Learning Discriminative Sparse Representations for Hyperspectral Image Classification

Peijun Du, *Senior Member, IEEE*, Zhaohui Xue, Jun Li, *Member, IEEE*, and Antonio Plaza, *Fellow, IEEE*

***Abstract*—In sparse representation (SR) driven hyperspectral image classification, signal-to-reconstruction rule-based classification may lack generalization performance. In order to overcome this limitation, we presents a new method for discriminative sparse representation of hyperspectral data by learning a reconstructive dictionary and a discriminative classifier in a SR model regularized with total variation (TV). The proposed method features the following components. First, we adopt a spectral unmixing by variable splitting augmented Lagrangian and TV method to guarantee the spatial homogeneity of sparse representations. Second, we embed dictionary learning in the method to enhance the representative power of sparse representations via gradient descent in a class-wise manner. Finally, we adopt a sparse multinomial logistic regression (SMLR) model and design a class-oriented optimization strategy to obtain a powerful classifier, which improves the performance of the learnt model for specific classes. The first two components are beneficial to produce discriminative sparse representations. Whereas, adopting SMLR allows for effectively modeling the discriminative information. Experimental results with both simulated and real hyperspectral data sets in a number of experimental comparisons with other related approaches demonstrate the superiority of the proposed method.**

***Index Terms*—Hyperspectral image classification, discriminative sparse representation (DSR), total variation (TV), dictionary learning, sparse multinomial logistic regression (SMLR).**

I. Introduction

HYPERSPECTRAL remote sensing sensors allow for the acquisition of hundreds of contiguous bands for the same area on the surface of the Earth, and provide plenty of useful information that increases the accurate discrimination of spectrally similar materials of interest [1]. Hyperspectral image (HSI) has been extensively and increasingly exploited in classification, unmixing, fusion, target detection, land physical and chemical parameter estimation, and fast computing [2], [3]. Among many processing tasks, classification has attracted plenty of attention in the last decades [4]. This type of processing aims at assigning each pixel with one thematic class for an object in a scene [1].

Recently, sparse representation (SR) [5] has emerged as an effective way in many HSI processing tasks, such as target detection [6] and spectral unmixing [7], [8]. Especially, SR can represent high-dimensional signals as linear composition of few non-zero coefficients based on a pre-defined dictionary, which is beneficial to yield state-of-the-art performance when applying SR to HSI classification [9]–[19].

In incorporating spatial information for SR-based classification, authors in [9] proposed a simultaneous subspace pursuit (SSP) method to incorporate SR model with spatial information for spectral-spatial HSI classification. Later, in [11], SR was extended into kernel space and spatial information was also considered. Similarly, authors in [12] proposed nonlocal weighted joint sparse representation method to improve the classification performance. In the exploration of sparse solver, the work in [10] exploited a Homotopy-based method to overcome the challenges which arise when limited labeled samples are available.

In SR-based spectral-spatial feature extraction, authors in [13] utilized the empirical mode decomposition and morphological wavelet transform to obtain spectral-spatial features, then a kernel-based sparse multitask learning was employed for classification. The work in [14] exploited sparse representation based on morphological attribute profiles for remotely sensed image classification. Authors in [15] considered that regions of different scales can incorporate complementary yet correlated information for classification, so they proposed a multiscale adaptive sparse representation model for HSI classification. More recently, our previous work [19] exploited HSI decomposition based on morphological component analysis and sparse representation, leading to accurate spectral-spatial classification result.

In SR-based manifold learning, authors in [16] used SR to build sparse graph, and they adopted sparse graph embedding technique to extract spectral feature. Similarly, the work in [17] incorporated local linear embedding and Laplacian eigenmap with SR in unified optimization problems for HSI classification. In SR-based feature selection, authors in [18] proposed a discriminative sparse multimodal learning method to integrate spectral and spatial features based on the concatenating strategy.

Manuscript received August 19, 2014; revised January 05, 2015 and March 23, 2015; accepted April 02, 2015. Date of publication April 15, 2015; date of current version August 12, 2015. This work was supported in part by the Jiangsu Provincial Natural Science Foundation (No. BK2012018), the Natural Science Foundation of China (No. 41471275), and the National Key Scientific Instrument and Equipment Development Project (No. 012YQ050250). The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Charles Creusere. *(Corresponding author: Zhaohui Xue.)*

P. Du and Z. Xue are with the Key Laboratory for Satellite Mapping Technology and Applications of National Administration of Surveying, Mapping and Geoinformation of China, with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, and also with the Collaborative Innovation Center of South China Sea Studies, Nanjing University, Nanjing 210023, China (e-mail: xzh2012@163.com).

J. Li is with the Guangdong Provincial Key Laboratory of Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China.

A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10071 Cáceres, Spain.

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

Digital Object Identifier 10.1109/JSTSP.2015.2423260

Generally, in SR based classification approaches, a given pixel can be sparsely represented by a few atoms from a given dictionary, and the obtained sparse representations carry out the class-label information. Therefore, the dictionary builds a bridge between the observed signal and its sparse code, which should have good discriminative power in order for SR to yield good performance. However, the dictionary in most of these applications is pre-defined, rather than learnt from a training set. A trivial strategy to build the dictionary is by using random sampling. In turn, learning a desired dictionary from the training set has gained popularity recently as it can further improve the effectiveness of SR [20], [21]. Some dictionary learning methods have been developed in terms of minimizing reconstruction error, such as K-SVD [22] which is generalized from K-means and has been widely used for natural image processing, and the majorization method (MM) which adopts a surrogate function to update dictionary in each step [23].

Another important issue in SR based HSI exploitation is the need for making sparse representations exhibit significant discriminative power, which is particularly crucial for its successful use in the classification stage. However, most available SR methods in HSI classification are based on the rule of signal-to-reconstruction error, where the label of an unknown pixel is assigned according to the label of the associated sub-dictionary that produces the minimum reconstruction error. This strategy may suffer from lack of generalization since it is dependent on the error measures and easily affected by noise. Separating dictionary learning from classification may result in a suboptimal dictionary for classification, so it is generally preferred to embed dictionary and the classifier learning in SR.

In computer vision, some advanced methods prefer to introduce a discriminative term in dictionary learning to produce discriminative sparse representations [24]–[27], allowing for a joint learning of the reconstructive and discriminative ingredients instead of simply using a reconstructive term. We call this method discriminative sparse representation (DSR).

The performance of DSR is highly related to the designed loss function and the adopted optimization strategy. In [24], a discriminative term with *square loss* is introduced into SR to model the classification error. Similarly, in [27] a label consistent term with *square loss* is designed to enhance the discriminative power of SR. These two studies then adopt K-SVD to find the optimal solution, and the two methods are termed D-KSVD and LC-KSVD, respectively. In [25], authors adopted *logistic loss* to build the discriminative term, and they alternatively optimized the dictionary and classifier via gradient descent. They also hints at the potential usage of *soft-max loss* to build the discriminative term. Based on graph theory, authors in [26] proposed a submodular dictionary learning (SDL) algorithm, where the entropy rate of random walk on a graph is used to design the discriminative term. The optimization of SDL is then considered as a graph partitioning problem, where the dictionary is updated by finding a graph topology that maximizes the objective function.

Particularly, some studies has exploited DSR in HSI processing. Authors in [28] modified an existing unsupervised learning method to learn the dictionary for HSI classification. Later, DSR was first exploited in [29], where block-structured dictionary learning and subpixel unsupervised abundance mapping were jointly considered. More recently, in [30] a *hinge loss* function inspired from learning vector quantization was designed to address the discriminative dictionary learning problem. In [31], a semi-supervised classification method is proposed by jointly learning the classifier and dictionary in a task-driven framework, where *logistic loss* function is adopted to build the discriminative term.

Despite the good performance of these methods, DSR still need to be further exploited in terms of designing effective *loss* function and optimization strategy when facing HSI. Currently, *square loss* and *logistic loss* are preferred due to the convenience in implementation. However, the generalization performance of *square loss* is limited and easily affected by data noise. Whereas, *logistic loss* is usually computationally intensive in HSI classification scenarios, where multi-class classification with high-dimensional space is often encountered [25]. Another observation is that, previous studies usually optimized the associated problems as a whole without balancing the generalization performance for different classes.

In this work, we develop a new DSR model that integrates dictionary and classifier learning for HSI classification. Specifically, sparse representation is driven by using a spectral unmixing by variable splitting augmented Lagrangian and total variation (SUnSAL-TV) method [32], dictionary learning is motivated by using gradient descent in a class-wise manner, and the discriminative information is modeled by optimizing a sparse multinomial logistic regression (SMLR) [33] classifier via a alternating direction method of multipliers (ADMM) [34] method also in a class-wise manner. The proposed approach, termed SMLR-DSR, learns discriminative sparse representations for HSI classification by alternatively conducting sparse representation, dictionary learning, and classifier learning.

HSI often exhibits spatial variability of the spectral signatures, which usually results in "salt-and-pepper" phenomena in pixel-wised classification [35]. Early studies incorporate spatial information in DSR via a Laplacian smoothness regularization term when conducting sparse representation [29]–[31]. In the proposed method, SUnSAL-TV is adopted to obtain the sparse representations, which can well model the spatial information by using total variation (TV) based spatial regularization. Our previous work has confirmed the effectiveness of TV in modeling spatial information [19] in spectral-spatial based HSI classification. This is one feature of our method. Another one is that we consider dictionary learning to produce more representative sparse representations via gradient descent, and we introduce SMLR into DSR to model the discriminative information. Previous work have highlighted the powerful performance of SMLR in various HSI classification scenarios [36]–[40]. However, previous methods usually adopt *square loss*, *logistic loss*, or *soft-max loss* to build the discriminative term in DSR. As the third feature, we design a class-oriented sub-problem optimization strategy to conduct dictionary and classifier learning. This optimization strategy can partially address the unbalanced classification issue which is very common in reality, which improves the generalization performance of the learnt model for specific class. Whereas, previous methods usually update all atoms or regressors as a whole in each iteration. Note that, these features compose our major contributions in this work and make the proposed method unique with regards to previously proposed approaches (e.g., [28]–[31]) in this area.

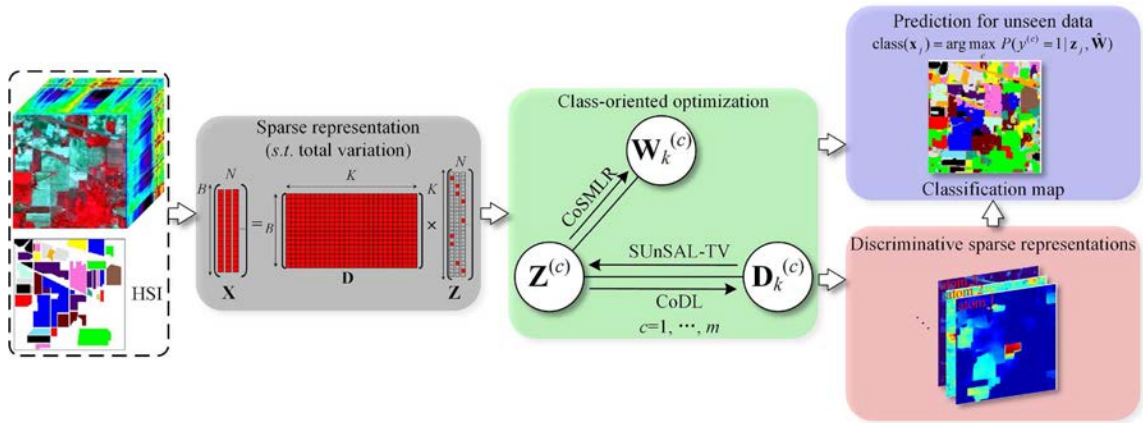


Fig. 1. Graphical illustration of the proposed method.

The remainder of this paper is organized as follows. Section II briefly introduces some related work. Section III formulates the proposed SMLR-DSR method. Section IV presents an experimental evaluation of the proposed method based on one simulated and two real hyperspectral data sets. Comparisons with some related techniques are also reported in this section. Section V concludes this paper with some remarks and hints at plausible future research lines. A graphical illustration of the proposed method is shown in Fig. 1.

