Sparse Graph-Based Discriminant Analysis for Hyperspectral Imagery

Nam Hoai Ly, *Student Member, IEEE*, Qian Du, *Senior Member, IEEE*, and James E. Fowler, *Senior Member, IEEE*

***Abstract*—Sparsity-preserving graph construction is investigated for the dimensionality reduction of hyperspectral imagery. In particular, a sparse graph-based discriminant analysis is proposed when labeled samples are available. By forcing the projection to be along the direction where a sample is clustered with within-class samples that best represented it, the discriminative power can be enhanced. The proposed method has no requirement on the number of labeled samples as in traditional linear discriminant analysis, and it can be solved by a simple generalized eigenproblem. The quality of the dimensionality reduction is evaluated by a support vector machine with a composite spatial-spectral kernel. Experimental results demonstrate that the proposed sparse graph-based discriminant analysis can yield superior classification performance with much lower dimensionality as compared to performance on the original data or on data transformed with other dimensionality-reduction approaches.**

***Index Terms*—Classification, dimensionality reduction (DR), hyperspectral imagery (HSI), sparse representation.**

I. Introduction

In HYPERSPECTRAL imagery (HSI), the dense spectral sampling of each pixel yields rich information content at the cost of high data-set dimensionality. Consequently, dimensionality reduction (DR) plays a critical role in HSI analysis, especially for the classification task when the number of available labeled training samples is limited. Pattern-classification systems often employ DR followed by a statistical classifier to learn models in the reduced-dimension feature space; those models are then used to classify test pixels. Common transform-based DR strategies include both unsupervised approaches—e.g., principal component analysis (PCA) [1] and the maximum-noise-fraction (MNF) transform [2]—as well as supervised techniques—e.g., linear discriminant analysis (LDA) [3] and local Fisher discriminant analysis (LFDA) [4]. We note that, while band selection [5]—which can also be unsupervised [6] or supervised [7]—can be considered to be another category of DR, in this paper, we limit consideration to transform-based DR.

A graph is a mathematical data representation that describes geometric structures of data. In a graph, data can be visualized as a finite collection of samples with one sample at each vertex. The weight associated with each edge connecting two vertices is often chosen to represent the similarity of the corresponding data samples, and analysis of the resulting graph can solve statistical learning problems. Recently, [8] proposed a general graph-embedding (GE) framework that describes many existing DR techniques. In this GE framework, each DR algorithm is considered to be an undirected weighted graph that embodies desired statistical or geometrical properties of a data set, coupled with scale-normalization constraints or a penalty graph that characterizes properties that the resulting DR should avoid. Construction of the graph is critical: an appropriate graph provides a high level of DR while preserving important information such as anomalies as well as manifold and multimodal structures. Common graph structures include k-nearest neighbor and ϵ-radius ball [9], both of which connect graph vertices with simple rules which are, however, highly sensitive to data-set noise and difficult to determine for real-world applications.

An alternative graph construction was proposed in [9]. Therein, concepts from the field of sparse representation were exploited—specifically, an ℓ_1-based optimization was employed to produce a graph whose edges are intended to be sparse. Additionally, the graph automatically inherited advantages of sparse representation which is increasingly being considered to be of significant benefit to the classification task (e.g., [10]–[12]). The resulting sparsity-based graph has been proposed for a number of machine-learning tasks, including DR, data clustering, and semi-supervised learning [9], [13]. We call DR based on this ℓ_1 approach "sparsity-preserving GE" (SPGE).

In this paper, we employ such a sparse graph to construct a supervised DR method which is referred to as sparse graph-based discriminant analysis (SGDA). By preserving the sparse connection in the manifold, the class-discriminative power can be significantly reinforced [14]. Moreover, this method does not need to actually evaluate the within-class scatter matrix $(\mathbf{S}_{\mathrm{w}})$ or between-class scatter matrix $(\mathbf{S}_{\mathrm{b}})$ as in the traditional Fisher's linear discriminant analysis (LDA); in particular, it does not have the requirement of a large number of training samples as is necessary to produce a full-rank $\mathbf{S}_{\mathrm{w}}$. The solution turns out to be a generalized eigenvalue problem, thereby greatly facilitating its practical implementation. To quantify DR performance, we adopt a support vector machine (SVM) [15] for classification of the transformed data. The SVM uses a composite kernel (CK) which is a composite of spectral and spatial kernels [16]. Using the resulting SVM-CK classifier, we present a battery of experimental results that demonstrate that

Manuscript received February 6, 2013; revised May 29, 2013 and July 6, 2013; accepted August 2, 2013. This material is based upon work supported by the National Science Foundation under Grant CCF-0915307.

The authors are with the Department of Electrical and Computer Engineering and the Geosystems Research Institute, Mississippi State University, Starkville, MS 39762, USA.

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/TGRS.2013.2277251

SGDA-based DR can outperform SPGE-based DR as well as other widely used transform-based DR methods.

It is worth mentioning that a sparse graph has been used previously with LDA [17]–[19]. However, our method is different in the following respects. First, the sparse representation that we use is generated from labeled samples; in particular, in a block-structured variant of SGDA that we describe later, the sparse representation arises from the use of samples within the same class, yielding a block-structured affinity matrix and resulting in significant performance improvement. Second, the sparse graph in our approach is decomposed directly instead of being added as a regularized term to an objective function, resulting in a parameter-free model with much lower computational cost.

The remainder of the discussion is organized as follows. We start with Section II which overviews the SPGE approach to DR as developed in [9]; we then describe our proposed approach in Section III. Section IV reports classification results based on real hyperspectral data sets, comparing to a variety of other state-of-the-art DR methods. Finally, several concluding remarks are made in Section V.

II. Background

Recently, [8] unified a number of DR techniques into a common GE framework involving undirected weighted graphs. Specifically, for data set $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M]$ of M vectors with $\mathbf{x}_m \in \mathbb{R}^N$, [8] imposes graph $\mathcal{G} = \{\mathbf{X}, \mathbf{W}_\mathrm{s}\}$, the "intrinsic graph," and, optionally, graph $\mathcal{G}_\mathrm{p} = \{\mathbf{X}, \mathbf{W}_\mathrm{p}\}$, the "penalty graph." In both graphs, the data set $\mathbf{X}$ in N-dimensional space is the vertex set, while $\mathbf{W}_\mathrm{s}$ and $\mathbf{W}_\mathrm{p}$ are matrices of edge weights between vertices. The framework in [8] argues that DR techniques effectively attempt to find a projection that emphasizes the similarity between vertices in the graph as expressed in the "similarity matrix" $\mathbf{W}_\mathrm{s}$. On the other hand, the "penalty matrix" $\mathbf{W}_\mathrm{p}$ captures similarity relationships that are suppressed by the DR. Employing $\mathcal{G}$ and $\mathcal{G}_\mathrm{p}$ with specific $\mathbf{W}_\mathrm{s}$ and $\mathbf{W}_\mathrm{p}$, [8] characterizes a number of popular DR techniques—including PCA [1], LDA [3], and local linear embedding (LLE) [20]—in terms of this GE framework.

