Supervised Bayesian Sparse Coding for Classification

Jinhua Xu, Li Ding and Shiliang Sun

***Abstract*—In this paper, we propose a supervised Bayesian sparse coding (SBSC) model for classification. The sparse coding with Laplacian scale mixture prior is formulated as a weighted l_1 minimization problem. Category-specific discriminative dictionaries and regularization parameters are learned using variational EM algorithm from the training samples of each category. Classification of a test sample is done using the MAP estimate of the sparse codes. We have tested the model on different recognition tasks and demonstrated the effectiveness of the model.**

I. Introduction

Sparse coding has been applied successfully in numerous classification tasks [1], [2], [3], [4], [5], [6], [7], [8], [9], [10].

There are two stages in classification application using sparse coding. In the first stage, sparse codes and dictionary(ies) are learned to represent the input signal. In the second stage, classification is done based on the reconstruction error or a classifier output. For unsupervised sparse coding and dictionary learning, the two stages are separated completely, and label information is not used in the first coding stage [1], [6]. However, the sparse codes and dictionaries learned via unsupervised learning are often lack of discrimination as they are optimal for reconstruction but not for classification. Recently, many algorithms have been proposed to enhance the discrimination of visual dictionaries through supervised learning, which can be divided into three categories.

The first class of approaches is supervised sparse coding, that is, the label information is used for sparse coding. In some previous work[5], [7], [11], [12], [13], [14], multiple category-specific dictionaries were learned to promote discrimination between classes. The simplest strategy consists of learning one dictionary for each class and estimating the class based on the reconstruction error. In [5], [11], the classical softmax discriminative cost function was combined with sparse reconstruction in the objective function and jointly optimized during dictionary learning. In [12], an incoherence promoting term was added to encourage dictionaries associated to different classes to be as independent as possible. In [2], [13], the Fisher discrimination criterion was imposed on the coding coefficients so that they have small within-class scatter but big between-class scatter. In [14], a joint dictionary learning algorithm was proposed to exploit the visual correlation within a group of visually similar object categories where a commonly shared dictionary and multiple category-specific dictionaries were accordingly modeled.

The second class of approaches combines the dictionary learning and classifier training into a single objective function, aiming at enhancing the discrimination of the learned dictionary by solving the unified optimization[4], [10], [15], [16]. The discrimination criteria include linear predictive classification error [15], [16] and logistic loss function with residual errors [4]. In [10], a label consistent constraint was introduced and combined with the reconstruction error and the linear predictive classification error to form a unified objective function.

The last type of approaches updates the dictionary by using backpropagation of the classification error [3], [9], [10], [17], [18], [19]. It was indicated that dictionaries learned via backpropagation yield better classification performances [3], [9], [10]. Bradley and Bagnell [3] introduced a differentiable KL prior as a smooth approximation of the sparse regularization and employed a backpropagation procedure to train the dictionary for sparse coding. In [9], supervised and semi-supervised dictionary learning was introduced to various tasks. It was shown that even for nonsmooth sparse regularization such as l_1, the resulting optimization problem is smooth under mild assumptions and can be solved efficiently using stochastic gradient descent. In [17], [18], discriminative dictionaries were learned through back propagation by minimizing the training error of the image level features, which are extracted by average pooling or max pooling over the sparse codes over larger neighborhoods within a spatial pyramid. In [19], a top-down saliency model was proposed that jointly learns a Conditional Random Field (CRF) and a discriminative dictionary. The dictionary was learned by minimizing the energy function.

Some recent research work suggested that image space is actually a smooth low dimensional sub-manifold embedded in a high dimensional ambient space. Standard sparse coding does not include locality constraints explicitly, and thus may be inaccurate in modeling the manifold[20]. Meanwhile, the over-completeness of the dictionary and the independent coding process may also result in the instability of sparse coding[21], that is, similar features may be encoded as totally different sparse codes. As suggested in [20], locality was more essential than sparsity, since locality can lead to sparsity while sparsity cannot cause locality. Therefore, some research has been done to address locality-preserving or similarity preserving during dictionary learning for image classification[8], [20], [21], [22], [23].

In this paper, we address the supervised sparse coding with locality preserving in a Bayesian framework. By assuming that observations from the same class have the same prior distribution, we build a Bayesian sparse coding model for each class independently, where Laplacian scales of coefficients are considered as random variables and sparse coding with Laplacian scale mixture prior is formulated. Discriminative dictionaries can then be learned with the regularization parameters using variational EM algorithms. Since each class has its own regularization parameters, similar observations

Jinhua Xu, Li Ding and Shiliang Sun are with the Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai (email: jhxu@cs.ecnu.edu.cn).

This work is supported by the National Natural Science Foundation of China under Project 61175116, and Shanghai Knowledge Service Platform for Trustworthy Internet of Things (No. ZF1213).

from the same class will be encoded in similar sparse codes, and the instability problem is alleviated. Different from the existing approaches where a discriminative term based on classification cost or Fisher discrimination criterion was added in the objective function for sparse coding, we make the learned dictionaries discriminative by Bayesian modeling of the coefficients for each class. To the best of our knowledge, this is the first work for discriminative dictionary learning and stable sparse coding through sparsity regularization design.

The paper is organized as follows: Section 2 reviews some related work. Section 3 describes the proposed supervised Bayesian sparse coding method. Section 4 presents experimental results on some well-known image databases. Finally, discussions and conclusions are drawn in Section 5.