II. Background

First of all, we introduce the notation that will be adopted throughout this paper. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{B\times N}$ be a hyperspectral data set with a B-dimensional signal (spectral signature) for each pixel $\mathbf{x}_j = [x_1, \ldots, x_B]^{\mathrm{T}}$, $j \in 1, \ldots, N$; let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{m\times N}$ represent the class label matrix for the input data, which uses a "1-of-m" encoding vector $\mathbf{y}_j = [y_j^{(1)}, y_j^{(2)}, \ldots, y_j^{(m)}]^{\mathrm{T}}$, such that $y_j^{(i)} = 1$ if $\mathbf{x}_j$ belongs to the class i and $y_j^{(i)} = 0$ otherwise; let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{B\times K}$ denote the dictionary; let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbb{R}^{K\times N}$ be the sparse codes for $\mathbf{X}$.

Generally, DSR problems take the form

$$\arg\min_{\mathbf{D},\mathbf{W},\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda_1\|\mathbf{Z}\|_{1,1} + \frac{1}{N}\sum_{i=1}^{N}\Gamma[\mathbf{y}_i, f(\mathbf{z}^\star(\mathbf{x}_i, \mathbf{D}), \mathbf{W})] + \gamma\Gamma_{\mathrm{reg}}(f), \quad (1)$$

where, the first row is the reconstruction term, whereas the second row formulate the discriminative term (F); λ_1 is a regularization parameter controlling the tradeoff between the reconstruction error and the *sparsity*; $\Gamma[\cdot]$ denotes the objective loss function with respect to the classifier $f(\cdot)$; $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]^{\mathrm{T}} \in \mathbb{R}^{m\times K}$ denotes the classifier parameter; $\Gamma_{\mathrm{reg}}(\cdot)$ is a regularizer controlling the generalization capacity of the classifier; γ is a tradeoff parameter between the loss and the regularization.

Based on (1), D-KSVD [24] adopts a *square loss* to build F, which is formulated as follows

$$F = \|\mathbf{Y} - \mathbf{W}\mathbf{Z}\|_2^2. \quad (2)$$

K-SVD can be used to optimize (2) based on augmented dictionary and input data.

LC-KSVD [27] further adds a label consistency regularization term to build F, which takes the form

$$F = \|\mathbf{Y} - \mathbf{W}\mathbf{Z}\|_2^2 + \alpha\|\mathbf{Q} - \mathbf{G}\mathbf{Z}\|_2^2, \quad (3)$$

where, α is a tradeoff parameter. Similar to D-KSVD, LC-KSVD can also be solved by using K-SVD.

In [25], [31], a *logistic loss* is used to build F, which can be formulated as follows

$$F = \frac{1}{N}\sum_{i=1}^{N}\log\left(1 + e^{-y_i\mathbf{w}^T\mathbf{z}_i^\star(\mathbf{x}_i,\mathbf{D})}\right) + \frac{\gamma}{2}\|\mathbf{W}\|_2^2. \quad (4)$$

Equation (4) solves binary classification problems. In order to conduct multiclass classification, this method should be incorporated with one-versus-rest or one-versus-one strategy.

III. Proposed Method

As aforementioned, the major differences of various DSR methods in the literature lie in the designed discriminative term and the optimization strategy. For instance, available techniques may use *square loss* [24], [27], *logistic loss* [25], [31], *soft-max loss* [41], and *hinge loss* [30] with Gaussian prior to build the discriminative term, and adopt K-SVD [27] and gradient descent [25], [29], [42] methods to optimize the associated problem. In this work, we adopt TV regularization to model the spatial information in SR, and adopt SMLR (MLR with Laplacian prior) to build the discriminative term. As for optimization, we design a class-oriented dictionary and classifier learning strategy to improve the performance.

A. Sparse Representation and Dictionary Learning via TV

Following [32], the sparse code of $\mathbf{X}$ with respect to dictionary $\mathbf{D}$ can be obtained by optimizing an ℓ_1-norm penalized problem with a nonisotropic TV for spatial regularization, which is given by

$$\mathbf{Z} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda_1\|\mathbf{Z}\|_{1,1} + \lambda_{TV}TV(\mathbf{Z}), \quad (5)$$

where, $\|\mathbf{Z}\|_{1,1} \equiv \sum_{j=1}^{N}\|\mathbf{z}_j\|_1$, $TV(\mathbf{Z}) \equiv \sum_{\{i,j\}\in\Omega}\|\mathbf{Z}_i - \mathbf{Z}_j\|_1$, Ω is a neighborhood, and λ_{TV} is another tradeoff parameter. Note that, the first term $\|\mathbf{Z}\|_{1,1}$ promotes *sparsity*

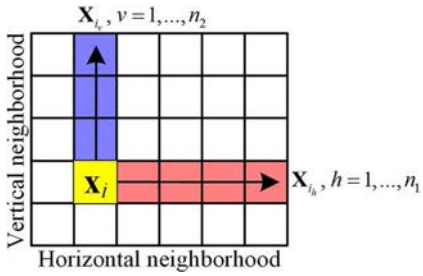


Fig. 2. Graphical illustration of the neighborhood adopted in TV regularization.

and the second term $TV(\mathbf{Z})$ guarantees piecewise smoothness in the sparse codes among neighboring pixels. In the optimization of (5), SUnSAL-TV method achieves state-of-the-art performance.

It is worth noting that TV plays a main role in the proposed method. In (5), $TV(\cdot)$ is a vector extension of the nonisotropic TV, which promotes smooth transitions in sparse representation. Ω denotes a set of horizontal and vertical neighborhoods in the image [see Fig. 2], which is free of size tuning. Precisely, for a single pixel $\mathbf{x}_i$, its horizontal neighborhood Ω_h denotes the pixel located in the right hand side of $\mathbf{x}_i$, i.e., $\mathbf{x}_{i_h}, h = 1,\ldots,n_1$. Whereas, the vertical neighborhood Ω_v is the pixel located in the top hand side of $\mathbf{x}_i$, i.e., $\mathbf{x}_{i_v}, v = 1,\ldots,n_2$. Thus, Ω is defined according to a cyclic boundary for pixels in the image. Let $\mathbf{H}_h$ denote a linear operator computing the horizontal differences between components in $\mathbf{Z}$ and the corresponding neighboring pixels, i.e., $\mathbf{H}_h\mathbf{Z} = [\mathbf{T}_1,\ldots,\mathbf{T}_{n_1}]$. Similar definition can be made for the vertical differences. Then, $TV(\mathbf{Z})$ has another equivalent form $\|\mathbf{HZ}\|_{1,1}$, where $\mathbf{HZ} = [\begin{matrix}\mathbf{H}_h\mathbf{Z}\\ \mathbf{H}_v\mathbf{Z}\end{matrix}]$. More details of TV-based SR can be found in [32] and references therein.

We further consider dictionary learning in (5) by solving

$$\langle \mathbf{D}, \mathbf{Z}\rangle = \arg\min_{\mathbf{D},\mathbf{Z}} \frac{1}{2}\|\mathbf{X} - \mathbf{DZ}\|_2^2 + \lambda_1\|\mathbf{Z}\|_{1,1} + \lambda_{TV}TV(\mathbf{Z}). \tag{6}$$

Equation (6) can be optimized by alternatively updating dictionary and sparse codes. The work in [43] theoretically proved the feasibility of TV dictionary model based on convex analysis and bounded variation functions. However, few studies in the literature address such dictionary learning problem.

B. SMLR-DSR

We adopt SMLR to model the discriminative information so that, for a given variable $\mathbf{z}_j$ (recall that, $\mathbf{z}_j$ is the sparse code obtained from SR), its probability belonging to class c is given by [44]

$$P\left(y^{(c)} = 1|\mathbf{z}_j, \mathbf{W}\right) = \frac{\exp\left(\mathbf{w}^{(c)^{\mathrm{T}}}\mathbf{h}(\mathbf{z}_j)\right)}{1 + \sum_{i=1}^{m}\exp\left(\mathbf{w}^{(i)^{\mathrm{T}}}\mathbf{h}(\mathbf{z}_j)\right)}, \tag{7}$$

where $\mathbf{h}(\cdot)$ is a mapping function. We adopt the Gaussian radial basis function (RBF) kernel [45] to improve the data separability in the output space, which is given by $\mathbf{h}(\mathbf{z}_j) = [1, K_{\mathbf{z}_j,\mathbf{z}_1},\ldots,K_{\mathbf{z}_j,\mathbf{z}_N}]$ with $K(\mathbf{z}_j,\mathbf{z}_i) = \exp(-\|\mathbf{z}_j - \mathbf{z}_i\|^2/(2\sigma^2))$, where $\sigma \in \mathbb{R}$ is a kernel parameter.

According to [33], $\mathbf{W}$ can be learnt by estimating its maximum a posteriori (MAP) as follows

$$\mathbf{W} = \arg\min_{\mathbf{W}} L(\mathbf{W}) = \arg\min_{\mathbf{W}} -[\ell(\mathbf{W}) + \log\psi(\mathbf{W})], \tag{8}$$

where $\ell(\mathbf{W})$ is the log-likelihood function given by

$$\ell(\mathbf{W}) = \sum_{j=1}^{N}\log P(y_j|\mathbf{z}_j, \mathbf{W}), \tag{9}$$

$\psi(\mathbf{W}) \propto \exp(-\lambda_2\|\mathbf{W}\|_1)$ is a Laplacian prior promoting *sparsity* on $\mathbf{W}$, and λ_2 acts as a regularization parameter controlling the degree of *sparsity*. By introducing (6) and (8) into (1), we obtain

$$\begin{aligned}\langle \mathbf{D}, \mathbf{W}, \mathbf{Z}\rangle =& \arg\min_{\mathbf{D},\mathbf{Z},\mathbf{W}} \min \frac{1}{2}\|\mathbf{X} - \mathbf{DZ}\|_2^2 + \lambda_1\|\mathbf{Z}\|_{1,1}\\ &- \ell(\mathbf{W}) + \lambda_2\|\mathbf{W}\|_1 + \lambda_{TV}TV(\mathbf{Z}).\end{aligned} \tag{10}$$

C. Class-Oriented Optimization

Optimizing problem (10) in a unified algorithm is very difficult since there is no explicit analytical link between $\mathbf{D}$ and $\mathbf{W}$ (or $\mathbf{Z}$ and $\mathbf{W}$). Therefore, we resort to alternating optimization between $\mathbf{D}$, $\mathbf{Z}$, and $\mathbf{W}$. Although alternating optimization cannot guarantee a global optimum, it allows to finding the optimum update in each step and good results have been reported both in early studies (e.g., [22], [23]) and our current work.

Precisely, we adopt alternating optimization with respect to $(\mathbf{Z}, \mathbf{D})$ and $\mathbf{W}$. When optimizing $(\mathbf{Z}, \mathbf{D})$, we first adopt SUnSAL-TV with respect to fixed $\mathbf{D}$ and then update $\mathbf{D}$ with fixed $\mathbf{Z}$ by using gradient descent in a class-wise manner. Finally, we adopt ADMM algorithm to class-wisely optimize $\mathbf{W}$ with fixed $\mathbf{Z}$. More details of this optimization strategy are given as follows.

1) Class-Oriented Dictionary Learning via Gradient Descent: The objective function of dictionary learning can be defined as follows

$$\mathcal{L}(\mathbf{D}|\widehat{\mathbf{W}}, \widehat{\mathbf{Z}}) \triangleq \frac{1}{2}\|\mathbf{X} - \mathbf{D}\widehat{\mathbf{Z}}\|_2^2 + C(\widehat{\mathbf{W}}, \widehat{\mathbf{Z}}), \tag{11}$$

where $C(\widehat{\mathbf{W}}, \widehat{\mathbf{Z}})$ is independent from $\mathbf{D}$. The gradient descent of $\mathcal{L}$ with respect to $\mathbf{D}$ can be obtained by

$$\nabla_{\mathbf{D}}\mathcal{L} = (\mathbf{X} - \mathbf{DZ})\mathbf{Z}^{\mathrm{T}}. \tag{12}$$

Then, the dictionary at iteration k can be updated according to $\mathbf{D}_{k+1} = \mathbf{D}_k + \rho\nabla_{\mathbf{D}}\mathcal{L}$, where $\rho = \min(\rho, \rho k_0/k)(k_0 = \text{maximum iterations}/10)$ is a heuristic rule following the work in [25], [27]. Finally, Algorithm 1 summarizes the class-oriented dictionary learning method.

2) Class-Oriented Classifier Learning via ADMM: ADMM serves as a useful optimization tool. However, it is computationally intensive when directly adopting ADMM to optimize $\mathbf{W}$ for HSI characterized with high-dimension and a large number of classes. To better illustrate this issue, we briefly illustrate the main step that adopts ADMM to optimize (8), where $\mathbf{W}$ is updated according to

$$\widehat{\mathbf{W}}_{k+1} = (\mathbf{B} - \mu\mathbf{I})^{-1}[\mathbf{B}\widehat{\mathbf{W}}_k - g(\widehat{\mathbf{W}}_k) - \mu(\mathbf{v}_k + \mathbf{b}_k)], \tag{13}$$

Algorithm 1 Class-oriented dictionary learning via gradient descent (CoDL).