More specifically, to reduce data set $\mathbf{X}$ from dimensionality N to K, DR seeks to find $N \times K$ projection matrix $\mathbf{P}$ which results in low-dimensional $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$. Assume $M \times M$ real symmetric similarity matrix $\mathbf{W}_\mathrm{s}$ has zeros on the diagonal, and let $(\mathbf{W}_\mathrm{s})_{m,m'}$ denote the entry at row m and column m'. Then, [8] finds the DR projection via an optimization aimed at preserving the intrinsic graph

$$\begin{aligned}\mathbf{P}^* &= \underset{\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P}=\mathbf{I}_{K\times K}}{\arg\min} \sum_{m,m'} \left\|\mathbf{P}^T\mathbf{x}_m - \mathbf{P}^T\mathbf{x}_{m'}\right\|^2 (\mathbf{W}_\mathrm{s})_{m,m'} \\ &= \underset{\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P}=\mathbf{I}_{K\times K}}{\arg\min} \mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{P}) \end{aligned} \tag{1}$$

where $\mathbf{L}_\mathrm{s}$ is the Laplacian matrix of the intrinsic graph such that $\mathbf{L}_\mathrm{s} = \mathbf{D}_\mathrm{s} - \mathbf{W}_\mathrm{s}$, and $\mathbf{D}_\mathrm{s}$ is a diagonal matrix with the sums of the rows of $\mathbf{W}_\mathrm{s}$ along the diagonal. If a penalty graph is used, $\mathbf{L}_\mathrm{p}$ may be the Laplacian matrix of the corresponding penalty matrix $\mathbf{W}_\mathrm{p}$; alternatively, it may reflect a simple scale-normalization constraint. In either case, [8] solves (1) as a generalized eigenvalue problem.

For example, [8] casts PCA into this GE framework, formulating PCA as an intrinsic graph with equal weights between vertices such that $(\mathbf{W}_\mathrm{s})_{m,m'} = 1/M (m \neq m')$ and scale normalization with $\mathbf{L}_\mathrm{p} = \mathbf{I}_{N\times N}$. Similarly, [8] expresses LDA as a GE with both intrinsic and penalty graphs—the intrinsic graph has a similarity matrix with class-dependent weights $(\mathbf{W}_\mathrm{s})_{m,m'} = \delta_{m,m'}/n_m$, where $\delta_{m,m'} = 1$ if the classes of $\mathbf{x}_m$ and $\mathbf{x}_{m'}$ are the same, and n_m is the number of vectors in $\mathbf{X}$ with class being the same as that of $\mathbf{x}_m$. For the penalty graph, LDA uses $(\mathbf{W}_\mathrm{p})_{m,m'} = 1/M (m \neq m')$, which is identical to the similarity matrix for PCA.

Given the generality of the GE paradigm, the key to DR by GE is thus the proper selection of the similarity and penalty matrices $\mathbf{W}_\mathrm{s}$ and $\mathbf{W}_\mathrm{p}$, respectively. Departing from simpler methods that construct $\mathbf{W}_\mathrm{s}$ and $\mathbf{W}_\mathrm{p}$ directly (as is the case for PCA and LDA above), [9] exploits recent concepts in the increasingly popular paradigm of sparse representation. That is, the graph weights are found such that each $\mathbf{x}_m$ is approximated using the other vectors in the data set; i.e.,

$$\widetilde{\mathbf{x}}_m = \sum_{m' \neq m} (\mathbf{W}_\mathrm{s})_{m,m'} \mathbf{x}_{m'}, \tag{2}$$

and we seek specifically a sparse representation such that $\mathbf{W}_\mathrm{s}$ is mostly zero within each row. As is common throughout sparse-representation literature, [9] achieves this desired sparsity via an ℓ_1 optimization that explicitly preserves sparsity in the locality relations between vertices. Once this sparse $\mathbf{W}_\mathrm{s}$ graph is determined, the desired DR projection is formulated so as to minimize the projection-domain reconstruction error due to this sparse representation, yielding an optimization similar to (1)

$$\begin{aligned}\mathbf{P}^* &= \underset{\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P}=\mathbf{I}_{K\times K}}{\arg\min} \sum_m \left\|\mathbf{P}^T\mathbf{x}_m - \sum_{m'} (\mathbf{W}_\mathrm{s})_{m,m'} \mathbf{P}^T\mathbf{x}_{m'}\right\|^2 \\ &= \underset{\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P}=\mathbf{I}_{K\times K}}{\arg\min} \mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{P}), \end{aligned} \tag{3}$$

where $\mathbf{L}_\mathrm{s} = (\mathbf{I}_{M\times M} - \mathbf{W}_\mathrm{s})^T(\mathbf{I}_{M\times M} - \mathbf{W}_\mathrm{s})$, and $\mathbf{L}_\mathrm{p} = \mathbf{I}_{M\times M}$. As with (1), (3) is solved via a generalized eigenvalue problem. We present the resulting DR algorithm as Algorithm 1, which we call SPGE.[1]

Algorithm 1 The SPGE Algorithm for DR (from [9])

1: **Input**: Data set $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M] \in \mathbb{R}^{N\times M}$, desired reduced dimensionality $K (K < N)$
2: **for** $m = 1$ **to** M **do**
3: Find sparse representation for $\mathbf{x}_m$ via ℓ_1 optimization

$$\min_{\boldsymbol{\alpha}_m} \|\boldsymbol{\alpha}_m\|_1 \quad \text{such that } \|\mathbf{X}_m\boldsymbol{\alpha}_m - \mathbf{x}_m\|_2^2 \leq \epsilon \tag{4}$$

where

$$\begin{aligned}\mathbf{X}_m &= [\mathbf{x}_1 \cdots \mathbf{x}_{m-1} \ \mathbf{x}_{m+1} \cdots \mathbf{x}_M] \in \mathbb{R}^{N\times(M-1)} \\ \boldsymbol{\alpha}_m &= [\alpha_{m,1} \cdots \alpha_{m,M-1}]^T \in \mathbb{R}^{M-1} \end{aligned} \tag{5}$$

and small tolerance $\epsilon > 0$.
4: **end for**

[1] SPGE was also proposed independently in [18].