II. Related work

In this section, we first review the standard sparse coding and dictionary learning model with Laplacian prior, then introduce the related work on sparsity regularization, dictionary learning and locality preserving sparse coding repspectively.

A. Standard Sparse coding with Laplacian prior

The graphical model of sparse coding and dictionary learning with Laplacian prior is depicted in Figure 1.

$$\mathbf{x}_i = \Phi S_i + \nu, \tag{1}$$

where $\mathbf{x}_i \in R^d (i = 1, \ldots, N)$ are observations. $\Phi = [\phi_1, \phi_2, \cdots, \phi_m] \in R^{d \times m}$ is an over-complete dictionary ($m > d$), and the columns ϕ_i are visual words or basis functions. m is the size of the dictionary. $S_i = [S_{i1}, \cdots, S_{im}]^T$ are the coefficients(sparse codes) which are independent with each other. λ is a deterministic scale parameter of Laplacian distribution for coefficients. $\nu \sim \mathcal{N}(0, \sigma^2 I_n)$ is small Gaussian noise.

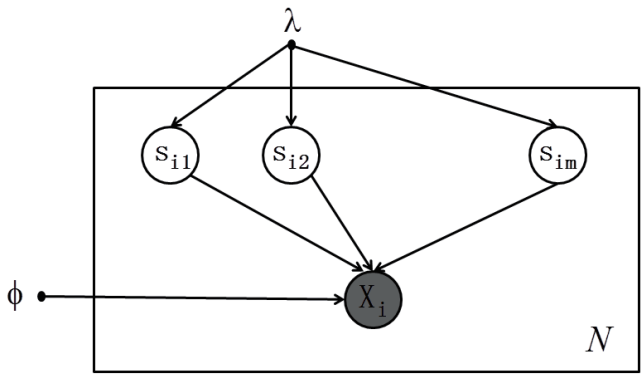


Fig. 1. The graphical model representation of sparse coding with Laplacian prior. Here random variables are denoted by open circles, observable variables by shaded circle, and deterministic parameters as smaller solid circle.

Denote $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, $S = \{S_i\}_{i=1}^N$. The joint distribution represented by the graphical model is

$$\begin{aligned} p(\mathcal{D}, S) &= \prod_{i=1}^{N} p(S_i|\lambda) p(\mathbf{x}_i|S_i, \Phi) \\ &= \prod_{i=1}^{N} \Big(\prod_{j=1}^{m} p(S_{ij}|\lambda) \Big) p(\mathbf{x}_i|S_i, \Phi), \end{aligned} \tag{2}$$

where $p(S_{ij}|\lambda)$ is a Laplace distribution

$$p(S_{ij}|\lambda) = \frac{\lambda}{2} \exp(-\lambda |S_{ij}|), \tag{3}$$

and the likelihood is a Gaussian distribution given by

$$p(\mathbf{x}_i|S_i, \Phi) = (2\pi\sigma^2)^{(-d/2)} \exp(-\frac{\|\mathbf{x}_i - \Phi S_i\|^2}{2\sigma^2}). \tag{4}$$

Using Bayesian rule $p(S|D) \propto p(D|S)p(S)$, and from (3) and (4), the MAP estimate $\hat{S}$ and dictionary $\hat{\Phi}$ are given by

$$< \hat{S}, \hat{\Phi} >= arg \min_{S_i, \Phi} \sum_i \{\|\mathbf{x}_i - \Phi S_i\|^2 + \mu \|S_i\|_1\}. \tag{5}$$

where the first term is the reconstruction error and the second term is the l_1 sparsity regularizer. $\mu = 2\sigma^2\lambda$ is a regularization parameter that controls the tradeoff between the reconstruction error and sparsity. Many efficient algorithms [24] have been developed to solve the l_1 minimization in (5).

After the dictionary is learned, the sparse coding of a new signal $\mathbf{x}$ can be obtained by MAP estimate:

$$\hat{s} = arg \max_s p(s|\mathbf{x}, \Phi) = arg \max_s p(\mathbf{x}, s|\Phi), \tag{6}$$

that is

$$\hat{s} = arg \min_s \|\mathbf{x} - \Phi s\|^2 + \mu \|s\|_1. \tag{7}$$

B. Sparsity regularization

The sparsity of sparse representation is controlled by a sparsity regularization term and its associated parameters. The choice of the functional form of the regularizer and its parameters is a challenging task [25]. Various regularizers have been proposed for different purposes and applications. The existing regularizers include l_0 norm or l_1 norm (Laplacian prior)[26][27], Gaussian prior[28], KL prior[3], Laplacian scale mixture prior [29] and reweighted l_1 norm [30], mixture of exponential (MOE) and Jeffreys mixture of exponentials (JOE) prior[25]. In [28], the EM algorithm was used to solve the sparse codes and the regularization parameters of Gaussian prior. In [31], the adaptive lasso was proposed, where adaptive weights were used for penalizing different coefficients in the l_1 penalty. In [30], reweighted l_1 minimization was analysed and demonstrated to enhance sparsity when compared with l_1 norm. In [25], sparsity regularization terms were designed based on the minimum description length (MDL) principle. A family of universal regularizer was proposed and shown to enjoy several desirable theoretical and practical properties such as statistical consistency, improved robustness to outliers in the data, and improved sparse signal recovery when compared with the traditional l_0 and l_1 norms.