Input : $\mathbf{X}$, $\mathbf{D}_0$, $\mathbf{Z}$, ρ, maximum iteration
Output: D
for $k = 0$ to maximum iteration **do**
  $\rho \leftarrow \min(\rho, \rho k_0/k)$
  for $c = 1$ to m **do**
    $\nabla_{\mathbf{D}}\mathcal{L} = (\mathbf{X} - \mathbf{D}_k^{(c)}\mathbf{Z}^{(c)})\mathbf{Z}^{(c)\mathrm{T}}$
    $\mathbf{D}_{k+1}^{(c)} = \mathbf{D}_k^{(c)} + \rho\nabla_{\mathbf{D}}\mathcal{L}$
  end for
end for
return $\mathbf{D}_{k+1}$

Algorithm 2 Class-oriented sparse multinomial logistic regression via ADMM (CoSMLR)

Input: $\mathbf{h}(\mathbf{Z})$, $\mathbf{Y}$, λ_2, maximum iteration
Output: $\mathbf{W}$
Initialization: $\mathbf{W}_0$, $\mathbf{B}$, μ, $\mathbf{v}_0 = \mathbf{W}_0$, $\mathbf{b}_0 = \mathbf{0}$
for $k = 0$ to maximum iteration **do**
  Calculate $\mathbf{P}$ according to (7)
  for $c = 1$ to $m - 1$ **do**
    $\nabla_{\mathbf{W}}\ell(\mathbf{W}) = \mathbf{h}(\mathbf{Z})(\mathbf{Y}^{(c)} - \mathbf{P}^{(c)})^{\mathrm{T}}$
    $\mathbf{W}_{k+1}^{(c)} = (\mathbf{B} - \mu\mathbf{I})^{-1}(\mathbf{B}\mathbf{W}_k^{(c)} - \nabla_{\mathbf{W}}\ell(\mathbf{W}) - \mu(\mathbf{v}_k^{(c)} + \mathbf{b}_k^{(c)}))$
    $\mathbf{v}_{k+1}^{(c)} = \mathrm{soft}(\mathbf{W}_{k+1}^{(c)} - \mathbf{b}_k^{(c)}, \lambda_2/\mu)$
    $\mathbf{b}_{k+1}^{(c)} = \mathbf{b}_k^{(c)} - \mathbf{W}_{k+1}^{(c)} + \mathbf{v}_{k+1}^{(c)}$
  end for
  $\mu \leftarrow 1.05\mu$
end for
return $\mathbf{W}_{k+1}$

where, $\mathbf{B} \triangleq -1/2[\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/m] \otimes \sum_{i=1}^{N} \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^{\mathrm{T}}$ is the Hessian matrix; $g(\widehat{\mathbf{W}}_k)$ is the gradient descent of $\ell(\mathbf{W})$ with respect to $\mathbf{W}$; $\mathbf{v}$ and $\mathbf{b}$ are two intermediate variables in the optimization. However, (13) is computationally intensive since the matrix inversion at each iteration requires $\mathcal{O}((dm)^3)$ (where d is the dimension of $\mathbf{h}$) operations, which results in the main computational complexity.

In order to make it numerically convenient, we propose to update $\mathbf{W}$ in a class-wise manner, which reduces the complexity to $\mathcal{O}(m(d)^3)$ for each iteration. Furthermore, inspired from the LORSAL (MLR via variable splitting and augmented Lagrangian) [36] algorithm, we approximate $\ell(\mathbf{W})$ by implementing a block-based Gauss-Seidel iterative procedure. On the one hand, the computational complexity can be reduced to $\mathcal{O}(md^2)$ since $\mathbf{B} \triangleq -(1/2)(1 - 1/m)\sum_{i=1}^{N} \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^{\mathrm{T}}$. On the other hand, it improves the generalization performance of the learnt model for specific classes. In this context, we reformulate this optimization strategy in Algorithm 2.

Actually, previous work has pointed out that component-wise updating procedure is an effective method in the optimization of SMLR [33], [46]. However, the choice of component in $\mathbf{W}$ to update at each iteration is one issue that has to be addressed. In this paper, we update those components relative to one class at each iteration. In addition, the objective function for learning SMLR is concave. Consequently, the component- or class-oriented weight update leads to monotonically decrease of the objective function and finds the global minimum without any risk of falling into local minimum (i.e., the ill-defined issue). Moreover, it has produced good results in our experiment.

D. Prediction for Unseen Data

After obtaining the learnt dictionary $\widehat{\mathbf{D}}$ and the classifier $\widehat{\mathbf{W}}$ from Algorithms 1 and 2 respectively, the label of a new incoming test sample $\mathbf{x}_j$ can be assigned by calculating the probability of the sample belonging to class $c \in \{1, \ldots, m\}$ based on its sparse code $\mathbf{z}_j$, which is given by

$$\mathrm{class}(\mathbf{x}_j) = \arg\max_c \; P\left(y^{(c)} = 1 | \mathbf{z}_j, \widehat{\mathbf{W}}\right). \tag{14}$$

The computational complexity of the proposed SMLR-DSR method comes from three parts. For each iteration, SUnSAL-TV has complexity of $\mathcal{O}(B^2N)$, CoDL costs $\mathcal{O}(BmN)$, and CoSMLR requires $\mathcal{O}(md^2)$. Therefore, the overall complexity of SMLR-DSR is $\mathcal{O}(B^2N)$ resulting from the SUnSAL-TV algorithm.

IV. Experiments

In this section, we evaluate the proposed approach by using one simulated hyperspectral data set generated from USGS library and two real hyperspectral data sets[1] respectively collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and the Reflective Optics Spectrographic Imaging System (ROSIS) instruments. The remainder of the section is organized as follows. First, we introduce the three hyperspectral data sets. Then, we describe the parameter settings and notations adopted in our experiments. Finally, we present several experiments with each considered data set.

A. Hyperspectral Data Sets

- The first simulated hyperspectral data set is generated from the USGS library, denoted splib06[2] and released in Sep. 2007. It comprises 224 spectral bands range from $0.4 - 2.5\,\mu$m. The simulated data cube is generated using a linear mixture model, with a certain number of randomly selected signatures as the endmembers and imposing abundance nonnegativity and sum-to-one constraints for each simulated pixel. Precisely, we first randomly select m signatures ($\mathbf{D} \in \mathbb{R}^{B\times m}$) from the library and generate a ground-truth map [see Fig. 3(b)]. Then, we randomly generate an abundance map for each class region $\{\mathbf{Z} = [\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}] \in \mathbb{R}^{m\times N}\}$, where the abundance values with respect to the specific class are set larger than others. Next, we generate the Gaussian iid noise ($\mathbf{E}$) according to a pre-defined signal-to-noise ratio (SNR in dB). At last, the observation data can be simulated according to the linear mixture model, i.e., $\mathbf{X} = \mathbf{D}\mathbf{Z} + \mathbf{E}$. Here, we set $m = 5$, $N = 128 \times 128$ (in pixel) and SNR $= 20$ dB[3]. A false color

[1]Available online: http://www.ehu.es/ccwintco/index.php.

[2]http://speclab.cr.usgs.gov/spectral.lib06.

[3]Note that, we generate a very noisy data on purpose.

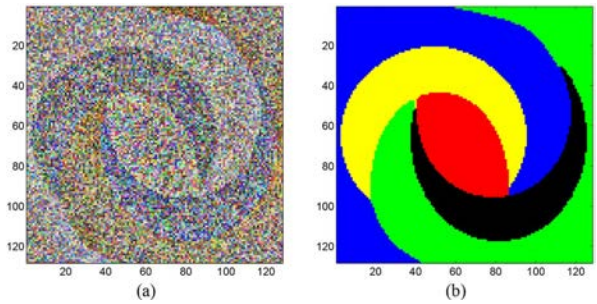


Fig. 3. Simulated hyperspectral data set (SNR $=$ 20 dB). (a) False color composite image (Red: 40, Green: 10, Blue: 200). (b) Ground-truth map containing 5 mutually exclusive classes.

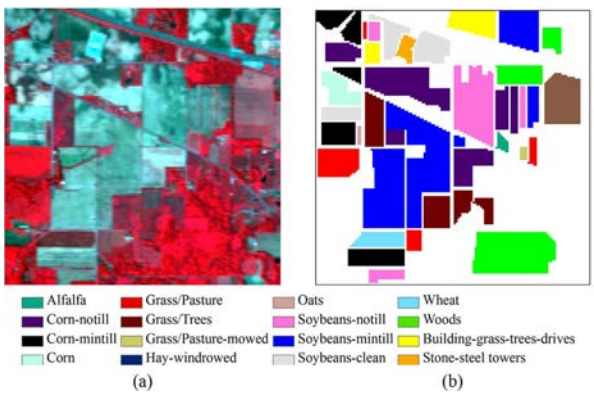


Fig. 4. AVIRIS Indian Pines data set. (a) False color composite image (Red: 57, Green: 27, Blue: 17). (b) Ground-truth map containing 16 mutually exclusive land-cover classes. The legend of this scene is shown at the bottom.

composite image of the simulated hyperspectral data set can be seen in Fig. 3(a).

- The second hyperspectral image used in the experiments was recorded by the AVIRIS sensor over the Indian Pines region in Northwestern Indiana in 1992. This scene, covers a mixed agricultural/forest area with 145×145 pixels and comprises 220 spectral bands in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, moderate spatial resolution of 20 m/pixel, and 16-b radiometric resolution. After an initial screening, several spectral bands were removed from the data set due to noise and water absorption phenomena, leaving a total of 200 radiance channels to be used in the experiments. A three-band false color composite image and the ground-truth map are shown in Fig. 4. A total of 10366 samples containing 16 classes are available, which are detailed in Table I. This scene constitutes a very challenging classification problem due to the significant presence of mixed pixels and unbalanced labeled classes.
- The third hyperspectral data set was acquired by the ROSIS sensor over the University of Pavia, Italy. The image size in pixels is 610×340, with very high spatial resolution of 1.3 m/pixel. The number of data channels in the acquired image is 103 (with spectral range from 0.43 to 0.86 μm). The ground-truth map contains nine classes of interest, with a total of 42776 labeled samples (see Table V). This training set is widely used in the HSI classification community and provided by the University of Pavia, who conducted the ground-truth data collection and labeled sample generation for this particular scene. A three-band false color composite image and the ground-truth map are shown in Fig. 5.

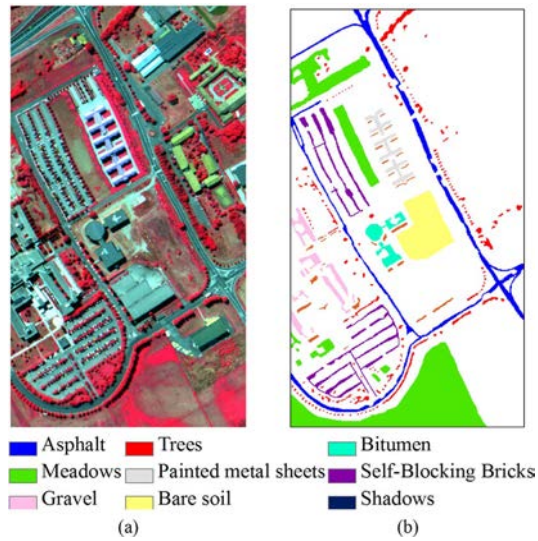


Fig. 5. ROSIS University of Pavia data set. (a) False color composite image (Red: 102, Green: 56, Blue: 31). (b) Ground-truth map containing 9 mutually exclusive land-cover classes. The legend of this scene is shown at the bottom.

TABLE I
16 GROUND-TRUTH CLASSES IN AVIRIS INDIAN PINES AND THE TRAINING AND TEST SETS FOR EACH CLASS

Class		#Samples	
No	Name	Train	Test
1	Alfalfa	5	49
2	Corn-no till	143	1291
3	Corn-min till	83	751
4	Corn	23	211
5	Grass/Pasture	50	447
6	Grass/Trees	75	672
7	Grass/Pasture-mowed	3	23
8	Hay-windrowed	49	440
9	Oats	2	18
10	Soybeans-no till	97	871
11	Soybeans-min till	247	2221
12	Soybean-clean till	61	553
13	Wheat	21	191
14	Woods	129	1165
15	Bldg-grass-tree-drives	38	342
16	Stone-steel towers	10	85
	Total	1036	9330

B. Experimental Setting

The parameter settings and notations adopted in our experiments are given as follows.