5: For the graph $\mathcal{G} = \{\mathbf{X}, \mathbf{W}_\mathrm{s}\}$, form similarity matrix $\mathbf{W}_\mathrm{s} \in \mathbb{R}^{M\times M}$:

$$(\mathbf{W}_\mathrm{s})_{m,m'} = \begin{cases} 0, & m = m' \\ \alpha_{m,m'}, & m > m' \\ \alpha_{m,m'-1}, & m < m'. \end{cases} \tag{6}$$

6: Solve the generalized eigenvalue problem

$$\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{p}_k = \lambda_k \mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{p}_k \tag{7}$$

where $\mathbf{p}_k \in \mathbb{R}^N$ is the eigenvector corresponding to the kth smallest eigenvalue λ_k, and

$$\mathbf{L}_\mathrm{s} = (\mathbf{I}_{M\times M} - \mathbf{W}_\mathrm{s})^T(\mathbf{I}_{M\times M} - \mathbf{W}_\mathrm{s}) \tag{8}$$

$$\mathrm{L}_\mathrm{p} = \mathbf{I}_{M\times M}. \tag{9}$$

7: Assemble projection matrix

$$\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K] \in \mathbb{R}^{N\times K}. \tag{10}$$

8: **Output**: $\mathbf{Y} = \mathbf{P}^T\mathbf{X} \in \mathbb{R}^{K\times M}$

The authors of [9] claim several advantages to the SPGE approach. First, motivated by manifold learning [21], the sparse-graph framework conveys information that is valuable for the following data analysis, such as classification, and automatic preservation of such sparsity within the DR process is thus desirable. This is in accordance with the observation that a sparse representation is naturally discriminative [10], [14], [22], [23]. Second, due to an overall context being incorporated into the weight matrix, in contrast to conventional pairwise Euclidean distance, SPGE possesses a high degree of noise robustness. Finally, the number of neighbors adjacent to each vertex in the graph is adaptive to the specific vertex; this property may be useful in applications with unevenly distributed data, such as HSI with heterogeneous spatial regions.

III. Sparse Graph-Based Discriminant Analysis

A. SGDA

SPGE as described above is strictly an unsupervised method for DR. It is limited to unsupervised use as it does not factor class-label information into the determination of the DR projection $\mathbf{P}$. We now formulate a supervised version of SPGE that incorporates class-label information for supervised discriminant analysis; as the resulting algorithm inherits the sparsity-preserving characteristics of SPGE, we call it sparse graph-based discriminant analysis (SGDA). Specifically, for each sample $\mathbf{x}_m$, we use class-label information to partition the data set into its constituent classes and subsequently examine the reconstruction error in the projected domain.

More precisely, suppose that, for data set $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$, we have corresponding class labels $\mathbf{Z} = [z_1 \cdots z_M]$, where the class of $\mathbf{x}_m$ is $z_m \in \{1, 2, \ldots, p\}$, and p is the total number of classes in the data set. Define a class-selector function to be

$$S_z(\mathbf{x}_m) = \begin{cases} 1, & \text{the class of } \mathbf{x}_m \text{ is } z_m = z \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Then, assuming that the ℓ_1 optimization of (4) has been solved for $\mathbf{X}$ resulting in the similarity matrix $\mathbf{W}_\mathrm{s}$ given by (6), we define the within-class reconstruction error, in the projected domain, by extending (3) as

$$S_\mathrm{w} = \sum_{z=1}^{p}\sum_{m=1}^{M} S_z(\mathbf{x}_m) \cdot \left\| \mathbf{P}^T\mathbf{x}_m - \sum_{m'=1}^{M} S_z(\mathbf{x}_{m'})(\mathbf{W}_\mathrm{s})_{m,m'}\mathbf{P}^T\mathbf{x}_{m'} \right\|^2. \tag{12}$$

If we define a modified similarity matrix $\mathbf{W}'_\mathrm{s}$ as

$$(\mathbf{W}'_\mathrm{s})_{m,m'} = \begin{cases} (\mathbf{W}_\mathrm{s})_{m,m'}, & z_m = z_{m'} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

then it is straightforward to derive that

$$S_\mathrm{w} = \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s}\right)^T\left(\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s}\right)\mathbf{X}^T\mathbf{P}\right). \tag{14}$$

In a similar fashion, we define the between-class reconstruction error as

$$S_\mathrm{b} = \sum_{z=1}^{p}\sum_{m=1}^{M} S_z(\mathbf{x}_m) \cdot \left\| \mathbf{P}^T\mathbf{x}_m - \sum_{m'=1}^{M} [1 - S_z(\mathbf{x}_{m'})](\mathbf{W}_\mathrm{s})_{m,m'}\mathbf{P}^T\mathbf{x}_{m'} \right\|^2 \tag{15}$$

which can be expressed as

$$S_\mathrm{b} = \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\left(\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s}\right)^T\left(\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s}\right)\mathbf{X}^T\mathbf{P}\right) \tag{16}$$

where

$$(\mathbf{W}''_\mathrm{s})_{m,m'} = \begin{cases} (\mathbf{W}_\mathrm{s})_{m,m'}, & z_m \neq z_{m'} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

We then find the projection $\mathbf{P}$ in an effort to minimize the within-class reconstruction error while maximizing the between-class reconstruction error. We thus minimize the ratio of S_w to S_b; i.e.,

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} \frac{S_\mathrm{w}}{S_\mathrm{b}} = \arg\min_{\mathbf{P}} \frac{\mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{P})}{\mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P})} \tag{18}$$

where

$$\mathbf{L}_\mathrm{s} = \left(\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s}\right)^T\left(\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s}\right) \tag{19}$$

$$\mathbf{L}_\mathrm{p} = \left(\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s}\right)^T\left(\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s}\right). \tag{20}$$

It is well-known that a trace-ratio problem in the form of (18) does not have a closed-form solution (see, e.g., [24]). Consequently, such problems are typically solved approximately with a simpler determinant-ratio problem—in our case

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} \frac{|\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{P}|}{|\mathbf{P}^T\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{P}|} \tag{21}$$

which, in turn, is solved via a generalized eigenvalue problem. The complete SGDA algorithm is detailed as Algorithm 2.

Algorithm 2 The SGDA Algorithm for DR

1: **Input**: Data set $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M] \in \mathbb{R}^{N\times M}$, class labels $\mathbf{Z} = [z_1 \ \cdots \ z_M]$, desired reduced dimensionality $K (K < N)$

2: **for** $m = 1$ **to** M **do**

3: Find sparse representation for $\mathbf{x}_m$ via ℓ_1 optimization

$$\min_{\boldsymbol{\alpha}_m} \|\boldsymbol{\alpha}_m\|_1 \quad \text{such that } \|\mathbf{X}_m\boldsymbol{\alpha}_m - \mathbf{x}_m\|_2^2 \le \epsilon \tag{22}$$

where

$$\mathbf{X}_m = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_{m-1} \ \mathbf{x}_{m+1} \ \cdots \ \mathbf{x}_M] \in \mathbb{R}^{N\times(M-1)}$$
$$\boldsymbol{\alpha}_m = [\alpha_{m,1} \ \cdots \ \alpha_{m,M-1}]^T \in \mathbb{R}^{M-1} \tag{23}$$

and small tolerance $\epsilon > 0$.