C. Dictionary learning

There are two categories of dictionary learning approaches, unsupervised data-driven dictionary learning and supervised dictionary learning. Unsupervised dictionary learning are designed to produce dictionaries useful for images reconstruction, e.g., the K-SVD algorithm [24], the method of optimal directions (MOD) [32], the least squares optimization[33] and gradient descent [34]. They do not utilize class information about images in the training set. Dictionaries learned from natural scenes were introduced in [34] for modeling the spatial receptive fields of simple cells in the mammalian visual cortex, and were used successfully for different recognition tasks[35]. Recently, many algorithms have been proposed to enhance the discrimination of visual dictionaries through supervised learning, as discussed in the last section.

D. Locality preserving sparse coding

Some recent research work suggested that image space is actually a smooth low dimensional sub-manifold embedded in a high dimensional ambient space. Standard sparse coding does not include locality constraints explicitly, thus the locality or the geometrical structure among the instances to be encoded are lost. Research has been done to address this problem by embedding the manifold structure into sparse coding algorithm as regularization terms [8], [21], [22], [23]. In [8], locality-constrained Linear Coding (LLC) was proposed. LLC incorporates locality constraint instead of the sparsity constraint. Using the K nearest neighbors to select the local bases from the codebook, a faster approximate LLC was implemented. In [23], the geometrical structures are encoded in two situations. When data points distribute on a single manifold,it is explicitly modeled by locally linear embedding algorithm combined with k-nearest neighbors. When data points often lie on multiple manifolds, sparse representation algorithm combined with k-nearest neighbors is utilized to construct the topological structures. After obtaining the local fitting relationship,these two topological structures are then embedded into sparse coding algorithm as regularization terms to formulate the corresponding objective functions of dictionary learning.

Another related work is group sparse coding. Group sparse coding was proposed in [36], [37], in which similar features are encoded simultaneously, other than encoding each feature one by one in sparse coding. A blockwise nonzero entry distribution constraint was imposed on the sparse codes matrix, thus the similarities among the features within the same group can be preserved. Furthermore, nonoverlapping group lasso [38], [39] and overlapping group lasso [40], [41] were proposed to deal with nonoverlapping groups and overlapping groups respectively.

III. Supervised Bayesian sparse coding

We propose a Bayesian sparse coding model for a multi-class classification problem.

Assume signals from different classes have different prior distributions, and signals from the same class have the same prior distribution. We build a model for each of the C classes. For the cth class, the graphical model of Bayesian sparse coding on the corresponding training data is depicted in Figure 2. $\mathbf{x}_i \in R^d (i = 1, \ldots, N_c)$ are observations of the cth class from the training set. Denote $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N_c}$, $\Lambda = [\lambda_1, \ldots, \lambda_m]^\top$, $S = \{S_i\}_{i=1}^{N_c}$. It should be pointed out that $\lambda_i (i = 1, \cdots, m)$ are hidden random variables here.

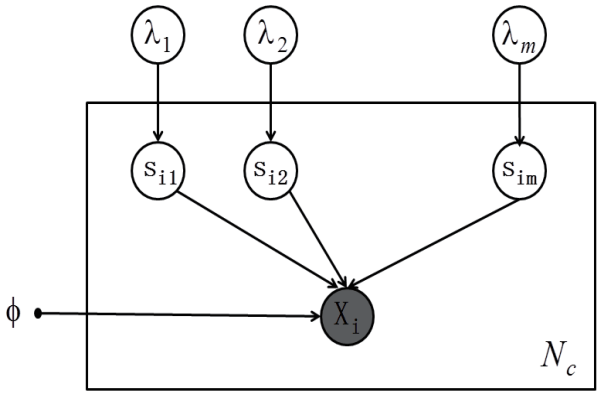


Fig. 2. The graphical model of Baysesian sparse coding and dictionary learning for the cth class.

The joint distribution represented by the graphical model is

$$\begin{aligned} p(\mathcal{D}, S, \Lambda) &= p(\Lambda) \prod_{i=1}^{N_c} p(S_i|\Lambda) p(\mathbf{x}_i|S_i, \Phi) \\ &= \prod_{j=1}^{m} p(\lambda_j) \prod_{i=1}^{N_c} \prod_{j=1}^{m} \Big(p(S_{ij}|\lambda_j) \Big) p(\mathbf{x}_i|S_i, \Phi), \end{aligned} \tag{8}$$

$p(\lambda_j)$ is a Gamma distribution with the form

$$p(\lambda_j) = \Gamma(\lambda_j|\alpha_j, \beta_j) = \frac{1}{\Gamma(\alpha_j)} \beta_j^{\alpha_j} \lambda_j^{\alpha_j - 1} e^{-\beta_j \lambda_j}, \lambda_j \in R^+ \tag{9}$$

where α_j and β_j are its shape and scale parameters, respectively. $p(S_{ij}|\lambda_j)$ is a Laplace distribution

$$p(S_{ij}|\lambda_j) = \frac{\lambda_j}{2} \exp(-\lambda_j |S_{ij}|), \tag{10}$$

and the likelihood is a Gaussian distribution given by

$$p(\mathbf{x}_i|S_i, \Phi) = (2\pi\sigma^2)^{(-d/2)} \exp(-\frac{\|\mathbf{x}_i - \Phi S_i\|^2}{2\sigma^2}). \tag{11}$$

The distribution over S_{ij} is a continuous mixture of Laplacian distributions with different inverse scale, and it can be computed by integrated out λ_j

$$p(S_{ij}) = \int p(S_{ij}, \lambda_j) d\lambda_j = \int p(S_{ij}|\lambda_j) p(\lambda_j) d\lambda_j. \tag{12}$$

Note that for most choices of $p(\lambda_j)$, we do not have an analytical expression for $p(S_{ij})$. Such a distribution is called a Laplacian Scale Mixture (LSM) [29].