- For training set generation, we first randomly select a subset of labeled samples from the available ground-truth map. Then, we randomly choose some samples from the selected training set to build the dictionary. For the parameters involved with SMLR-DSR, we set $\lambda_1 = 1e-5$, $\lambda_2 = 1e-5$, $\lambda_{TV} = 1e-3$, and $\sigma = 0.8$.
- For classification, we report the overall (OA), average (AA), individual class accuracies (%), kappa statistic (κ), standard deviation, and computational time derived from averaging the results after conducting ten independent Monte Carlo runs with respect to the initial training set.
- For performance comparison, some strongly related DSR methods including D-KSVD, LC-KSVD, and SDL have

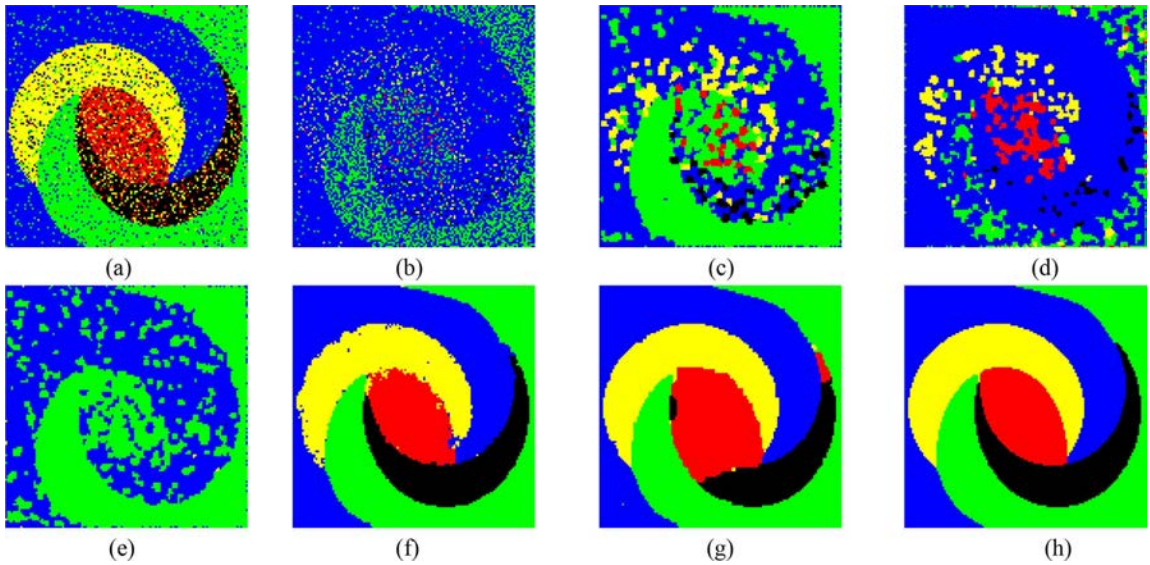


Fig. 6. Classification results obtained by different methods for the simulated hyperspectral data set. The OA in each case is reported in the parentheses. (a) SVM (84.84%). (b) SRC (44.26%). (c) D-KSVD (60.76%). (d) LC-KSVD (45.60%). (e) SDL (56.22%). (f) SMLR-DSR (98.03%). (g) LORSAL-MLL (95.64%). (h) Ground-truth.

been implemented[4]. Since the proposed method involves spatial information (by adopting the TV regularization), some state-of-the-art SSR methods including simultaneous subspace pursuit (SSP) [9] and simultaneous orthogonal matching pursuit (SOMP) [47] are also compared. Furthermore, some advanced spectral-spatial classification methods, such as LORSAL-MLL (LORSAL with graph-cut) [36] and MLR-GCK [39], are also included in this comparison. As the basic classifiers, support vector machine (SVM) [48] and sparse representation based classification (SRC) [49] methods are also compared.

- Finally, it should be noted that all the implementations were carried out using Matlab R2012b in a desktop PC equipped with an Intel Core i7 CPU (at 3.4 GHz) and 32 GB of RAM. We conduct cross-validation for selecting the parameters involved in the proposed method. The associated parameters for other considered methods are set to their recommended values, and we empirically found that these settings led to good performance.

C. Experiments With Simulated Hyperspectral Data Set

Experiment 1: We first test the proposed method for the simulated hyperspectral data set. 5% of labeled samples (a total of 820 labeled samples) are randomly chosen for training. A total of 250 samples from the training set are used to build the dictionary. Fig. 6 visually depicts the classification maps obtained by different methods, where the proposed method SMLR-DSR significantly outperforms others. Note that, LORSAL-MLL provides competitive result due to adopting graph-cut for spatial regularization.

Experiment 2: We then test the proposed method under different training scenarios. Fig. 7 plots the overall accuracy as a function of different number of labeled samples per class ($NL_{\rm per}$). As we can see from the figure, the proposed method outperforms others in each case. It is worth noting that even with very small training set (5 samples per class), SMLR-DSR can also provide good result (OA $>$ 90%). Again, LORSAL-MLL provides competitive result. However, it is no so stable for this data set.

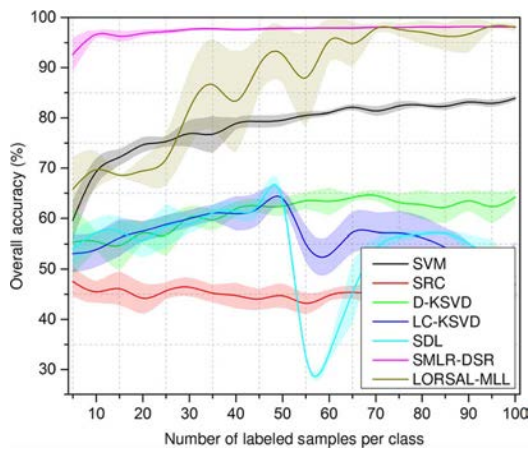


Fig. 7. Overall accuracy (OA) with standard deviation (colored area) as a function of different number of labeled samples per class ($NL_{\rm per}$) for the simulated hyperspectral data set.

Experiment 3: In the last experiment for this simulated data set, we test the robustness-to-noise of the proposed method. As shown in Fig. 8, the proposed method not always performs good. However, when SNR $>$ 12.5 (dB), SMLR-DSR exhibits significant superiority. Another observation is that SMLR-DSR and LORSAL-MLL converge to the same level due to the fact that the noise level becomes lower with the increase of SNR.

D. Experiments With AVIRIS Indian Pines Data Set

Experiment 1: In the first set of experiments, we estimate the quality of sparse code obtained by the proposed method. For this purpose, we randomly select 10% labeled training samples per class to build the initial training set, and the remaining samples are used for test (see Table I). Among the training set, 15 labeled samples per class are used to build the dictionary.

[4]We conduct a simple spatial regularization by majority voting based on the obtained probability for these considered methods.

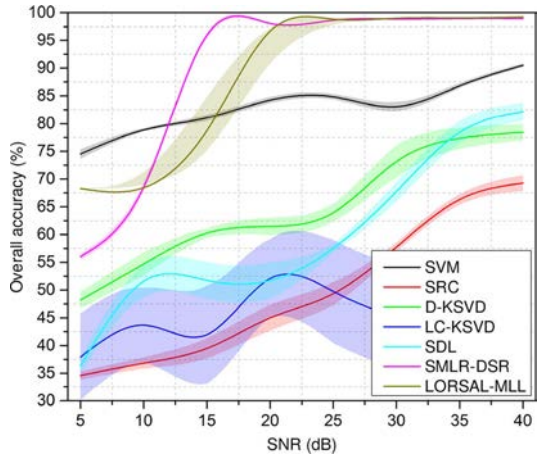


Fig. 8. Overall accuracy (OA) with standard deviation (colored area) as a function of different values of SNR for the simulated hyperspectral data set.

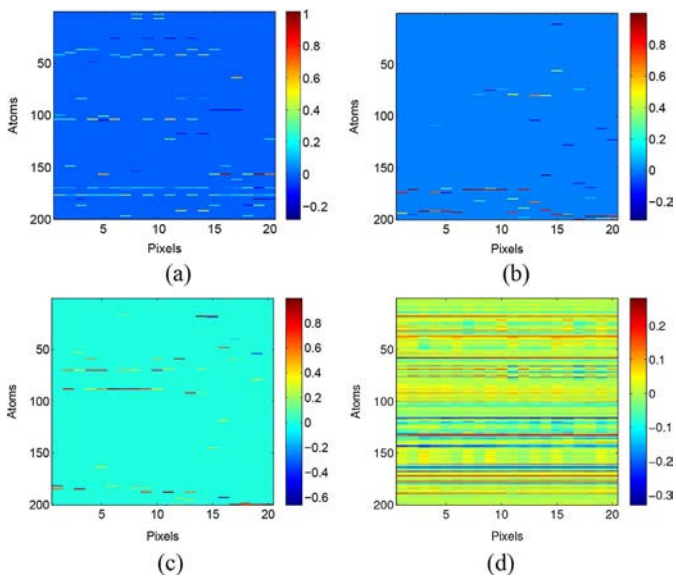


Fig. 9. Normalized sparse codes for the class *oats* obtained by different methods for the AVIRIS Indian Pines data set. (a) D-KSVD. (b) LC-KSVD. (c) SDL. (d) SMLR-DSR.

We first illustrate the distribution of sparse codes obtained by different methods. Fig. 9 depicts the normalized sparse codes taking the class *oats* as an example. As shown in the figure, the sparse codes obtained by the proposed SMLR-DSR method are more homogeneous as compared with others. This is due to the fact that SMLR-DSR adopts the TV spatial regularization. However, it seems that the *sparsity* is not very good. Fig. 10 depicts the average reconstruction error (ARE) obtained by different methods. Obviously, SMLR-DSR leads to equal values of ARE for different classes due to the fact that SMLR-DSR adopts class-oriented dictionary learning, allowing to produce more representative sparse codes relative to specific classes.

Then, we graphically illustrate the image reconstruction quality measured by signal-to-reconstruction error (SRE[5] in dB) [50] in Fig. 11, where we again take the class *oats* (atom 100) and the spectral band 57 as an example. As shown in the figure, the proposed method show some distinct characteristics: 1) the

[5]SRE $\triangleq E[\|x\|_2^2]/E[\|x-\widehat{x}\|_2^2]$ and SRE(dB) $\triangleq 10\log_{10}(\text{SRE})$. Higher value of SRE represents better reconstruction quality.

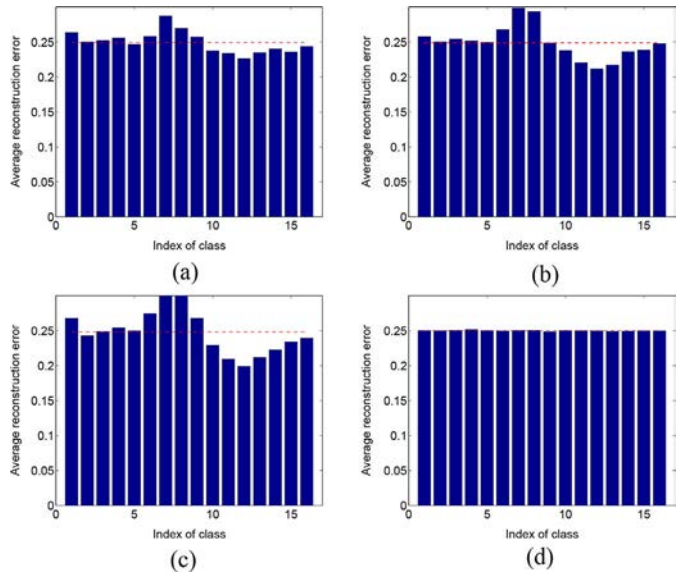


Fig. 10. Average reconstruction error obtained by different methods for the AVIRIS Indian Pines data set. (a) D-KSVD. (b) LC-KSVD. (c) SDL. (d) SMLR-DSR.

sparse code for the class *oats* (atom 100) is more representative and homogenous than others; 2) the SRE value is higher than others. These two observations reveal that the proposed method shows good performance in terms of representativeness, reconstruction ability, and spatial homogeneity, which hints at its potential use for classification purpose. It is interesting to note that, SRE is highly related to OA by comparing the results illustrated in Fig. 11 and Table II. Precisely, different DSR methods are in the same rank with respect to SRE and OA.