4: **end for**

5: Form similarity matrices $\mathbf{W}_\mathrm{s}, \mathbf{W}'_\mathrm{s}, \mathbf{W}''_\mathrm{s} \in \mathbb{R}^{M\times M}$

$$(\mathbf{W}_\mathrm{s})_{m,m'} = \begin{cases} 0, & m = m' \\ \alpha_{m,m'}, & m > m' \\ \alpha_{m,m'-1}, & m < m' \end{cases} \tag{24}$$

$$(\mathbf{W}'_\mathrm{s})_{m,m'} = \begin{cases} (\mathbf{W}_\mathrm{s})_{m,m'}, & z_m = z_{m'} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

$$(\mathbf{W}''_\mathrm{s})_{m,m'} = \begin{cases} (\mathbf{W}_\mathrm{s})_{m,m'}, & z_m \ne z_{m'} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

6: Solve the generalized eigenvalue problem

$$\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{p}_k = \lambda_k\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{p}_k, \tag{27}$$

where $\mathbf{p}_k \in \mathbb{R}^N$ is the eigenvector corresponding to the kth smallest eigenvalue λ_k, and

$$\mathbf{L}_\mathrm{s} = (\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s})^T(\mathbf{I}_{M\times M} - \mathbf{W}'_\mathrm{s}) \tag{28}$$
$$\mathbf{L}_\mathrm{p} = (\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s})^T(\mathbf{I}_{M\times M} - \mathbf{W}''_\mathrm{s}). \tag{29}$$

7: Assemble projection matrix

$$\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K] \in \mathbb{R}^{N\times K}. \tag{30}$$

8: **Output**: $\mathbf{Y} = \mathbf{P}^T\mathbf{X} \in \mathbb{R}^{K\times M}$

B. SGDA With Block-Structured Similarity Matrix

In SGDA as described in the previous session, label information is not used directly when estimating the $\mathbf{W}_\mathrm{s}$ matrix since all samples are employed when finding the sparse representation of a single sample. However, it is straightforward to devise a variant of SGDA that finds the sparse representation of a sample using only the labeled samples in the same class. Assume that the samples are ordered, as is common, in terms of their class labels. In this case, the resulting $\mathbf{W}_\mathrm{s}$ matrix has a block-diagonal structure, i.e.,

$$\mathbf{W}_\mathrm{s} = \begin{bmatrix} \mathbf{W}^{(1)} & 0 & \cdots & 0 \\ 0 & \mathbf{W}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{W}^{(p)} \end{bmatrix} \tag{31}$$

where $\mathbf{W}^{(i)}$ is the sparse representation matrix of size $M_i \times M_i$ for samples in the ith class C_i using the M_i samples belonging to C_i only. Note that the diagonal of each $\mathbf{W}^{(i)}$ is zero to avoid self-similarity. We call the resulting algorithm block SGDA (BSGDA) which is shown in Algorithm 3.

Algorithm 3 The BSGDA Algorithm for DR

1: **Input**: Data set $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M] \in \mathbb{R}^{N\times M}$, class labels $\mathbf{Z} = [z_1 \ \cdots \ z_M]$, desired reduced dimensionality $K (K < N)$. Assume there are p classes, and the number of samples in the ith class C_i is M_i, i.e., $\sum_{i=1}^p M_i = M$. The samples are ordered in terms of their class labels, i.e., $\{z_i\}|_{i=1}^{M_1} = C_1, \{z_i\}|_{I=M_1+1}^{M_1+M_2} = C_2, \cdots, \{z_i\}|_{i=M-M_{p-1}+1}^{M} = C_p$. $\mathbf{W}_\mathrm{s}$ has a block structure, i.e., coefficients $\{w_{ij}\}|_{i=1:M_1}^{j=1:M_1}$, $\{w_{ij}\}|_{i=M_1+1:M_2}^{j=M_1+1:M_2}, \cdots, \{w_{ij}\}|_{i=M-M_{p-1}+1:M}^{j=M-M_{p-1}+1:M}$ may have non-zero values; all other elements are zero.

2: **for** $i = 1$ **to** p **do**

3: **for** $j = 1$ **to** M_i **do**

4: Find sparse representation for $\mathbf{x}_j$ via ℓ_1 optimization:

$$\min_{\boldsymbol{\alpha}_j} \|\boldsymbol{\alpha}_j\|_1 \text{ such that } \left\|\mathbf{X}_j^{C_i}\boldsymbol{\alpha}_j - \mathbf{x}_j^{C_i}\right\|_2^2 \le \epsilon \tag{32}$$

where $\mathbf{X}_j^{C_i} = \{\mathbf{x}_m \in C_i, \mathbf{x}_m \ne \mathbf{x}_j\}$, and small tolerance $\epsilon > 0$. The zero-padding version is $\boldsymbol{\alpha}'_j = [0 \ \cdots \ 0 \ \boldsymbol{\alpha}_j^T \ 0 \ \cdots \ 0]^T$, where the numbers of zeros being added before and after $\boldsymbol{\alpha}_j$ being $\sum_{k=1}^{i-1} M_k$ and $\sum_{k=i+1}^{p} M_k$, respectively.

5: **end for**

6: **end for**

7: For the graph $\mathcal{G} = \{\mathbf{X}, \mathbf{W}_\mathrm{s}\}$, form similarity matrices

$$(\mathbf{W}_\mathrm{s})_{m,m'} = \begin{cases} 0, & m = m' \\ \alpha_{m,m'}, & m > m' \\ \alpha_{m,m'-1}, & m < m. \end{cases} \tag{33}$$

8: Solve the generalized eigenvalue problem

$$\mathbf{X}\mathbf{L}_\mathrm{s}\mathbf{X}^T\mathbf{p}_k = \lambda_k\mathbf{X}\mathbf{L}_\mathrm{p}\mathbf{X}^T\mathbf{p}_k \tag{34}$$

to find the kth smallest eigenvalue λ_k, where $\mathbf{L}_\mathrm{s} = (\mathbf{I} - \mathbf{W}_\mathrm{s})^T(\mathbf{I} - \mathbf{W}_\mathrm{s})$ and $\mathbf{L}_\mathrm{p} = \mathbf{I}$.

9: Assemble projection matrix

$$\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K] \in \mathbb{R}^{N\times K}. \tag{35}$$

10: **Output**: $\mathbf{Y} = \mathbf{P}^T\mathbf{X} \in \mathbb{R}^{K\times M}$

In BSGDA, we intentionally disconnect between-class samples in the sparse graph. This operation is critical, particularly when the samples in the dictionary possess some coherence, resulting in an unstable sparse solution for $\mathbf{W}_\mathrm{s}$. Since the objective is to preserve the sparse representation from the samples in the same class in the projected space, and any partial representation from other classes is impossible under this setting, within-class samples are further clustered in the projected space. Consequently, class separability can be enhanced. Note that, in this case, even when within-class samples are coherent, class separability is not degraded because a sample is permitted to be represented by only the samples in the same class,

TABLE I
GROUND-TRUTH CLASSES AND CORRESPONDING TRAINING- AND TESTING-SET SIZES FOR INDIAN PINES

	Classes		
No.	*Name*	*train*	*test*
1	Alfalfa	5	41
2	Corn-notill	143	1285
3	Corn-min	83	747
4	Corn	24	213
5	Grass/Pasture	48	435
6	Grass/Trees	73	657
7	Grass/Pasture-mowed	3	25
8	Hay-windrowed	48	430
9	Oats	2	18
10	Soybean-notill	97	875
11	Soybean-min	246	2209
12	Soybean-clean	59	534
13	Wheats	21	184
14	Woods	127	1140
15	Building-Grass-Trees-Drives	39	347
16	Stone-steel Towers	9	84
	Total	1027	9222

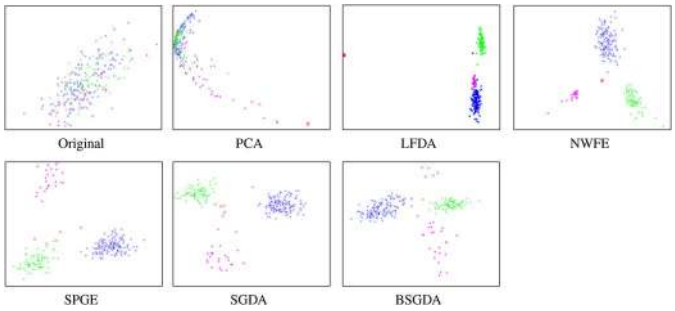

Fig. 1. DR to 2-D space for a subscene of the Indian Pines data set.

meaning that it will be assigned the same class label anyway. In subsequent experiments, we will show that BSGDA can even outperform SGDA.