A. Reweighted l_1 minimization

In this subsection, we will introduce an analytical solution to the Bayesian sparse coding, which results in a nonconvex log-sum regularizer.

With a conjugate Gamma prior distribution in (9), we can compute $p(S_{ij})$ analytically [29].

$$p(S_{ij}) = \int p(S_{ij}|\lambda_j)p(\lambda_j)d\lambda_j = \frac{\alpha_j \beta_j^{\alpha_j}}{2(\beta_j + |S_{ij}|)^{\alpha_j+1}}. \quad (13)$$

As in the standard sparse coding and dictionary learning, from (11) and (13), the MAP estimate $\hat{S}$ and dictionary is given by

$$< \hat{S}, \hat{\Phi} >= arg \max_{S_i,\Phi} \sum_i \{\log p(\mathbf{x}_i|S_i, \Phi) + \log p(S_i)\}, \quad (14)$$

that is

$$\begin{aligned} < \hat{S}, \hat{\Phi} > &= arg \min_{S_i,\Phi} \sum_i \{\|\mathbf{x}_i - \Phi S_i\|^2 \\ &+ 2\sigma^2 \sum_{j=1}^{m} (\alpha_j + 1) \log(\beta_j + |S_{ij}|)\}. \quad (15) \end{aligned}$$

The nonconvex log-sum regularizer in (15) was also proposed in [25], which was called mixture of exponential (MOE) prior. As discussed in [25], the parameters α and β are noninformative, in the sense that the probability distribution of the sparse codes does not depend on their choice. Therefore, all shape parameters α_j and scale parameter β_j were set to be same, that is, $\alpha_j = \alpha$ and $\beta_j = \beta$ for $j = 1, \cdots, m$. Thus the graphical model is simplified as in Fig.3, where all coefficients have the same prior distribution. The MAP estimate $\hat{S}$ and dictionary $\hat{\Phi}$ are given by

$$\begin{aligned} < \hat{S}, \hat{\Phi} > &= arg \min_{S_i,\Phi} \sum_i \{\|\mathbf{x}_i - \Phi S_i\|^2 \\ &+ 2\sigma^2(\alpha + 1) \sum_{j=1}^{m} \log(\beta + |S_{ij}|)\}. \quad (16) \end{aligned}$$

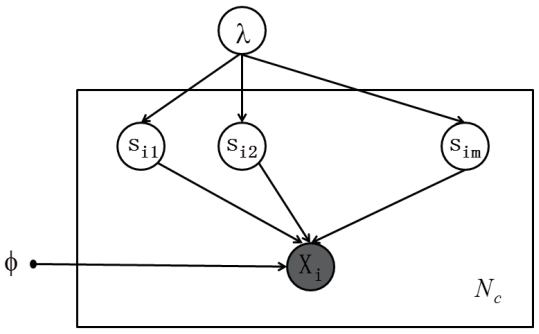


Fig. 3. The graphical model of sparse coding with log-sum regularizer.

After the dictionary is learned, the sparse coding of a new signal $\mathbf{x}$ can be obtained by MAP estimate:

$$\hat{s} = arg \min_s \|\mathbf{x} - \Phi s\|^2 + \gamma \sum_{j=1}^{m} \log(\beta + |s_j|). \quad (17)$$

where $\gamma = 2\sigma^2(\alpha+1)$. Some work has been done to solve the sparse coding problem with the nonconvex regularizer in (17) [42], [43]. The iterative reweighted l_1 algorithm was used in [30], [25], which consists of solving a sequence of weighted l_1 minimization problems where the weights used for the next iteration are computed from the value of the current solution.

$$s^{t+1} = arg \min_s \|\mathbf{x} - \Phi s\|^2 + \gamma \sum_{j=1}^{m} \frac{1}{\beta + |s_j|^t|} |s_j|. \quad (18)$$

B. Weighted l_1 minimization

In this subsection, we introduce a variational approach to the Bayesian sparse coding, which results in a convex weighted l_1 regularizer.

Instead of integrating λ out as in (13), we use variational inference to calculate $p(S)$.

$$\log p(S) = \mathcal{L}(q, S) + KL(q(\Lambda)\|p(\Lambda|S)), \quad (19)$$

where

$$\begin{aligned} \mathcal{L}(q, S) &= \int_\Lambda q(\Lambda) \log \frac{p(S, \Lambda)}{q(\Lambda)} d\Lambda, \quad (20) \\ KL(q(\Lambda)\|p(\Lambda|S)) &= \int_\Lambda q(\Lambda) \log \frac{q(\Lambda)}{p(\Lambda|S)} d\Lambda. \quad (21) \end{aligned}$$

Here $q(\Lambda)$ is an approximate distribution of $p(\Lambda|S)$, which can be any probability distribution. From (19) we know

$$\log p(S) \geq \mathcal{L}(q, S),$$

and the equality holds only when $q(\Lambda) = p(\Lambda|S)$.

We use EM algorithm for the MAP estimate of the coefficients. In the E step, we update the approximate distribution $q(\Lambda) = \prod_j q(\lambda_j)$. In the M step, we update the coefficients and dictionary with the current $q(\Lambda)$.