Experiment 2: In our second set of experiments with this scene, we evaluate the classification performance of the proposed approach using balanced training samples per class (recall that 10% samples used for training and 15 samples per class used for building the dictionary).

Table II reports the OA, AA, individual classification accuracies (%), κ statistic, standard deviation, and computational time in seconds. We mark in bold typeface the best result for each case (i.e., the highest OA, AA, κ statistic, and the lowest computational time). It is remarkable that the proposed method (SMLR-DSR) outperforms other considered DSR methods in terms of classification accuracy. For instance, SMLR-DSR obtained an OA of 97.71%, which is a remarkable result for this scene. Furthermore, SMLR-DSR still outperforms other state-of-the-art SSR methods. When compared with two recently proposed spectral-spatial classification methods such as LORSAL-MLL and MLR-GCK, SMLR-DSR also achieves superior performance even if MLR-GCK provides competitive results. It is interesting to note that, for small class such as *oats* (highlighted with gray color in the 9-th row of the table), the proposed method exhibits very good generalization performance with an OA of 100%, which well validates our former statement that CoSMLR can improve the performance of the learnt model for specific class. Another important observation is that the proposed method is not the most computationally efficient. However, the computational time of SMLR-DSR is bearable.

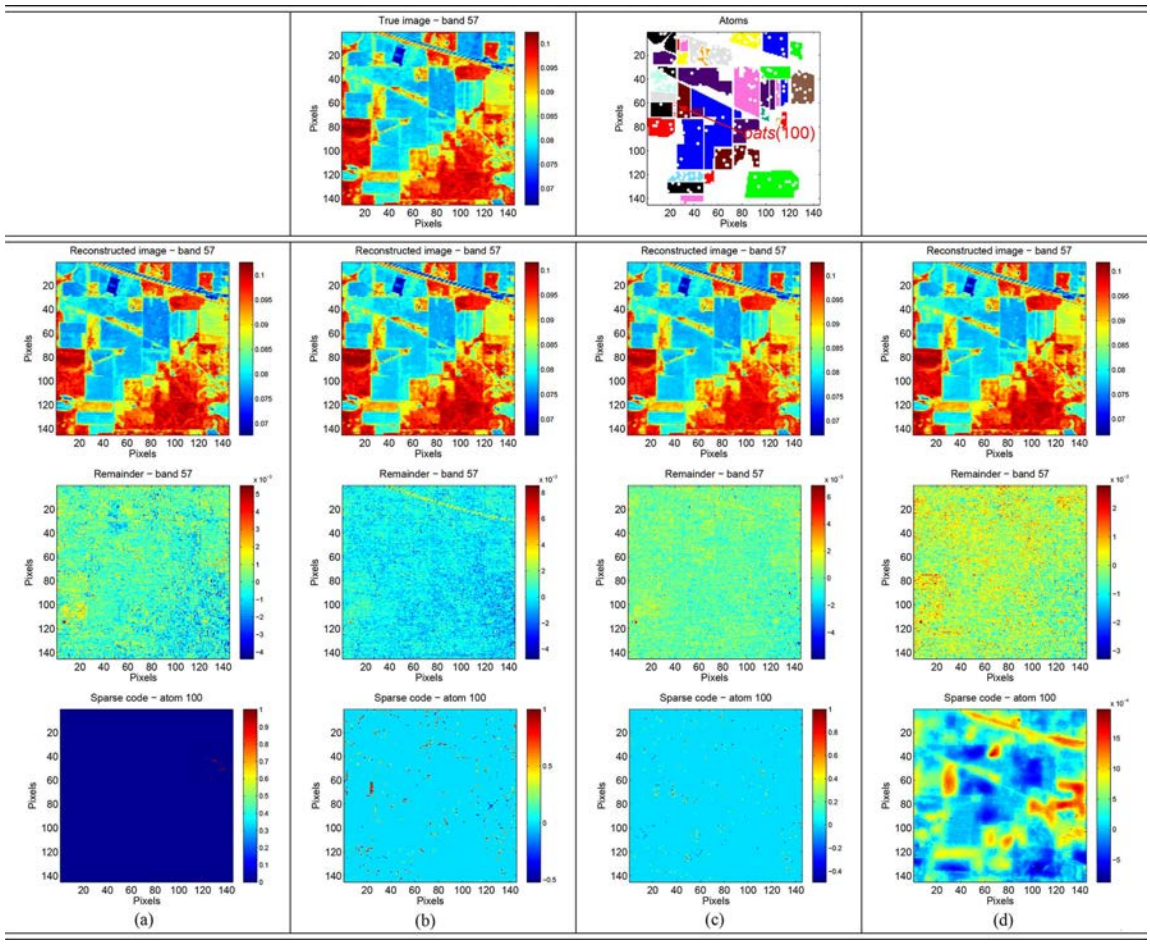


Fig. 11. Graphical illustration of the image reconstruction quality measured by the SRE(dB). The figures in the first row respectively represent the true normalized image of band 57 (left) and the chosen atoms (right), and the figures in the next three rows respectively represent the reconstructed image (second row), remainder (third row), and sparse code of atom 100 (fourth row) for different methods. SRE is reported in the parenthesis for each case. (a) D-KSVD (SRE $\approx$ 36.1 dB). (b) LC-KSVD (SRE $\approx$ 38.8 dB). (c) SDL (SRE $\approx$ 39.4 dB). (d) SMLR-DSR (SRE $\approx$ 42.0 dB).

The aforementioned observations can be explained as follows: 1) SMLR-DSR adopts SMLR with Laplacian prior to model the discriminative information which is nontrivial [33]. Nevertheless, other DSR methods adopt *square loss* with Gaussian prior; 2) the sparse representations is further mapped into the kernel space by using the RBF kernel, which greatly improves data separability; 3) SUnSAL-TV enhances the spatial homogeneity for the obtained sparse representations due to the adopted TV regularization.

The advantages obtained by adopting the proposed method with respect to other considered methods can be visually appreciated in the classification maps displayed in Fig. 12, which also reports the OAs in the parentheses for each case. These classification maps correspond to one of the ten Monte Carlo runs that were averaged in order to generate the classification scores reported in Table II.

Experiment 3: In order to illustrate the convergence and performance of the proposed algorithm under different training scenarios, in this set of experiments we analyze the impact of using unbalanced training sets on the obtained classification accuracy.

Firstly, we evaluate the impact of different number of labeled samples per class (NL_{per}) on OA with a fixed dictionary. As depicted in Fig. 13(a), the OAs increase as the training set becomes bigger, and SMLR-DSR outperforms others in different cases. Another interesting observation is that LC-KSVD and SDL converge to a stable and low value. This is due to the fact that the dictionary and classifier are sub-optimal since the optimization is motivated by feeding the atoms into the learning phase, in which context a larger dictionary is preferred in order to obtain better results. However, in this set of experiments we initialize a small dictionary on purpose.

Secondly, we evaluate the impact of different number of atoms per class (NA_{per}) on OA with a fixed training set. As shown in Fig. 13(b), the proposed method is insensitive to the dictionary size, which can be explained by the fact that only a few accurately obtained non-zero values are valid in the matrix computing process when learning the regressor. Another relevant observation is that a small dictionary still produces good results, which is in accordance with the results obtained by LC-KSVD [27] and SDL [26].

In summary, this set of experiments illustrate the good performance of the proposed method under complex training scenarios. We can also conclude that the method is insensitive to dictionary size.

Experiment 4: In this experiment, we investigated the effects of parameters λ_1 and λ_{TV} on OA. Note that, we set a small value of $\lambda_2 = 1e-5$ for learning the regressor in SMLR, which is recommended in [36]. Fig. 14 graphically illustrates the

TABLE II
OVERALL (OA), AVERAGE (AA) AND INDIVIDUAL CLASS ACCURACIES (%), KAPPA STATISTIC (κ), STANDARD DEVIATION OF TEN CONDUCTED MONTE CARLO RUNS, AND COMPUTATIONAL TIME IN SECONDS OBTAINED FOR DIFFERENT CLASSIFICATION METHODS FOR THE AVIRIS INDIAN PINES DATA SET WITH A BALANCED TRAINING SET (10% LABELED SAMPLES PER CLASS USED FOR TRAINING FOR A TOTAL OF 1036 SAMPLES AND THE REMAINING LABELED SAMPLES USED FOR TESTING)

Class	SVM[£]	SRC[#]	SSP[&]	SOMP[&]	D-KSVD[⋆]	LC-KSVD[¶]	SDL[$]	SMLR-DSR[§]	LORSAL-MLL[†]	MLR-GCK[‡]
1	70.00±10.18	74.08±7.07	89.58	85.42	60.63±8.21	48.16±12.57	67.35±8.73	88.57±4.83	67.14±13.38	**93.98±2.28**
2	83.30±1.58	40.17±5.26	95.04	94.88	49.65±3.88	50.15±4.47	62.28±4.20	**96.57±1.11**	93.65±2.29	94.61±1.15
3	75.11±4.29	53.13±4.04	92.93	94.93	48.49±1.80	48.12±6.05	65.20±4.35	**97.72±2.61**	88.12±5.41	95.32±1.01
4	72.65±7.35	61.04±6.28	85.24	91.43	32.71±8.97	36.18±5.80	46.75±11.50	91.52±7.33	**93.22±8.62**	92.64±3.01
5	92.98±2.45	80.72±5.39	92.17	89.49	72.08±8.07	69.80±12.88	78.78±9.05	**97.67±1.50**	94.14±4.42	96.90±1.19
6	96.76±0.47	85.10±5.77	98.81	98.51	82.32±3.35	86.80±1.89	85.73±3.04	**99.23±0.89**	99.14±1.04	99.00±0.56
7	82.61±8.45	85.65±5.44	73.91	91.30	55.90±18.77	49.42±21.14	64.18±18.24	**96.09±1.37**	55.65±48.28	93.41±3.58
8	98.45±0.61	93.00±2.36	99.55	99.55	93.13±1.87	92.68±1.58	94.56±2.09	99.59±0.10	99.52±0.23	**99.69±0.15**
9	68.33±20.12	60.56±9.96	0	0	58.24±7.63	47.08±10.98	49.26±30.18	**100.00±0.00**	11.67±24.77	75.80±9.94
10	79.08±3.09	69.56±4.65	**98.98**	89.44	59.21±2.36	56.72±4.83	70.38±4.39	97.24±1.56	89.61±4.31	93.21±0.54
11	86.71±1.28	46.35±4.34	97.34	97.34	64.83±1.93	66.50±1.63	74.59±1.52	**98.99±0.71**	97.12±1.17	96.68±0.35
12	84.36±2.32	55.23±6.34	86.59	83.22	43.23±6.91	47.69±4.70	72.44±4.92	94.92±2.50	**96.47±1.34**	93.64±1.13
13	99.11±0.70	98.22±1.62	99.47	100	87.79±2.33	90.90±3.12	87.90±4.54	**99.69±0.27**	99.53±0.17	99.53±0.17
14	95.91±1.27	80.27±5.01	98.88	99.14	88.61±3.21	88.70±4.19	91.23±1.89	**99.79±0.17**	97.15±2.13	99.70±0.14
15	59.91±5.34	43.86±5.14	97.37	99.12	33.60±9.00	43.30±6.17	43.86±12.19	**97.89±1.53**	88.19±4.51	96.96±0.99
16	86.82±5.73	91.88±5.36	85.88	96.47	**92.67±4.48**	88.64±5.08	84.68±7.84	70.35±6.94	76.94±14.22	87.48±4.77
AA	83.26±1.67	69.93±1.81	86.36	88.45	61.88±0.00	63.18±2.00	71.66±3.68	**95.36±0.70**	84.20±3.47	94.29±0.93
OA	86.00±0.70	61.60±1.98	94.79	95.28	64.24±0.98	65.35±1.38	74.01±1.31	**97.71±0.30**	94.32±0.91	96.29±0.27
κ	0.840±0.01	0.571±0.02	0.940	0.946	0.593±0.01	0.603±0.02	0.702±0.02	**0.974±0.00**	0.935±0.01	0.958±0.00
Time (s)	29.4±0.87	**0.7±0.03**	-	-	1.6±0.10	2.7±0.05	2.0±0.21	32.5±1.52	4.1±0.11	28.0±1.50

[£] SVM results are obtained by using SVM optimized by particle swarm optimization [51];
[#] SRC is a standard sparse representation-based classification (SRC) method [49];
[&] SSP and SOMP results are taken from [9], which respectively extend SP and OMP to address SSR within a window;
[⋆] D-KSVD is designed by extending KSVD with a discriminative term;
[¶] LC-KSVD extends from KSVD by adding a label consistent term to improve the generality of the learnt model [27];
[$] SDL adopts graph partition to build the discriminative term and shrinks from a large dictionary in the learning phase [26];
[§] SMLR-DSR is the proposed method;
[†] LORSAL-MLL algorithm implements the MAP segmentation by using SMLR via LORSAL and a multilevel logistic (MLL) prior in [36];
[‡] MLR-GCK algorithm integrates the spectral and spatial information via composite kernel to improve the classification performance [39].