C. Evaluation With SVM-CK

The performance of DR can be evaluated with classification applied to the data resulting from DR, and we use SVMs [15] due to their popularity. Since spectral signatures of spatially adjacent hyperspectral pixels are highly correlated, neighboring pixels often belong to the same class. Consequently, the coupling of spatial context to the spectral signature can significantly improve classification accuracy [25]. A well-known approach for incorporating such spectral-spatial information into classification is via the use of a CK within SVM [16] (called SVM-CK). As in [16], we define the spatial feature vector for hyperspectral pixel $\mathbf{x}_m$ to be $\bar{\mathbf{x}}_m$, the average vector over a spatial window of size 5×5 surrounding $\mathbf{x}_m$. The spectral feature vector is simply $\mathbf{x}_m$ itself. Then, using weight parameter μ, the weighted-summation CK is

$$K(\mathbf{x}_m, \mathbf{x}_{m'}) = \mu K_{\mathrm{s}}(\bar{\mathbf{x}}_m, \bar{\mathbf{x}}_{m'}) + (1-\mu) K_{\omega}(\mathbf{x}_m, \mathbf{x}_{m'}). \quad (36)$$

For spatial kernel K_{s}, we use a polynomial kernel while a radial-basis-function (RBF) kernel is used for the spectral kernel K_{ω}. Note that this is opposite of [16] which originally used a polynomial kernel spectrally and an RBF kernel spatially—we have found a 2 to 3% better overall accuracy in experimental observations for the CK that we propose here. We

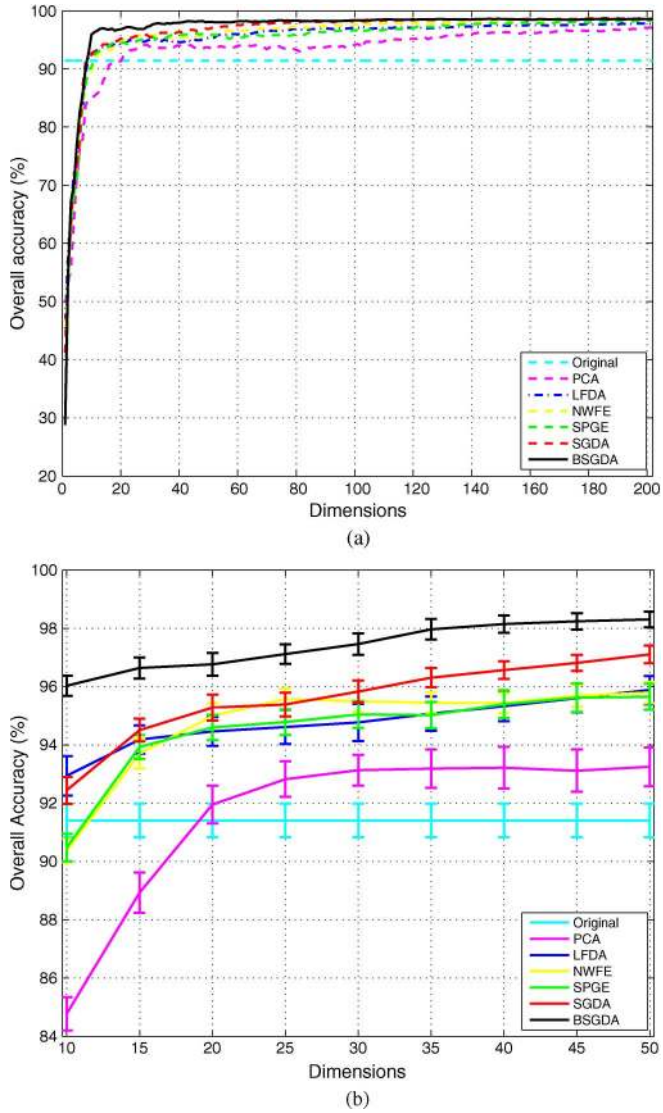

Fig. 2. SVM-CK classification performance (overall accuracy) under DR for Indian Pines. Error bars indicate standard deviation. (a) Mean accuracy. (b) Mean accuracy with standard deviation.

Fig. 3. Weight matrices for SPGE ($\mathbf{W}_{\mathrm{s}}$), SGDA ($\mathbf{W}'_{\mathrm{s}}$), and BSGDA ($\mathbf{W}_{\mathrm{s}}$).

note that, in this work, SVM-CK is applied to the data after DR; i.e., $\mathbf{x}_m$ and $\bar{\mathbf{x}}_m$ in (36) are of reduced dimensionality.

IV. EXPERIMENTAL RESULTS

We now demonstrate the effectiveness of the proposed SGDA and BSGDA algorithms on three hyperspectral images. The classifier parameters (RBF kernel parameter γ, SVM regularization parameter ς, and CK weight μ) are obtained via cross-validation. The one-against-one strategy is employed for

TABLE II
SVM-CK CLASSIFICATION ACCURACY (%) FOR INDIAN PINES ($K = 30$)