E step: The Gamma distribution and Laplacian distribution are conjugate, that is, the posterior probability of λ_j given S_{ij} is also a Gamma distribution. Hence, the posterior of λ_j given S is a Gamma distribution with parameters $\alpha_j + N_c$ and $\beta_j + \sum_i |S_{ij}|$. Hence, we have

$$q(\lambda_j) = p(\lambda_j|S) = \Gamma(\alpha_j + N_c, \beta_j + \sum_i |S_{ij}|). \quad (22)$$

Then the expectation of λ_j is

$$\mathbb{E}_q[\lambda_j] = \frac{\alpha_j + N_c}{\beta_j + \sum_i |S_{ij}|}. \quad (23)$$

When there are enough training examples, $N_c >> \alpha_j$, the influence of the hyperparameters α_j and β_j are neglectable and therefore can be set as constants. Then the expectation can be simplified as

$$\mathbb{E}_q[\lambda_j] \approx \frac{1}{\beta_0 + \{\sum_i |S_{ij}|\}/N_c}. \quad (24)$$

Here β_0 is a small positive number to make the denominator nonzero.

M step: The complete log-likelihood $\log p(S, \Lambda)$ can be calculated analytically

$$\begin{aligned}\log p(S,\Lambda) &= \sum_i \log p(S_i,\Lambda) = \sum_{i,j} \log p(S_{ij},\lambda_j) \\ &= -\sum_{i,j}\lambda_j|S_{ij}| + N_c\sum_j\{\log\frac{\lambda_j}{2} + \log p(\lambda_j)\}.\end{aligned} \tag{25}$$

Then from (20) and (25)

$$\begin{aligned}\mathcal{L}(q,S) &= -\mathbb{E}_q[\sum_{i,j}\lambda_j|S_{ij}|] + f_0 \\ &= -\sum_{i,j}\mathbb{E}_q[\lambda_j]|S_{ij}| + f_0,\end{aligned} \tag{26}$$

here $f_0(\Lambda) = N_c\sum_j \mathbb{E}_q[\{\log\frac{\lambda_j}{2} + \log p(\lambda_j)\}] - \mathbb{E}_q[q(\Lambda)]$ is a constant which is irrelevant to S.

Since $\log p(S) \geq \mathcal{L}(q,S)$, that is, $\mathcal{L}(q,S)$ is the lower bound of $\log p(S)$. The optimization problem in (14) can be reformulated as

$$< \hat{S}, \hat{\Phi} >= arg\max_{S_i,\Phi}\sum_i\{\log p(\mathbf{x}_i|S_i,\Phi) + \mathcal{L}(q,S)\}. \tag{27}$$

Insert (26) to (27), we have

$$< \hat{S}, \Phi >= arg\min_{S,\Phi}\sum_i\{\|\mathbf{x}_i - \Phi S_i\|_2^2 + 2\sigma^2\sum_j\mathbb{E}_q[\lambda_j]|S_{ij}|\}. \tag{28}$$

The weighting factor $\mathbb{E}_q[\lambda_j]$ in (24) has an appealing intuitive interpretation. When the mean of the jth coefficient is small, $\mathbb{E}_q[\lambda_j]$ will be large, which increase the chance that it will be smaller in the next iteration. On the other hand, when the mean of the jth coefficient is large, $\mathbb{E}_q[\lambda_j]$ will be small, such that the jth component is not penalized to be large in the next iteration.

The minimization problem in (28) can be solved using an iterative method[24]. There are two main steps in each iteration. The first step is sparse coding based on the current dictionary. The second step is dictionary update with the current coefficients. For a given dictionary Φ, the MAP estimate $\hat{s}$ of the coefficients in (28) can be solved by the existing l_1 weighted minimization algorithms [30]:

$$\hat{S}_i = arg\min_{S_i}\|\mathbf{x}_i - \Phi S_i\|_2^2 + 2\sigma^2\sum_j\mathbb{E}_q[\lambda_j]|S_{ij}|. \tag{29}$$

Now we consider the dictionary learning in (28). There are many algorithms proposed for the dictionary learning, e.g. the gradient descent [34], method of optimal directions (MOD) [44] and K_SVD [24]. Here we use a modified MOD for dictionary learning.

The MOD was formulated as

$$\Phi = XS^T(SS^T)^{-1}. \tag{30}$$

Here $X \in R^{d\times N_c}$, and $S \in R^{m\times N_c}$. Since the coefficients of some components(words) for the training samples in one class may be all zeros, that means these words are not used to reconstruct the signal in this class, which are called inactive words. It is unnecessary to update these words. The remaining words are active words for this class, and denoted as Φ_{act}. The corresponding coefficients of active words are denoted as S_{act}. The modified MOD can be described as follows.

$$\Phi_{act} = XS_{act}^T(S_{act}S_{act}^T + \rho I)^{-1}. \tag{31}$$

Here ρ is a small constant to make the matrix invertible.

The proposed Bayesian sparse coding and dictionary learning algorithm with weighted l_1 regularizer can be summarized as follows.

Initialization.
  Initialize $\mathbb{E}_q[\lambda_j]$ and Φ;
  Calculate the MAP estimate of S_i.
E step.
  Calculate expectation of Λ using (24) .
M step.
  Update MAP estimate of S_i by solving (29);
  update dictionary Φ using (31).
Repeat the E step and M step.

The sparse coding in (28) is similar to group sparse coding in the way that features from a group (class) are encoded simultaneously. But different from group sparse coding, where a blockwise nonzero entry distribution constraint was imposed on the sparse codes matrix, we use the regularization parameters to preserve the similarities among the features within the same group.

