LC-KSVD1	93.61
LC-KSVD2	94.48
D-KSVD	93.58
ORDL	89.17

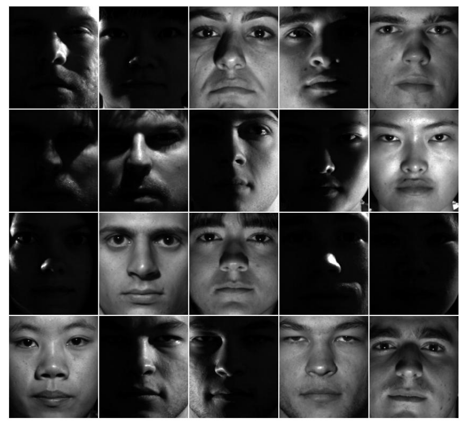


Figure 5: The outliers found by our method when training dictionary on an randomly selected training set from extended Yale B dataset.

in Fig. 4(b) [1], the AR face dataset includes more facial variations, including different illumination conditions, different expressions, and different facial disguises (sunglasses and scarves). Following the standard evaluation procedure from [Wright *et al.*, 2009], we used a subset of the database consisting of 2,600 images from 50 male subjects and 50 female subjects. For each person, 20 images were randomly selected for training and the remaining images are for testing. Each face image is cropped to 165×120 and then projected into a 540-dimensional feature vector.

Similar to the experiments on Yale B dataset, we also ran all methods in 10 different splits of training and testing set. The results are presented in Table 2. We can see that our method achieved the best performance among all dictionary learning algorithms. We show the 17 outliers found by our algorithm in Fig 6. We can see that these outliers are faces with glasses or scarves. Our algorithm identified these faces as outliers that will hurt the quality of dictionary and did not use them in the training process.

4.3 Face recognition with occlusion

In order to study the property of robustness to outliers, we carried out experiments on extended Yale B dataset with syn-

[1]We use grayscale images in our experiments.

Table 2: Average classification accuracies(%) on AR face dataset.

Method	Accuracy (%)
Capped Norm	**97.48**
Traditional DL	97.25
KSVD	95.03
LC-KSVD1	94.56
LC-KSVD2	94.33
D-KSVD	88.18
ORDL	91.72

Figure 6: The 17 outliers found by our method when training dictionary on an randomly selected training set from AR face dataset.

thetic outliers. If the samples for training the dictionary contain outliers, traditional dictionary learning will try to find dictionary that could express these outliers. But it is of no use to express the outliers, hence the quality of dictionary will be effected for traditional dictionary learning. In this experiment, we generate outliers by adding block occlusion. Other ways of generating outliers will also lead to the similar phenomena described below.

We selected a fraction of images (outlier ratio) from training set, then we added block occlusion of size 96×96 into these images at random positions. Some samples are shown in Fig. 7. We set outlier ratio as 10%, 20%, 30% and 40% to get 4 different datasets with synthetic outliers in the training sets. The images from testing sets are not added any occlusions. Under this setting, the effect on the performance of classification will come from the quality of dictionaries. Similar to the experiments we have done above, we also ran experiments on 10 different datasets and the average accuracies are reported in Table 3. We can see that the classification accuracies of all methods decreased as the outlier ratio increased. Compared with performances without block occlusion in Table 1, the performance of traditional dictionary learning dropped a lot (from 95.70 to 90.00 with 10% outliers). But our method did not drop so much (from 96.91 to 96.29). And our method also performed the best on all datasets. Hence we can conclude that our capped ℓ_1-norm loss does provide resistance to such kind of noise.

The outliers found by our algorithm from an synthetic dataset with 10% corruption are shown in Fig. 8. We can see that most outliers are images that are too dark or corrupted and are difficult to recognize. In our method, the dictionaries were trained without these samples, hence they were better than dictionaries trained in other ways, which can be proved from the comparisons of performances in Table 3.

Figure 7: Sample faces with occlusion from extended Yale B dataset.

Table 3: Average classification accuracies(%) on extended Yale B dataset with block occlusion of different levels.

Method	10%	20%	30%	40%
Capped Norm	**96.29**	**95.49**	**95.43**	**93.88**
Traditional DL	90.00	88.37	87.61	86.17
KSVD	93.74	93.26	92.99	92.30
LC-KSVD1	94.10	93.47	93.14	92.42
LC-KSVD2	94.27	93.65	93.21	92.27
D-KSVD	91.19	90.14	89.85	89.21
ORDL	87.00	85.53	84.95	83.38

Figure 8: The 93 outliers found by our method when training dictionary on an randomly selected training set with 10% samples corrupted with block occlusion from extended Yale B dataset.

5 Conclusion

In this paper, we presented a robust dictionary learning model based on capped ℓ_1-norm and an efficient algorithm to find local solutions. One important advantage of our method is its robustness to outliers. By iteratively assigning weights to samples, our algorithm succeeded in finding outliers and reduced their effects on training the dictionaries. The proposed method was extensively evaluated on face recognition jobs on different real word datasets and synthetic datasets. The experimental results demonstrated that our method outperforms previous state-of-the-art dictionary learning methods.

References

[Aharon *et al.*, 2006] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

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

[Georghiades *et al.*, 2001] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[Gong *et al.*, 2013] Pinghua Gong, Jieping Ye, and Changshui Zhang. Multi-stage multi-task feature learning. *The Journal of Machine Learning Research*, 14(1):2979–3010, 2013.

[Grosse *et al.*, 2007] Roger Grosse, Rajat Raina, Helen Kwong, and Andrew Y. Ng. Shift-Invariant Sparse Coding for Audio Classification. In *Uncertainty in Artificial Intelligence*, 2007.

[Jiang *et al.*, 2013] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label consistent k-svd: learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[Lee *et al.*, 2007] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, Cambridge, MA, 2007.

[Lu *et al.*, 2013] Cewu Lu, Jiaping Shi, and Jiaya Jia. Online robust dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*, pages 415–422, 2013.

[Mairal *et al.*, 2009a] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th International Conference on Machine Learning*, pages 689–696, Montreal, June 2009. Omnipress.

[Mairal *et al.*, 2009b] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1033–1040. 2009.

[Mallat, 1999] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[Martinez, 1998] Aleix M Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

[Nie *et al.*, 2014] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1062–1070, 2014.

[Olshausen and Field, 1997] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[Raina *et al.*, 2007] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[Tosic and Frossard, 2011] I. Tosic and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, March 2011.

[Wang *et al.*, 2013a] Hua Wang, Feiping Nie, Weidong Cai, and Heng Huang. Semi-supervised robust dictionary learning via efficient $l_{2,0+}$-norms minimization. *International Conference on Computer Vision (ICCV 2013)*, pages 1145–1152, 2013.

[Wang *et al.*, 2013b] Hua Wang, Feiping Nie, and Heng Huang. Robust and discriminative self-taught learning. *The 30th International Conference on Machine Learning (ICML 2013), Journal of Machine Learning Research, W&CP*, 28(3):298–306, 2013.

[Wang *et al.*, 2013c] Naiyan Wang, Jingdong Wang, and Dit-Yan Yeung. Online robust non-negative dictionary learning for visual tracking. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 657–664, 2013.

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

[Zhang and Li, 2010] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*, pages 2691–2698. IEEE, 2010.

[Zhang, 2010] Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010.

[Zhang, 2013] Tong Zhang. Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293, 11 2013.