Class	Original (SVM)	Original (SVM-CK)	PCA	LFDA	NWFE	SPGE	SGDA	BSGDA
1	67.5 ± 13.3	72.4 ± 16.9	82.9 ± 10.8	79.4 ± 12.5	92.4 ± 7.1	86.8 ± 6.9	92.5 ± 4.9	$\mathbf{95.6 \pm 4.3}$
2	81.9 ± 2.0	91.6 ± 1.6	93.9 ± 1.1	94.9 ± 1.1	95.3 ± 1.3	94.5 ± 1.3	95.3 ± 0.8	$\mathbf{97.3 \pm 0.9}$
3	70.9 ± 2.7	90.8 ± 2.7	91.1 ± 2.3	93.5 ± 2.0	94.5 ± 2.1	93.3 ± 2.1	96.6 ± 1.5	$\mathbf{98.1 \pm 1.1}$
4	68.1 ± 7.7	80.8 ± 5.9	83.3 ± 5.8	83.8 ± 7.3	94.3 ± 2.9	85.5 ± 4.6	96.6 ± 1.8	$\mathbf{97.7 \pm 2.2}$
5	91.3 ± 2.3	93.2 ± 2.1	94.8 ± 1.5	96.8 ± 1.7	96.8 ± 1.5	96.4 ± 1.6	97.1 ± 1.6	$\mathbf{97.9 \pm 1.8}$
6	95.1 ± 1.8	97.9 ± 1.0	99.0 ± 0.6	$\mathbf{99.8 \pm 0.2}$	99.3 ± 0.4	$\mathbf{99.8 \pm 0.2}$	99.0 ± 0.5	99.2 ± 0.4
7	68.3 ± 17.9	84.6 ± 16.4	77.3 ± 17.2	79.8 ± 19.0	90.5 ± 7.6	94.8 ± 5.1	96.1 ± 6.8	$\mathbf{99.2 \pm 2.1}$
8	98.2 ± 1.0	98.4 ± 1.2	99.8 ± 0.3	$\mathbf{100.0 \pm 0.0}$	99.9 ± 0.1	99.9 ± 0.1	100.0 ± 0.1	$\mathbf{100.0 \pm 0.0}$
9	35.8 ± 18.6	56.8 ± 20.5	55.0 ± 22.4	29.4 ± 18.9	42.1 ± 21.4	63.3 ± 19.0	47.5 ± 22.4	$\mathbf{92.1 \pm 7.8}$
10	73.6 ± 3.3	85.5 ± 2.2	86.8 ± 2.7	88.1 ± 2.5	90.1 ± 2.4	89.5 ± 2.3	89.6 ± 2.3	$\mathbf{94.1 \pm 1.7}$
11	82.4 ± 1.9	91.5 ± 1.5	93.1 ± 1.5	95.3 ± 1.3	94.2 ± 1.2	94.8 ± 1.0	94.7 ± 1.5	$\mathbf{96.5 \pm 1.2}$
12	78.9 ± 3.3	89.7 ± 2.3	91.2 ± 2.7	95.7 ± 2.3	97.2 ± 1.1	95.0 ± 1.6	95.8 ± 1.6	97.2 ± 1.5
13	97.9 ± 1.3	98.7 ± 1.1	99.3 ± 0.4	$\mathbf{99.7 \pm 0.3}$	$\mathbf{99.7 \pm 0.3}$	99.2 ± 0.7	99.2 ± 0.8	99.1 ± 1.0
14	94.0 ± 1.9	95.6 ± 1.6	97.1 ± 1.5	98.1 ± 0.8	98.7 ± 0.7	98.7 ± 0.7	99.0 ± 0.6	$\mathbf{99.1 \pm 0.7}$
15	60.7 ± 5.3	78.0 ± 4.4	85.6 ± 5.6	89.9 ± 5.6	94.2 ± 3.0	93.3 ± 3.2	95.8 ± 1.9	$\mathbf{98.3 \pm 1.7}$
16	90.1 ± 4.7	96.7 ± 4.5	96.6 ± 3.6	97.4 ± 3.6	97.2 ± 3.2	98.4 ± 2.7	$\mathbf{99.0 \pm 0.9}$	98.3 ± 2.2
Overall	82.9 ± 0.5	91.4 ± 0.6	93.1 ± 0.5	94.8 ± 0.6	95.5 ± 0.3	95.0 ± 0.5	95.8 ± 0.4	$\mathbf{97.5 \pm 0.4}$
Average	78.4 ± 2.2	87.6 ± 2.2	89.2 ± 2.2	88.8 ± 2.2	92.3 ± 1.3	92.7 ± 1.3	93.4 ± 1.6	$\mathbf{97.5 \pm 0.6}$
κ	80.4 ± 0.6	90.2 ± 0.7	92.2 ± 0.6	94.0 ± 0.7	94.9 ± 0.4	94.4 ± 0.5	95.2 ± 0.4	$\mathbf{97.1 \pm 0.4}$

classification using SVM-CK. We use the popular libSVM toolkit[2] for SVM, and ℓ_1-optimization problems are solved using SPGL1[3] [26]. For each image, the number of training samples is randomly selected according to [27], and we run the simulation 100 times and report the average as well as the standard deviation. Throughout, we report results for a variety of state-of-the-art DR methods, comparing SPGE, SGDA, and BSGDA to PCA, LFDA, as well as nonparametric weighted feature extraction (NWFE) [28].

A. Discriminant Capability From a 2-D Visualizaton

To demonstrate the efficacy of SGDA and BSGDA, we perform a simple visualization experiment on a small spatial region from the AVIRIS Indian Pines data set (see Section IV-B for a complete description of the data set) which includes the alfalfa, corn-notill, corn-min, and corn classes (the first four classes in Table I); 202 features; and 255 samples as the training set. Fig. 1 illustrates scatter plots of the original data (the first two dimensions), wherein classes are highly mixed, as well as the data after DR to two dimensions by PCA, LFDA, NWFE, SPGE, SGDA, and BSGDA.

From Fig. 1, we draw several conclusions. First, SPGE, SGDA, and BSGDA all yield DR superior to that of PCA since the Euclidean pairwise distance employed by PCA fails to identify the real local structure. Moreover, with SGDA and BSGDA, we see that the four classes can be perfectly separated, which can be explained from the imposed sparsity-preserving and discriminative information provided by each sample in terms of the whole training data set. SGDA and BSGDA therefore preserve the local structure (i.e., multiple modes) better than the other methods. Second, as compared to LFDA, SGDA and BSGDA are nonparametric and do not require the setting of global parameters as needed by LFDA for the computation of the affinity and local scaling. In particular, LFDA may, in fact, fail in the case of non-evenly distributed data; this can be seen in this experiment in the case of the alfalfa class which is represented by only five samples. Thus, in the case that the data is severely non-Gaussian—a prevalent occurrence in real-world HSI applications—it is anticipated that SGDA and BSGDA will outperform other techniques for DR projection. We note also that, in this demonstration, NWFE also produces perfect separation of the four classes.

B. AVIRIS Data Set: Indian Pines

We start our experimental evaluation with the popular AVIRIS Indian Pines[4] data set. The AVIRIS sensor generates 220 bands across the 0.2- to 2.4-μm spectral range; however, we remove 18 water-absorption bands, resulting in an original-image dimensionality of 202. The Indian Pines image has a spatial coverage of 145×145 pixels at a spatial resolution of 20 m. It contains 16 ground-truth classes which are tabulated in Table I. For each of the 16 classes, we randomly choose 10% of the labeled samples for training and the remaining 90% for testing for each class; Table I gives the resulting number of training and testing samples for each class.

The parameters of SVM and SVM-CK, $(\varsigma, \gamma, \mu) = (256, 0.3536, 0.7)$ for the original data, are obtained by five-fold cross-validation. The parameters for the data after DR are only slightly different from these values. The effect of DR on the overall classification accuracy is shown in Fig. 2. In this figure, SVM-CK in the original dimensionality is used as a baseline, while the three sparsity-based algorithms—SPGE, SGDA, and BSGDA—are compared to PCA, LFDA, and NWFE. The results show that BSGDA gives the best overall accuracy for low dimensionality. Fig. 3 illustrates the weight matrices for SPGE, SGDA, and BSGDA, where it can be seen that BSGDA matrix is sparser than the other two.