After training, we have learned the dictionary Φ^c and weighting factors $\Lambda^c = [\lambda_1^c, \cdots, \lambda_m^c]$ for the cth class. The model of the cth class can be simplified as shown in Fig.4. The probability distributions of the coefficients are still Laplacian distribution

$$p(s_j) = \frac{\lambda_j^c}{2}\exp(-\lambda_j^c|),$$

where

$$\lambda_j^c = \mathbb{E}_q[\lambda_j] = \frac{1}{\beta_0 + \{\sum_i|S_{ij}^c|\}/N_c}. \tag{32}$$

Here λ_j^c means the weighting factor of the jth component for the cth class, and S_{ij}^c means the coefficients of the jth component for the ith training samples from the cth class. It should be noted that each class has its own set of regularization parameters, which will determine the discriminative basis for each class. If λ_j^c is small, it means the jth basis is more discriminative for the cth class.

For a new test signal $\mathbf{x}$, we need to predict its category. That is

$$c^* = arg\max_c p(\mathbf{x}|\Phi^c,\Lambda^c), \tag{33}$$

here

$$p(\mathbf{x}|\Phi^c,\Lambda^c) = \int p(\mathbf{x},\mathbf{s}|\Phi^c,\Lambda^c)d\mathbf{s} = \int p(\mathbf{x}|\mathbf{s},\Phi^c)p(\mathbf{s}|\Lambda^c)d\mathbf{s}. \tag{34}$$

There does not exist the analytic formulation for the above marginalization. We can first get the MAP estimate $\mathbf{s}^{*c}$,

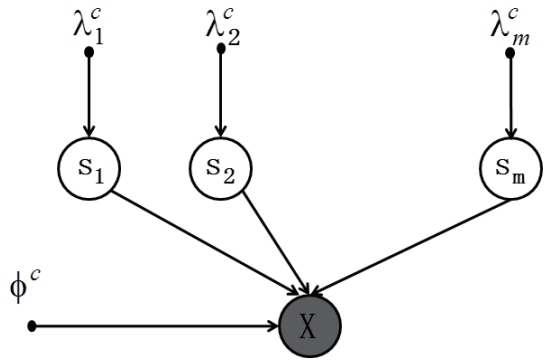


Fig. 4. The graphical model of sparse coding with weighted l_1 regularization for the c class.

then plug the $\mathbf{s}^{*c}$ into (34) to give a point estimate of the $p(\mathbf{x}|\Phi^c, \Lambda^c)$ by $p(\mathbf{x}|\mathbf{s}^{*c}, \Phi^c)p(\mathbf{s}^{*c}|\Lambda^c)$.

The MAP estimate of the coefficients of the signal $\mathbf{x}$ from the c class can be written as

$$\mathbf{s}^{*c} = arg\min_{\mathbf{s}}\{\|\mathbf{x} - \Phi^c\mathbf{s}\|_2^2 + \sum_j \lambda_j^c|\mathbf{s}_j|\}. \quad (35)$$

Therefore we have

$$c^* = arg\min_c \|\mathbf{x} - \Phi^c\mathbf{s}^{*c}\|_2^2 + \sum_j \lambda_j^c|\mathbf{s}_j^{*c}|. \quad (36)$$

IV. Experiments

In this section, we test the proposed model with different recognition tasks, including handwritten digits and handwritten English characters. SIFT features are used in all experiments. We used sparse modeling toolbox SPAMS to solve the sparse coding with weighted l_1 norm, which is downloadable from http://spams-devel.gforge.inria.fr/downloads.html. We compare the proposed sparse coding algorithms, the weighted l_1 (WL_1) in (35), with the reweighted l_1 ($reWL_1$) in (18) and L_1 in (7). For WL_1, We use (24) to estimate the regularization parameters, and β_0 is set as 0.01 in all experiments. For $reWL_1$ in (18), γ was set as 0.01 and β as 0.05. For L_1 in (7), μ was set as 0.15.

A. USPS dataset

The USPS dataset has 7291 training images and 2007 test images of size 16x16.

We extracted 128 dimensional SIFT feature for each digit image. The dictionary size is 256. The initial dictionary was learned from all training features using (5). The category-specific dictionaries and regularization parameters were learned using the EM algorithm discussed in the last section.

First, we investigated the discrimination of the learned dictionary and stability of the proposed SBSC. We compared the sparse codes obtained from our proposed supervised Bayesian sparse coding WL_1 (35) with standard sparse coding with l_1 regularizer (7) and $reWL_1$ in (18). The averages of sparse codes for all test examples of digits 0 and digit 6 were shown in Fig.5. Note the scale difference of the y axis. It can be seen that almost all words in the dictionary were used to represent the test examples using standard sparse coding L_1 in Fig.5(top), therefore the averages of sparse codes are very small (< 0.02). As discussed in [37], the codes for the group or category are not sparse, even though the code for each sample is sparse. However for our supervised Bayesian sparse coding in Fig.5(bottom), the averages of a small set of coefficients are much larger than others, thus only a small set of words were used to represent all test examples in this category. This means these words are discriminative for this category. Codes of WL_1 are also more sparse than codes of $reWL_1$ in Fig.5(middle). To investigate the stability of coding further, two similar images were shown in top panels of Fig.6. The sparse codes from standard sparse coding (7) and $reWL_1$ were shown in Fig.6(b) and Fig.6(c) respectively. It can be seen that as discussed in [30], $reWL1$ improved the sparsity of L_1, but the codes for two similar images are still dissimilar to each other. This is because similarity preserving was not considered in these two coding methods. The sparse codes from our proposed SBSC WL_1 were shown in Fig.6(d), which are very similar. This demonstrates the stability of the proposed SBSC. From Fig.6, we can see that the codes for one image from our proposed SBSC are less sparse than that from standard sparse coding. This is because that training samples from the same category may still be very different, therefore more regularization parameter λ_j^c in (32) will be small. As suggested in [20], locality was more essential than sparsity, therefore we focus on stability rather than sparsity.