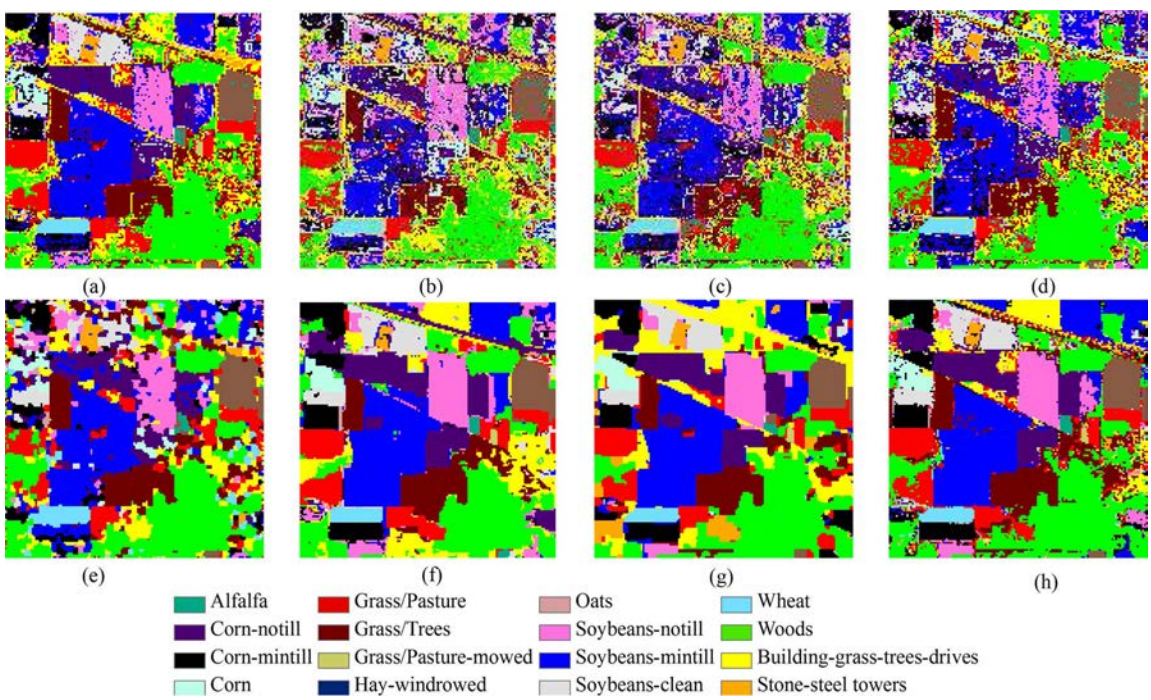


Fig. 12. Classification results obtained by different methods for the AVIRIS Indian Pines data set. The OA in each case is reported in the parentheses. (a) SVM (86.00%). (b) SRC (61.60%). (c) D-KSVD (64.24%). (d) LC-KSVD (65.35%). (e) SDL (74.01%). (f) SMLR-DSR (97.71%). (g) LORSAL-MLL (94.32%). (h) MLR-GCK (96.29%).

relationships between OA, λ_1, and λ_{TV}. As we can see from Fig. 14, when $\lambda_{TV} \geq 1e-3$ and $\lambda_1 \leq 1e-5$, our method can yield good results. This observation can be explained that larger values of λ_{TV} help to produce more homogenous sparse representations, which is better for learning effective classifier. In addition, smaller values of λ_1 lead to more sparse regressors that are considered to be more powerful in classification. To sum up, a small value of λ_1 and a higher value of λ_{TV} are recommended when applying SMLR-DSR to HSI classification.

Experiment 5: In this fifth experiment conducted with the AVIRIS Indian Pines data set, we evaluate the contribution of different components (i.e., loss function, TV regularization, dictionary learning, nonnegativity constraint) of SMLR-DSR to the improvement of classification accuracy. We compare the

TABLE III
THE CONTRIBUTION OF DIFFERENT COMPONENTS OF SMLR-DSR IN TERMS OF CLASSIFICATION ACCURACY (%) AND COMPUTATIONAL TIME (S)

	SMLR-DSR[§]	*square loss*[£]	*logistic loss*[⋆]	non-TV[&]	non-neg[¶]	non-DL[$]
AA	**95.36±0.70**	88.60±1.20	89.44±1.17	50.89±0.82	92.56±1.56	91.46±0.96
OA	**97.71±0.30**	96.44±0.31	93.13±0.71	70.52±0.88	96.49±0.46	95.27±0.62
κ	**0.974±0.00**	0.959±0.00	0.922±0.01	0.659±0.01	0.960±0.01	0.946±0.01
Time (s)	32.5±1.52	31.2±6.92	65.21±1.62	**5.3±0.26**	34.0±0.53	32.7±8.90

[§] SMLR-DSR is the proposed method that integrates sparse multinomial logistic regression (SMLR) with Laplacian prior, total variation (TV) based regularization, and dictionary learning (DL);
[£] *square loss* means the one adopting a *square loss* objective function with Gaussion prior, which is also known as the ridged regression model [52] that has been adopted in D-KSVD [24] and LC-KSVD [27];
[⋆] *logistic loss* represents the one that adopts a *logistic loss* objective function, which has been used in [25], [31]. Here, to ensure a fair comparison, we implemented *logistic loss* with Laplacian prior and conducted one-versus-rest for multiclass classification [53].
[&] non-TV is the one without TV spatial regularization;
[¶] non-neg denotes the one with nonnegativity constraint on sparse representations;
[$] non-DL refers to the one without performing dictionary learning.
Note that, each of the considered method is a variation of SMLR-DSR by specific adaption mentioned above.

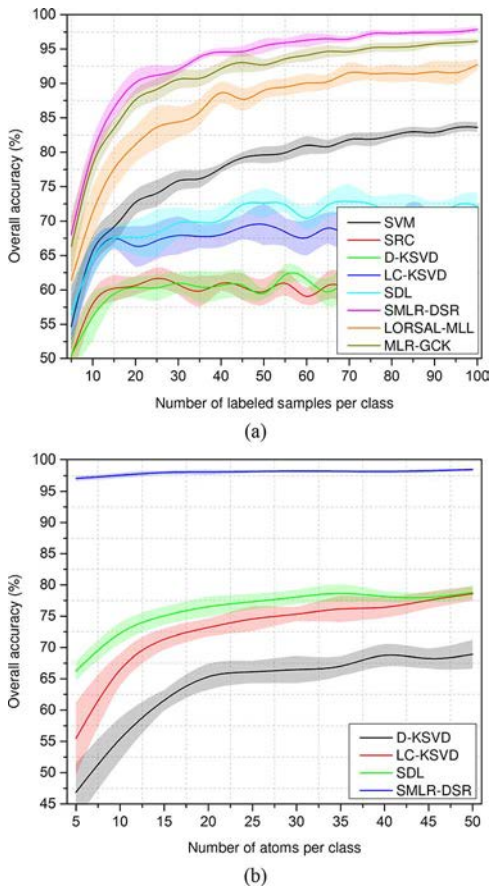


Fig. 13. Overall accuracy (OA) with standard deviation (colored area) as a function of (a) number of labeled samples per class ($NL_{\rm per}$) and (b) number of atoms per class ($NA_{\rm per}$) for the AVIRIS Indian Pines data set.

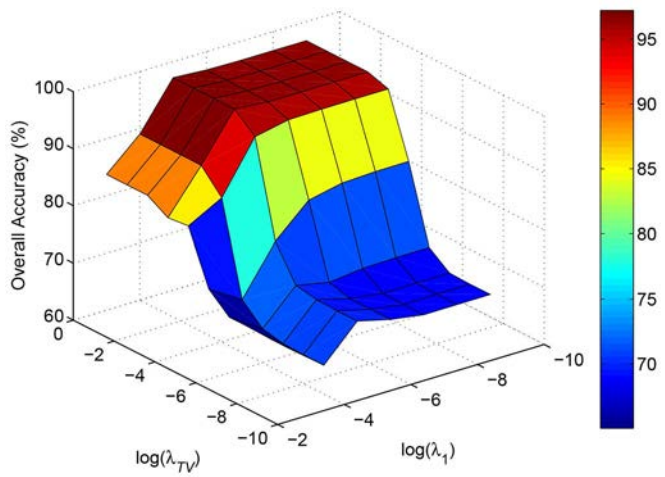


Fig. 14. Overall accuracy (OA) (%) (as a function of different values of λ_1 and λ_{TV}) obtained by the proposed method with the AVIRIS Indian Pines scene.

proposed method with others determined by some modifications based on SMLR-DSR (see Table III for more details). Table III reports the results obtained by different methods.

Some observations can be made from the results: 1) SMLR-DSR outperforms others in terms of classification accuracy; 2) *square loss* function provides competitive performance, and *logistic loss* function has higher computational complexity which also sacrifices the classification accuracy; 3) TV makes a major contribution in improving the classification accuracy; 4) nonnegativity constraint is not beneficial to produce better classification results even it guarantees the uniqueness of spare representation as stated in [54]; 5) dictionary learning is important in SMLR-DSR since it is beneficial to produce more representative sparse codes.

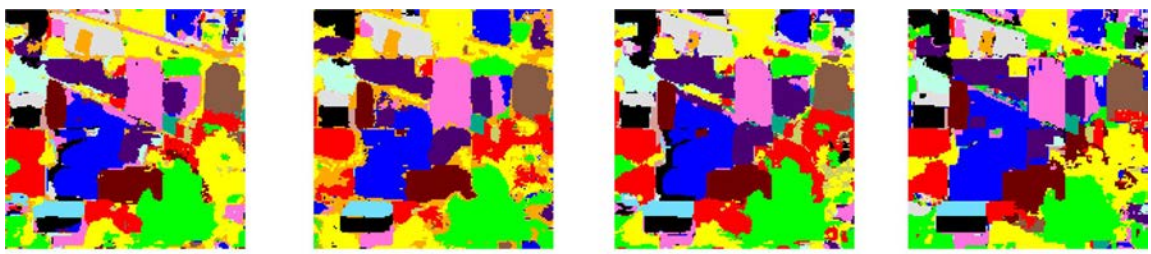
Experiment 6: In the final experiment, we analyze the performance of solving class-oriented sub-problem (sub-dictionary and sub-classifier learning) in terms of classification accuracy and computational time by comparing the results obtained by solving the whole-problem. To this end, we implement a range of methods derived from SMLR-DSR by adopting whole- or sub-problem optimization strategies. By using coupled subscript "whole, sub", we mean solving whole-problem in dictionary learning while solving class-oriented sub-problem in classifier learning. Similar definitions can be made for "whole, whole", "sub, whole", and "sub, sub".