From Fig. 2, we choose a reduced dimensionality of $K = 30$ which well represents the classification performance of each algorithm. We tabulate the classification accuracy (mean and standard deviation) for each class, overall accuracy, average accuracy, and κ coefficient in Table II. The overall accuracy

[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3]http://www.cs.ubc.ca/labs/scl/spgl1
[4]ftp://ftp.ecn.purdue.edu/biehl/MultiSpec

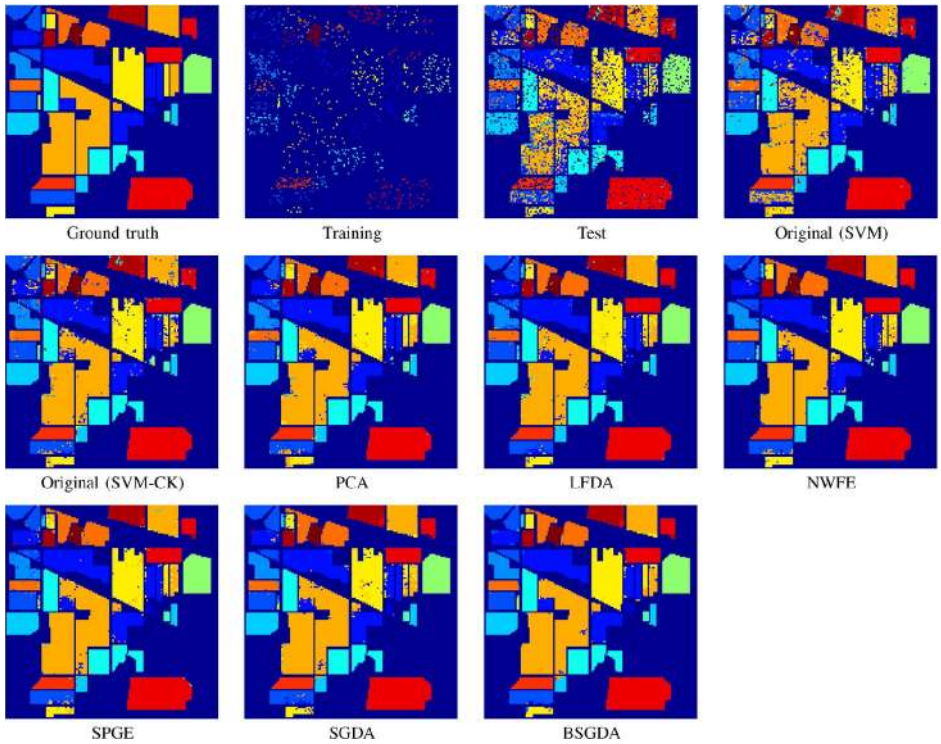

Fig. 4. Ground-truth and SVM-CK classification maps for Indian Pines ($K = 30$).

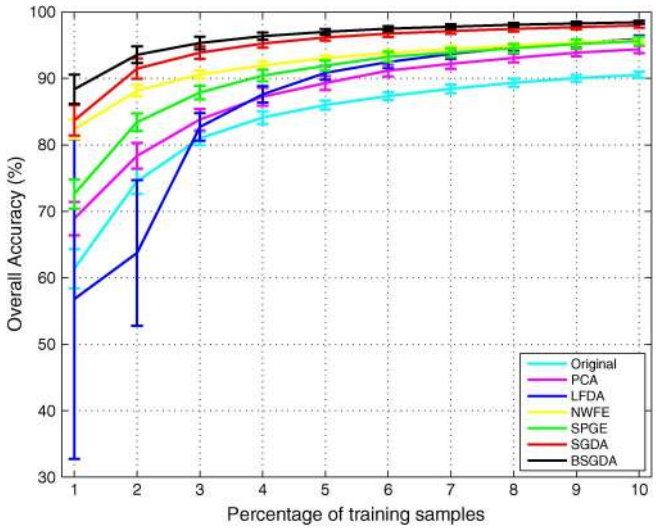

Fig. 5. SVM-CK classification performance (overall accuracy) as the number of training samples varies for Indian Pines. Error bars indicate standard deviation.

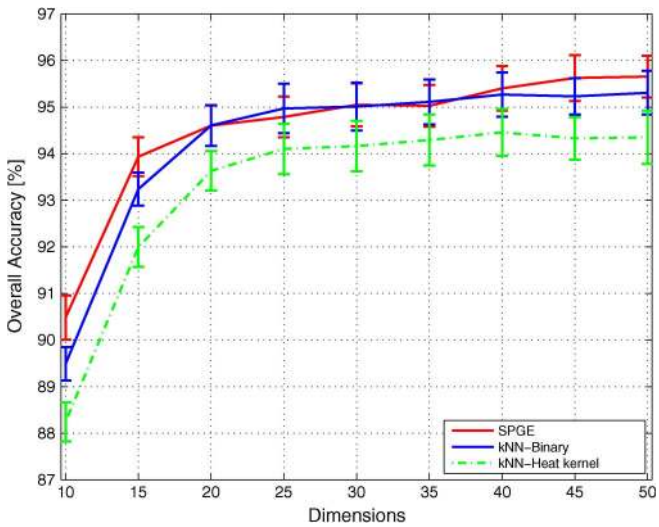

Fig. 6. Comparison with k-NN graphs for Indian Pines. Error bars indicate standard deviation.

is computed as the ratio of correctly classified test samples to the total number of test samples; the average accuracy is the mean of the 16 individual class accuracies; and the Cohen-κ coefficient is computed by weighting the measured accuracies which show a robust measure of the degree of agreement. In most cases, the proposed BSGDA outperforms the original SVM-CK, as well as other DR methods coupled with SVM-CK. Overall, BSGDA provides the best performance, especially in the extreme case (e.g., classes 1, 7, and 9 which have only 5, 3, and 2 training samples, respectively). Due to the high cost of training data, such performance at low numbers of training data is important in many applications.

One might expect that, as the reduced dimensionality K approaches the full data set dimensionality N (e.g., Fig. 2(a) as $K \to 202$), the classification performance of each of the DR methods would become identical and converge to the performance of the original data set. However, as we see in Fig. 2(a), this is not the case; in fact, when $K = N$, the performance of the DR-based methods is even better than that of the original data set, despite the fact that no DR actually takes place.[5] Additionally, the different DR-based methods yield somewhat different performance even for $K = N$. We note that this is a

[5]For example, when $K = N$, PCA becomes a unitary rotation instead of a dimension-reducing projection.