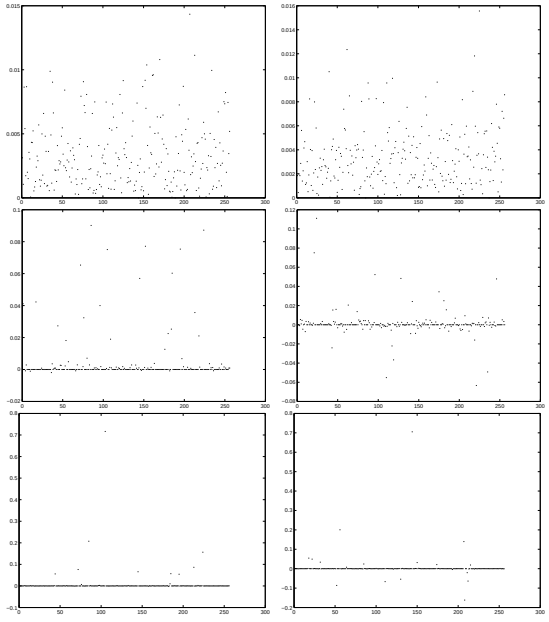


Fig. 5. Comparison of the proposed WL_1 (bottom) with standard sparse coding L_1 (top) and $reWL_1$ (middle). The average sparse codes of all test examples of digit 0 (left panel) and digit 6 (right panel) are displayed.

Secondly, we tested our model for classification. The accuracy on the test set are 98.90% using WL_1, which is slightly better than that of $reWL_1$, 98.36%. But since sparse coding for $reWL_1$ in (18) has to be solved iteratively, the test time for $reWL_1$ is much longer than WL_1. We have observed that ten iterations are needed to converge. Thus, the cost of

sparse coding with the reweighted L_1 regularizers, is ten times that of the WL_1. We also tested the standard sparse coding $L1$. The accuracy of L_1 is 98.36%, same as $reWL_1$.

Finally, we compare our results with Mairal et al.'s unsupervised and supervised approaches in [9]. The best results of these approaches were shown in Table I. Our error rate is significantly better than theirs on USPS dataset. But it should be noted that we used SIFT feature in all our experiments, rather than the raw patch in [9].

TABLE I. ERROR RATE FOR USPS FOR DIFFERENT APPROACHES

approaches	Mairal et al. [9] (unsupervised)	Mairal et al. [9] (supervised)	$reWL_1$ [25]	SBSC (WL_1)
Error rate	4.58	2.84	1.64	**1.10**

B. Handwritten character dataset

We tested the proposed model on handwritten character dataset downloaded from (http://ai.stanford.edu/∼btaskar/ocr/). The dataset has 52152 English characters. The image size is 16×8 pixels. The dictionary size is 128. We randomly selected M images as the training set, the rest as the test set. The average accuracies of 10 runs for different M were shown in Table II. We also tested the $reWL1$ on this dataset. Our results are better for all training numbers than $reWL1$.

We compared our results with the previous work [3], in which a differentiable smooth KL prior for sparse coding was proposed to improve the prediction performance over L1-prior, and supervised dictionary learning through back-propagation further improved the performance. As shown in Table II, our results are better than their KL prior and back-propagation when the number of training samples is less than 20000, but not as good as their results of back-propagation for $M = 20000$. This may be because the codes obtained by the proposed supervised Bayesian learning are less sparse and discriminative when the training set is too large.

TABLE II. RECOGNITION ACCURACY FOR HANDWRITTEN CHARACTER DATASET

Training	L1[3]	KL[3]	KL+BP[3]	$reWL_1$	SBSC
100	44.0	49.4	50.7	30.21	**51.69**
500	63.7	69.2	69.9	63.54	**72.73**
1000	69.5	75.0	76.4	71.11	**78.58**
5000	78.9	82.5	84.2	84.89	**85.57**
20000	83.3	86.0	**89.1**	87.63	87.88

V. CONCLUSIONS

We made three main contributions in this paper. First, we built a Bayesian sparse coding model for each class, and sparse coding was formulated as a weighted l_1 minimization problem. It is about ten times faster than the reweighted l_1 minimization[25]. Second, we proposed a novel discriminative dictionary learning approach, which can be learned with the regularization parameters using variational EM algorithm. Finally, the instablity problem of sparse coding with l_1 and reweighted l_1 minimization was alleviated, and similar features can be encoded similarly using the same regularization parameters.

Since we assume that images from the same class have the same prior distribution, our method can only be applied to patch-level images, such as digits, characters and faces, where the images are taken as patches. For other images such as natural scenes, different regions in a image have different appearance and therefore different prior distributions, our method is not applicable.

REFERENCES

[1] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proc. International conference on Machine Learning*.

[2] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. Advances in neural information processing systems(NIPS)*, 2006, pp. 609–616.

[3] D. M. Bradley and J. A. Bagnell, "Differential sparse coding," in *Proc. Advances in neural information processing systems(NIPS)*, 2008.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Advances in neural information processing systems(NIPS)*, 2009.

[5] ——, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

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

[8] J. Wang, J. Yang, K. Yu, F. Lv, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

[9] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, 2012.