Table IV reports the averaged classification accuracies and computational time with standard deviation (as a function of different number of labeled samples per class) obtained by different variants of SMLR-DSR. Several observations can be made from the results: 1) the proposed SMLR-DSR method obtains the highest OAs in each case, i.e., the OA reaches to 95.44% when $NL_{\rm per} = 50$; 2) class-oriented dictionary learning greatly improves the classification accuracies by comparing the results obtained by SMLR-DSR$_{\rm whole,whole}$

TABLE IV
A PERFORMANCE COMPARISON OF SOLVING CLASS-ORIENTED SUB-PROBLEMS AND WHOLE-PROBLEMS IN TERMS OF CLASSIFICATION ACCURACY (%) AND COMPUTATIONAL TIME (S)

NL_{per}	SMLR-DSR$_{\text{whole,whole}}$		SMLR-DSR$_{\text{sub,whole}}$		SMLR-DSR$_{\text{whole,sub}}$		SMLR-DSR$^{\star}_{\text{sub,sub}}$	
	OA	time	OA	time	OA	time	OA	time
5	62.65±3.58	**12.4±0.1**	60.53±3.74	13.9±0.3	63.22±2.07	13.7±0.2	**68.10±2.56**	13.5±0.1
10	71.76±1.77	**23.2±0.2**	74.83±2.30	27.0±1.8	75.45±2.01	26.1±0.7	**79.88±2.05**	25.2±0.2
15	76.54±2.15	**34.3±0.2**	80.75±3.08	**34.3±0.2**	77.00±2.16	38.6±0.7	**86.46±1.83**	37.0±0.1
20	78.98±0.96	35.7±0.5	84.94±1.93	**34.4±0.2**	80.67±1.90	38.6±0.7	**90.10±1.82**	37.3±0.1
25	83.85±2.07	34.8±0.2	87.88±1.70	**34.5±0.2**	85.25±1.21	39.4±1.0	**91.17±1.46**	37.5±0.2
30	83.67±1.23	35.0±0.4	90.89±0.75	**34.7±0.3**	84.99±1.92	39.2±1.4	**92.02±0.67**	38.7±1.8
35	82.77±1.54	**34.9±0.2**	91.48±0.75	35.0±0.2	83.78±1.55	38.7±0.2	**93.81±0.48**	38.5±0.2
40	84.18±1.25	**34.9±0.2**	91.98±0.98	35.0±0.3	83.02±1.86	38.8±0.1	**94.55±0.57**	38.6±0.3
45	84.34±2.37	**34.9±0.2**	92.49±1.28	35.4±0.3	86.31±1.21	39.0±0.2	**94.62±0.78**	39.8±1.3
50	85.49±2.23	35.4±0.3	93.81±1.17	**35.3±0.2**	87.35±1.79	39.2±0.1	**95.44±1.08**	41.0±1.9

* SMLR-DSR$_{\text{sub,sub}}$ is exactly the proposed method SMLR-DSR which solves class-oriented sub-problems when optimizing dictionary and classifier.



(a) (b) (c) (d)

Fig. 15. Classification results obtained by different variants of SMLR-DSR for the AVIRIS Indian Pines data set ($NL_{\text{per}} = 50$). The OA in each case is reported in the parentheses. (a) SMLR-DSR$_{\text{whole,whole}}$ (85.49%). (b) SMLR-DSR$_{\text{sub,whole}}$ (93.81%). (c) SMLR-DSR$_{\text{whole,sub}}$ (87.35%). (d) SMLR-DSR$_{\text{sub,sub}}$ (95.44%).

TABLE V
9 GROUND-TRUTH CLASSES IN ROSIS UNIVERSITY OF PAVIA AND THE TRAINING AND TEST SETS FOR EACH CLASS

Class		#Samples	
No	Name	Train	Test
1	Asphalt	332	6299
2	Meadows	932	17717
3	Gravel	105	1994
4	Trees	153	2911
5	Painted metal sheets	67	1278
6	Bare soil	251	4778
7	Bitumen	67	1263
8	Self-Blocking Bricks	184	3498
9	Shadows	47	900
Total		2138	40638

and SMLR-DSR$_{\text{sub,whole}}$; 3) class-oriented classifier learning obtains higher classification accuracies when comparing SMLR-DSR$_{\text{whole,whole}}$ and SMLR-DSR$_{\text{whole,sub}}$, or comparing SMLR-DSR$_{\text{sub,whole}}$ and SMLR-DSR$_{\text{sub,sub}}$.

Especially, class-oriented classifier learning cannot obviously reduce the computational time, which even increases the time cost with around 4–6s. The reason behind is that, we directly employ LORSAL to optimize SMLR in a whole-problem and report the results. Whereas, we cannot exactly report the results obtained by SMLR optimized via ADMM in whole-problem optimization without any numerically convenient trick since the time cost is huge as mentioned before. However, considering the trade-off between the greatly improved classification accuracy and the small sacrificed time cost, the proposed method exhibits good performance. For visually inspection purpose, Fig. 15 shows the associated classification maps obtained by different variants of SMLR-DSR.

E. Experiments With ROSIS University of Pavia Data Set

Experiment 1: In the first set of experiments, we evaluate the classification results obtained by the proposed method using a balanced training set with 5% labeled samples randomly selected per class, and the remaining samples are used for test (see Table V). Among the training set, 15 samples per class are used for building the dictionary. Table VI reports the obtained classification results in this case. As shown in the table, the classification accuracy obtained by the proposed algorithm is always superior to those reported for the other methods. For instance, the proposed algorithm provides the best classification accuracy with an OA of 98.69%, which is very high considering the small training set used for training. For illustrative purposes, Fig. 16 shows some of the classification maps obtained by different methods. These maps reveal clear advantages obtained by using the proposed algorithm. It's worth noting that the proposed method achieves competitive classification accuracy for the class Bare soil (highlighted with gray color in the 6-th row of the table), where great differences can be seen in the classification maps.

Experiment 2: In the second set of experiments, we again evaluate the convergence and performance of the proposed algorithm under different training scenarios. Fig. 17 illustrates the OA as a function of different number of labeled training samples per class. As shown in the figure, the OAs obtained by the proposed method increase as the training set becomes larger, which are always higher than that obtained by the other methods. Whereas, LORSAL-MLL produces competitive results in this case, and the other two DSR methods are not as effective.

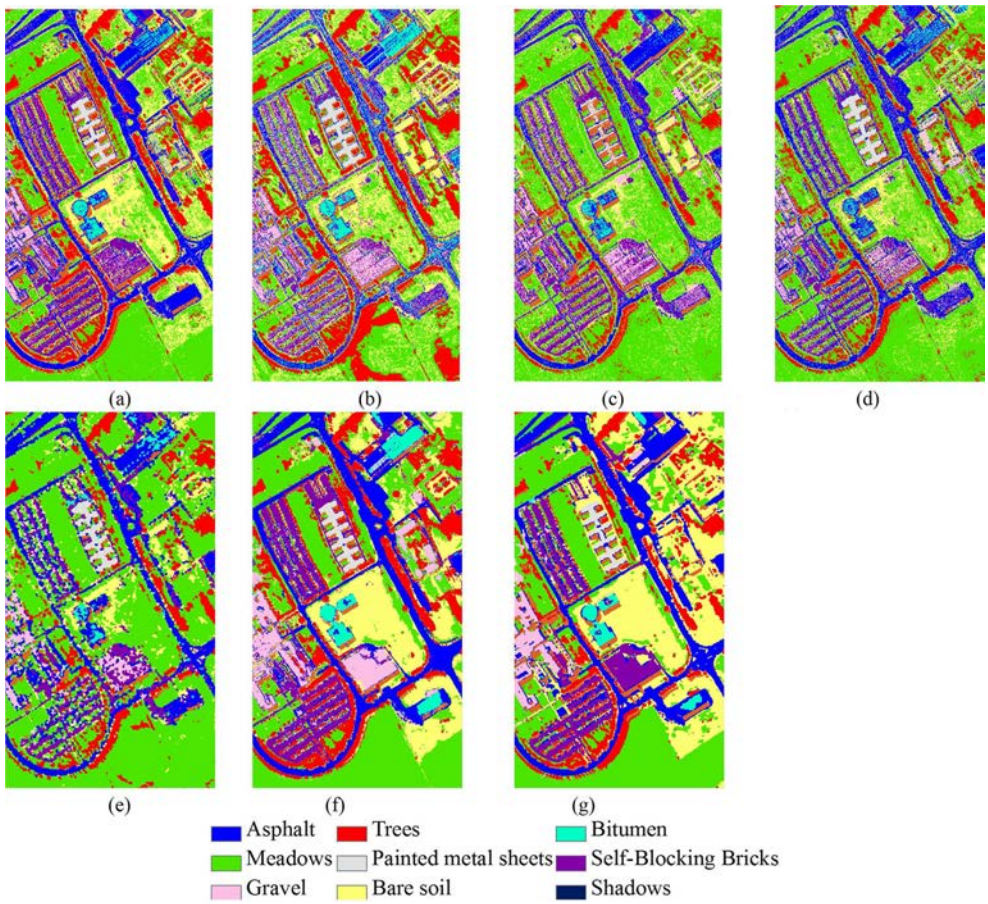


Fig. 16. Classification results obtained by different methods for the ROSIS University of Pavia data set. The OA in each case is reported in the parentheses. (a) SVM (89.90%). (b) SRC (64.27%). (c) D-KSVD (73.15%). (d) LC-KSVD (73.68%). (e) SDL (78.16%). (f) SMLR-DSR (98.69%). (g) LORSAL-MLL (97.75%).

TABLE VI
OVERALL (OA), AVERAGE (AA) AND INDIVIDUAL CLASS ACCURACIES (%), KAPPA STATISTIC (κ), STANDARD DEVIATION OF TEN CONDUCTED MONTE CARLO RUNS, AND COMPUTATIONAL TIME IN SECOND OBTAINED FOR DIFFERENT CLASSIFICATION METHODS FOR THE ROSIS UNIVERSITY OF PAVIA DATA SET WITH A BALANCED TRAINING SET (5% LABELED SAMPLES PER CLASS USED FOR TRAINING, FOR A TOTAL OF 2138 SAMPLES AND THE REMAINING LABELED SAMPLES USED FOR TESTING)

Class	SVM	SRC	SSP[&]	SOMP[&]	D-KSVD	LC-KSVD	SDL	SMLR-DSR	LORSAL-MLL
1	89.70±1.04	47.90±4.68	69.59	59.33	75.53±3.08	73.48±2.54	83.75±1.90	98.77±0.57	**98.83±0.52**
2	96.92±0.68	63.56±8.72	72.31	78.15	85.49±2.06	84.75±2.51	87.45±0.72	99.68±0.08	**99.89±0.06**
3	69.74±2.71	64.76±4.48	74.10	83.53	51.56±14.13	49.62±7.23	65.36±3.61	**95.51±1.21**	79.18±6.29
4	91.71±1.82	86.31±8.02	95.33	96.91	75.95±6.23	76.52±9.60	74.88±6.25	**98.63±0.37**	96.11±0.97
5	99.13±0.33	99.05±0.49	99.73	99.46	91.12±9.29	98.32±0.82	97.18±2.23	**99.78±0.10**	99.30±0.33
6	72.44±3.19	56.93±6.93	86.72	77.41	44.18±2.17	47.32±4.93	51.03±4.56	98.63±0.44	**99.53±0.27**
7	81.13±2.65	83.84±3.86	90.32	98.57	62.10±3.86	50.87±5.75	67.26±7.46	**98.75±0.44**	92.07±4.12
8	85.88±2.89	61.29±4.18	90.46	89.09	60.80±4.26	59.73±4.31	64.96±2.72	**96.23±1.00**	95.45±0.94
9	99.31±0.83	94.16±1.11	90.94	91.95	81.97±5.30	92.95±2.28	86.85±4.17	94.53±2.53	**99.71±0.27**
AA	87.33±0.44	73.09±1.36	85.50	86.04	68.80±2.96	70.40±1.77	75.41±1.78	**97.83±0.52**	95.56±1.15
OA	89.90±0.32	64.27±3.35	78.39	79.00	73.15±2.29	73.68±2.04	78.16±1.07	**98.69±0.30**	97.75±0.48
κ	0.865±0.00	0.553±0.03	0.724	0.728	0.638±0.03	0.647±0.03	0.696±0.02	**0.983±0.00**	0.970±0.01
Time (s)	98.1±8.07	**2.6±0.04**	-	-	9.8±0.10	10.9±0.83	22.6±5.35	477.3±70.67	33.43±1.38

[&] SSP and SOMP results are taken from [9] based on 3921 labeled samples for training.

V. CONCLUSION

In this paper, we have developed a new DSR method to learn discriminative sparse representations for hyperspectral image classification. The proposed method achieves an organic integration of SMLR with Laplacian prior and sparse representation with TV regularization. The method class-wisely learns the classifier and dictionary, which is beneficial to achieve more accurate classification results. Being a SR-based classification method, it considers TV-regularization and exploits dictionary learning in SR to produce more discriminative sparse representations, which is beneficial to learn powerful classifier by adopting SMLR. In the method, SUnSAL-TV, CoDL, and CoSMLR jointly contribute to the good performance.