[10] Z. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[11] J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce, "Discriminative sparse image models for class-specific edage detection and image interpretation," in *Proc. European Conference on Computer Vision(ECCV)*, 2008.

[12] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010.

[13] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. International Conference on Computer Vision(ICCV)*, 2011.

[14] N. Zhou, Y. Shen, J. Peng, and J. Fan, "Learning inter-related visual dictionay for object recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2012.

[15] D. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2008.

[16] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010.

[17] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

[18] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010.

[19] J. Yang and M. Yang, "Top-down visual saliency via joint crf and dictionary learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2012.

[20] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Advances in neural information processing systems(NIPS)*, 2009, pp. 2223–2231.

[21] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, 2013.

[22] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, 2011.

[23] B. Liu, Y. Wang, Y. Zhang, and B. Shen, "Learning dictionary on manifolds for image classification," *Pattern Recognition*, vol. 46, pp. 1879–1890, 2013.

[24] M. Aharon, M. Elad, and A. Bruckstein, "The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[25] I. Ramirez and G. Sapiro, "Universal regularizers for robust sparse coding and modeling," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3850–3864, 2012.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J R Stat Soc Series B Stat Methodol*, vol. 58, no. 1, pp. 267–288, 1996.

[27] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[28] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153C–2164, 2004.

[29] P. Garrigues and B. Olshausen, "Group sparse coding with a laplacian scale mixture prior," in *Proc. Advances in neural information processing systems(NIPS)*, 2010.

[30] E. J. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[31] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[32] K. Engan, S. O. Aase, and J. H. Hakon-Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*

[33] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Advances in neural information processing systems(NIPS)*, 2006, pp. 801–808.

[34] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Res.*, vol. 37, pp. 3311–3325, 1997.

[35] H. Shan and G. W. Cottrell, "Looking around the backyard helps to recognize faces and digits," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[36] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE Conference on Computer Vision (ICCV)*, 2009.

[37] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Proc. Advances in Neural Information Processing Systems(NIPS)*, 2009.

[38] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J R Stat Soc Series B Stat Methodol*, vol. 68, pp. 49C–67, 2006.

[39] J. Friedman, T. Hastie, and R. Tibshiraniz, "A note on the group lasso and a sparse group lasso," *technical report,arXiv:1001.0736v1*, 2010.

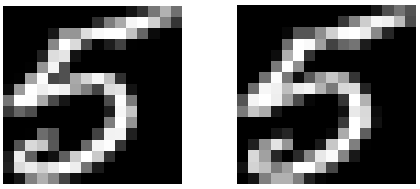
[40] L. Jacob, G. Obozinski, and J. P. Vert, "Group lasso with overlap and graph lasso," in *Proc. International Conference on Machine Learning(ICML)*, 2009.

[41] S. Mosci, S. Villa, A. Verri, and L. Rosasco, "A primal-dual algorithm for group sparse regularization with overlapping groups," in *Proc. Advances in neural information processing systems(NIPS)*, 2010.

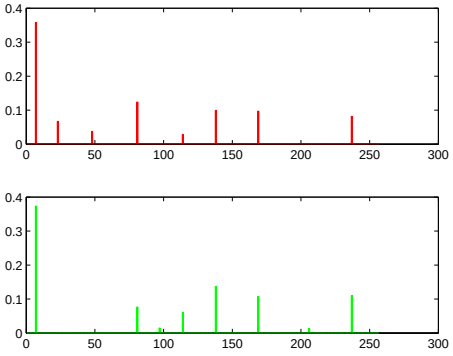
[42] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Stat.*, vol. 36, no. 4, p. 1509C1533, 2008.

[43] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with non-convex penalties and dc programming," *IEEE Trans. Signal Process.*, vol. 57, no. 12, p. 4686C4698, 2009.

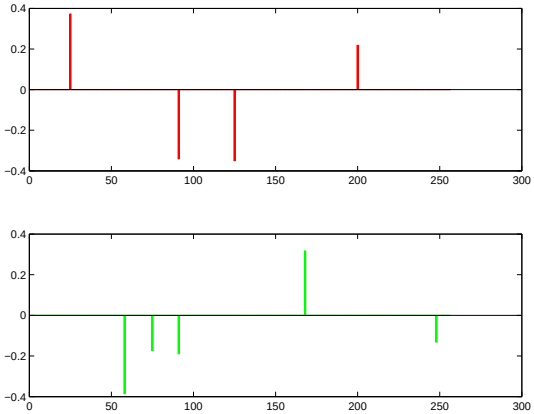
[44] K. Engan, S. O. Aase, and J. H. Hakon-Husoy, "Multi-frame compression: Theory and design," *Signal Process*, vol. 80, no. 10, p. 2121C2140, 2000.



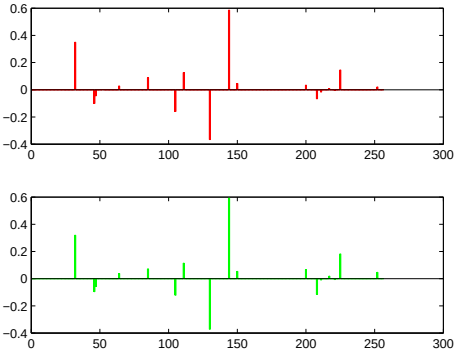
(a) Two similar images.



(b) Sparse codes of L_1.



(c) Sparse codes of $reWL_1$.



(d) Sparse codes of the proposed WL_1.

Fig. 6. Sparse codes for two similar images from different coding methods. The proposed supervised Bayesian sparse coding has similarity preserving property.