The experimental results, conducted with one simulated and two widely used hyperspectral data sets, indicate that the proposed SMLR-DSR method outperforms other state-of-the-art methods in this community. Being a DSR method, it achieves

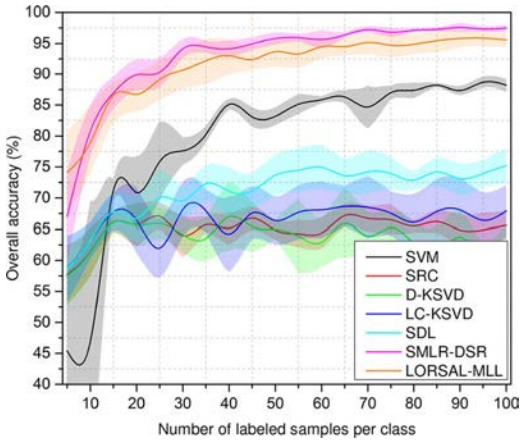


Fig. 17. Overall accuracy (OA) with standard deviation (colored area) as a function of the number of labeled samples per class (NL_{per}) obtained by different methods for the ROSIS University of Pavia data set.

better results compared to others. Being a pixel-wise method, it yields higher classification accuracy than other advanced spectral-spatial classification methods. Being a supervised method, it shows significant superiority compared to other basic classifiers. Although the results are very encouraging, further experiments with additional scenes and comparison methods should be conducted in future.

Furthermore, we also envisage several perspectives for the development of the current work:

1) First of all, more technical validations should be made in our future work to give a solid theoretical foundation for the good classification performance of the class-oriented optimization strategy.
2) Second, given the flexibility of the proposed method, in future work we will consider adopting Markov random field to achieve smoother classification results by introducing the spatial information based on probability, which is similar to the graph-cut method adopted in [36].
3) Last but not least, we are also planning to exploit structured *sparsity*-inducing norm, which is beneficial to improve the interpretability, stability, and identifiability of the learnt model as described in [55].

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for making the Airborne Visible/Infrared Imaging Spectrometer Indian Pines hyperspectral data set available to the community and Prof. P. Gamba for providing the Reflective Optics Spectrographic Imaging System data over Pavia, Italy, along with the training and test data sets. Last but not least, the authors would like to take this opportunity to thank the Associated Editor and the two Anonymous Reviewers for their comments and suggestions, which greatly helped us to improve the technical quality and presentation of our manuscript.

REFERENCES

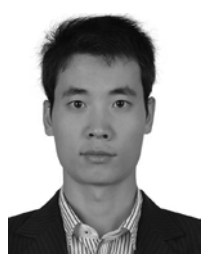
[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York, NY, USA: Wiley, 2003.
[2] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, Sep. 2009.
[3] J. Bioucas-Dias, A. Plaza, G. Camps-Valls, S. Paul, N. M. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.
[4] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.
[5] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. C. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.
[6] N. M. Nasrabadi, "Hyperspectral target detection," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 34–44, Jan. 2014.
[7] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.
[8] W. K. Ma, J. M. Bioucas-Dias, T. H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C. Y. Chi, "A signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan. 2014.
[9] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
[10] Q. S. Ul Haq, L. M. Tao, F. C. Sun, and S. Q. Yang, "A fast and robust sparse approach for hyperspectral data classification using a few labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2287–2302, Jun. 2012.
[11] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 217–231, Jan. 2013.
[12] H. Y. Zhang, J. Y. Li, Y. C. Huang, and L. P. Zhang, "A nonlocal weighted joint sparse representation classification method for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 7, no. 6, pp. 2056–2065, Jun. 2014.
[13] Z. He, Q. Wang, Y. Shen, and M. J. Sun, "Kernel sparse multitask learning for hyperspectral image classification with empirical mode decomposition and morphological wavelet-based features," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5150–5163, Aug. 2014.
[14] B. Q. Song, J. Li, M. D. Mura, P. J. Li, A. Plaza, J. M. Bioucas-Dias, J. A. Benediktsson, and J. Chanussot, "Remotely sensed image classification using sparse representations of morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5122–5136, Aug. 2014.
[15] L. Y. Fang, S. T. Li, X. D. Kang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.
[16] N. H. Ly, Q. Du, and J. E. Fowler, "Sparse graph-based discriminant analysis for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3872–3884, Jul. 2014.
[17] Y. Y. Tang, H. L. Yuan, and L. Q. Li, "Manifold-based sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7606–7618, Dec. 2014.
[18] Q. Zhang, Y. Tian, Y. Yang, and C. Pan, "Automatic spatial-spectral feature selection for hyperspectral image via discriminative sparse multimodal learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 261–279, Jan. 2015.
[19] Z. H. Xue, J. Li, L. Cheng, and P. J. Du, "Spectral-spatial classification of hyperspectral data via morphological component analysis-based image separation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 70–84, Jan. 2015.
[20] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.
[21] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[22] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[23] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2178–2191, Jun. 2009.
[24] Q. A. Zhang and B. X. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2010, pp. 2691–2698.
[25] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[26] Z. L. Jiang, G. X. Zhang, and L. S. Davis, "Submodular dictionary learning for sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2012, pp. 3418–3425.
[27] Z. L. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.
[28] A. S. Charles, B. A. Olshausen, and C. J. Rozell, "Learning sparse codes for hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 963–978, Sep. 2011.
[29] A. Castrodad, Z. M. Xing, J. B. Greer, E. Bosch, L. Carin, and G. Sapiro, "Learning discriminative sparse representations for modeling, source separation, and mapping of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4263–4281, Nov. 2011.
[30] Z. W. Wang, N. M. Nasrabadi, and T. S. Huang, "Spatial-spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4808–4822, Aug. 2014.
[31] Z. Y. Wang, N. M. Nasrabadi, and T. S. Huang, "Semisupervised hyperspectral classification using task-driven dictionary learning with Laplacian regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1161–1173, Mar. 2015.
[32] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4484–4502, Nov. 2012.
[33] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.
[34] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 1–3, pp. 293–318, 1992.
[35] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.
[36] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3947–3960, Oct. 2011.
[37] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.
[38] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.
[39] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.
[40] J. Li, X. Huang, P. Gamba, J. M. Bioucas-Dias, L. P. Zhang, J. A. Benediktsson, and A. Plaza, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015.
[41] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR)*, 2008, vol. 1–12, pp. 2415–2422.
[42] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.
[43] T. Y. Zeng and M. K. Ng, "On the total variation dictionary model," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 821–825, Mar. 2010.
[44] D. Böhning, "Multinomial logistic-regression algorithm," *Ann. Inst. Statist. Math.*, vol. 44, no. 1, pp. 197–200, Mar. 1992.
[45] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.
[46] T. Zhang and F. J. Oles, "Text categorization based on regularized linear classification methods," *Inf. Retrieval*, vol. 4, pp. 5–31, 2000.
[47] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.
[48] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[49] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[50] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, Jun. 2011.
[51] Z. H. Xue, P. J. Du, and H. J. Su, "Harmonic analysis for hyperspectral image classification integrated with PSO optimized SVM," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 7, no. 6, pp. 2131–2146, Jun. 2014.
[52] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," *SIAM J. Matrix Anal. Applicat.*, vol. 21, no. 1, pp. 185–194, 1999.
[53] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *Neuroimage*, vol. 42, no. 4, pp. 1414–1429, Oct. 1, 2008.
[54] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.
[55] X. X. Sun, Q. Qu, N. M. Nasrabadi, and T. D. Tran, "Structured priors for sparse-representation-based hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 7, pp. 1235–1239, Jul. 2014.

Peijun Du (M'07–SM'12) is a Professor of Remote Sensing at the Department of Geographic Information Sciences, Nanjing University, and the Deputy Director of the Key Laboratory for Satellite Mapping Technology and Applications of National Administration of Surveying, Mapping and Geoinformation (NASG), China. After receiving his Ph.D. degree from China University of Mining and Technology in 2001, he had been employed by the same university until he joined Nanjing University in 2011. He was a Postdoctoral Fellow at Shanghai JiaoTong University from February 2002 to March 2004, and was a Senior Visiting Scholar at the University of Nottingham, UK, and the GIPSA-Lab, Grenoble Institute of Technology, France.

His research interests focus on remote sensing image processing and pattern recognition, hyperspectral remote sensing, and applications of geospatial information technologies. He has published more than 40 articles in international peer-reviewed journals, and more than 100 papers in international conferences and Chinese journals.

Dr. Du has been the Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL) since 2009. He was the Guest Editor of 3 special issues IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATION AND REMOTE SENSING (JSTARS). He also served as the Co-chair of the Technical Committee of URBAN 2009, EORSA 2014 and IAPR-PRRS 2012, the Co-chair of the Local Organizing Committee of JURSE 2009, WHISPERS 2012 and EORSA 2012, and the member of Scientific Committee or Technical Committee of other international conferences, for example, Spatial Accuracy 2008, ACRS 2009, WHISPERS (2010–2014), URBAN (2011, 2013 and 2015), MultiTemp (2011, 2013 and 2015), ISDIF 2011, SPIE European Conference on Image and Signal Processing for Remote Sensing (2012–2014).



Zhaohui Xue received the B.S. degree in geomatics engineering from Shandong Agriculture University, Taian, China, in 2009 and the M.E. degree in remote sensing from China University of Mining & Technology, Beijing, China, in 2012. He has been honored as an outstanding graduate for B.S. and M.E. in 2009 and 2012, respectively.

He is currently pursuing the Ph.D. degree in cartography and Geographic Information System at Nanjing University, Nanjing, China. His research interests include hyperspectral image classification, time series image analysis, pattern recognition, and machine learning.

Mr. Xue was a recipient of the National Scholarship for Doctoral Graduate Students granted by the Ministry of Education of the People's Republic of China in 2014. He has been a Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Jun Li received the B.S. degree in geographic information systems from Hunan Normal University, Changsha, China, in 2004, the M.E. degree in remote sensing from Peking University, Beijing, China, in 2007, and the Ph.D. degree in electrical engineering from the Instituto de Telecomunicações, Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Lisbon, Portugal, in 2011.

From 2007 to 2011, she was a Marie Curie Research Fellow with the Departamento de Engenharia Electrotécnica e de Computadores and the Instituto de Telecomunicações, IST, Universidade Técnica de Lisboa, in the framework of the European Doctorate for Signal Processing (SIGNAL). She has also been actively involved in the Hyperspectral Imaging Network, a Marie Curie Research Training Network involving 15 partners in 12 countries and intended to foster research, training, and cooperation on hyperspectral imaging at the European level. Since 2011, she has been a Postdoctoral Researcher with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres, Spain. She has been a Reviewer of several journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, *Pattern Recognition*, *Optical Engineering*, *Journal of Applied Remote Sensing*, and *Inverse Problems and Imaging*.

Dr. Li is an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. Her research interests include hyperspectral image classification and segmentation, spectral unmixing, signal processing, and remote sensing.

Antonio Plaza (M'05–SM'07–F'15) was born in Cáceres, Spain, in 1975. He is an Associate Professor (with accreditation for Full Professor) with the Department of Technology of Computers and Communications, University of Extremadura, where he is the Head of the Hyperspectral Computing Laboratory (HyperComp). His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has been the advisor of 12 Ph.D. dissertations and more than 30 M.Sc. dissertations. He was the Coordinator of the Hyperspectral Imaging Network, a European project with total funding of 2.8 million Euro. He has authored more than 400 publications, including 140 journal papers (90 in IEEE journals), 20 book chapters, and over 240 peer-reviewed conference proceeding papers (94 in IEEE conferences). He has edited a book on High-Performance Computing in Remote Sensing for CRC Press/Taylor and Francis and guest edited 8 special issues on hyperspectral remote sensing for different journals.

Dr. Plaza is a Fellow of IEEE "for contributions to hyperspectral data processing and parallel computing of Earth observation data." He is a recipient of the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (in 2009) and a recipient of the recognition of Best Reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (in 2010), a journal for which he served as Associate Editor in 2007–2012. He is also an Associate Editor for IEEE Access, and was a member of the Editorial Board of the IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER (2011–2012) and the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE (2013). He was also a member of the steering committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is a recipient of the 2013 Best Paper Award of the JSTARS journal, and a recipient of the most highly cited paper (2005–2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He is a recipient of the Best Ph.D. Dissertation award at University of Extremadura, a recognition also received by five of his Ph.D. students. He served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) in 2011–2012, and is currently serving as President of the Spanish Chapter of IEEE GRSS (since November 2012). He has served as a proposal evaluator for the European Commission, the National Science Foundation, the European Space Agency, the Belgium Science Policy, the Israel Science Foundation, and the Spanish Ministry of Science and Innovation. He has reviewed more than 500 manuscripts for over 50 different journals. He is currently serving as the Editor-in-Chief of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING journal. Additional information: http://www.umbc.edu/rssipl/people/aplaza.