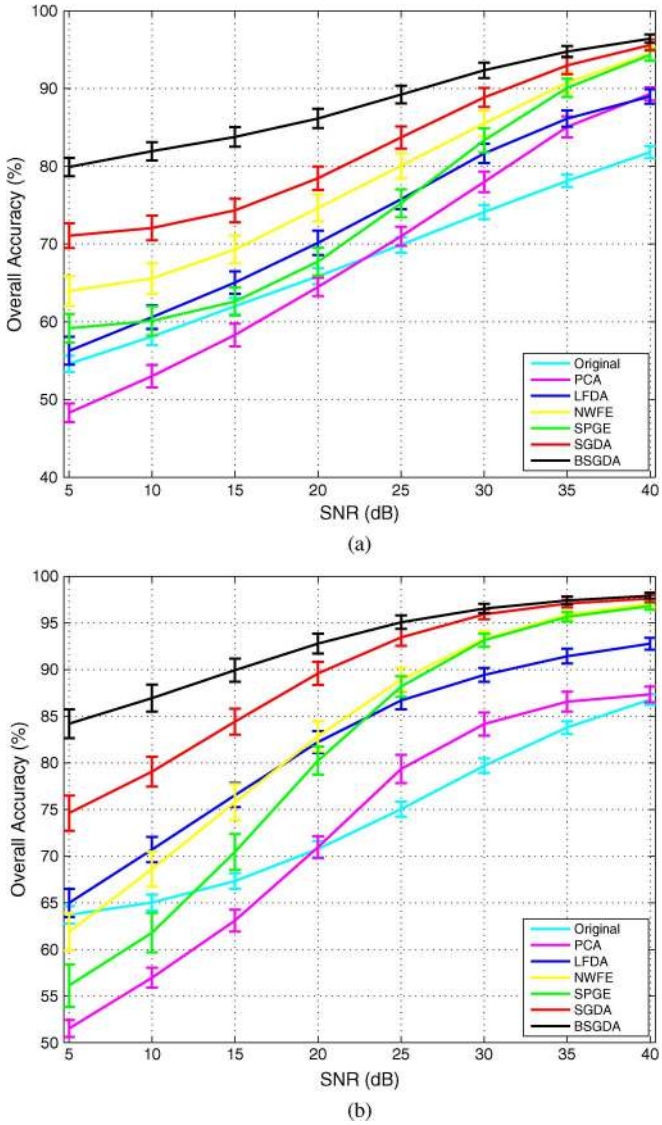

Fig. 7. SVM-CK classification performance (overall accuracy) under added noise for Indian Pines. Error bars indicate standard deviation. (a) Gaussian. (b) Uniform.

phenomenon that arises when SVM-CK is the classifier due to the interaction between its incorporation of spatial features and the transformation represented by the DR method.

Classification maps on labeled pixels obtained from the various algorithms are shown in Fig. 4. One can see that, by incorporating contextual information, BSGDA provides a much smoother classification map than do the other methods.

Fig. 5 demonstrates how the number of training samples affects the classification performance for the various algorithms. For this figure, the various parameters are fixed to be the same as used for Fig. 4. For each test, we randomly choose 1 to 10% of the labeled data in each class as the training samples and the remaining samples for testing. The overall accuracy is averaged over 100 simulations at each training rate so as to avoid any bias induced by random sampling. We observe that the overall accuracy of BSGDA monotonically increases as the training rate increases in all cases.

Fig. 6 compares with two k-nearest-neighbor (k-NN) graphs with weights assigned by both binary values as well as a

TABLE III
GROUND-TRUTH CLASSES AND CORRESPONDING TRAINING- AND TESTING-SET SIZES FOR SALINAS

Classes			
No.	*Name*	*train*	*test*
1	Broccoli-green-weeds-1	100	1909
2	Broccoli-green-weeds-2	186	3540
3	Fallow	99	1877
4	Fallow-rough-plow	70	1324
5	Fallow-smooth	134	2544
6	Stubble	198	3761
7	Celery	179	3400
8	Grapes-untrained	564	10707
9	Sell-vineyard-develop	310	5893
10	Corn-senesced-green-weeds	164	3114
11	Lettuce-romain-4-weeks	53	1015
12	Lettuce-romain-5-weeks	96	1831
13	Lettuce-romain-6-weeks	46	870
14	Lettuce-romain-7-weeks	54	1016
15	Vineyard-untrained	363	6905
16	Vineyard-vertical-trellis	90	1717
	Total	2706	51423

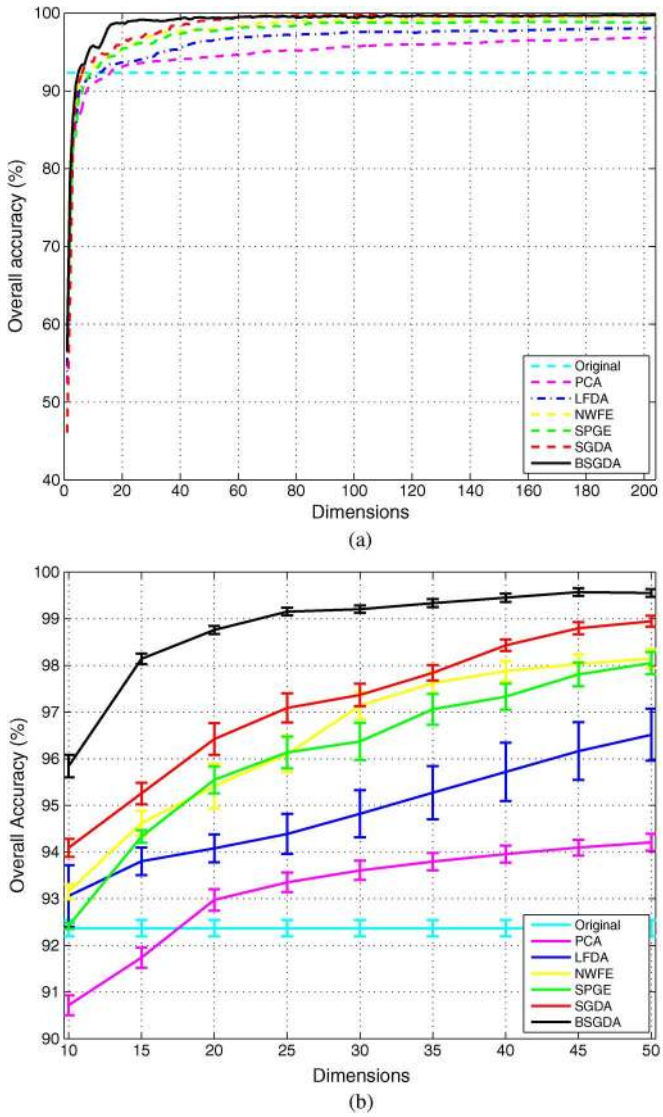

Fig. 8. SVM-CK classification performance (overall accuracy) under DR for Salinas. Error bars indicate standard deviation. (a) Mean accuracy. (b) Mean accuracy with standard deviation.

TABLE IV
SVM-CK CLASSIFICATION ACCURACY (%) FOR SALINAS ($K = 30$)

Class	Original (SVM)	Original (SVM-CK)	PCA	LFDA	NWFE	SPGE	SGDA	BSGDA
1	98.8 ± 0.6	99.1 ± 0.5	99.4 ± 0.2	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
2	99.8 ± 0.2	99.8 ± 0.2	99.9 ± 0.1	$\mathbf{100.0 \pm 0.0}$	99.9 ± 0.0	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
3	99.2 ± 0.7	99.4 ± 0.5	97.8 ± 1.4	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
4	99.3 ± 0.2	99.3 ± 0.2	$\mathbf{99.4 \pm 0.1}$	98.7 ± 0.6	98.2 ± 0.9	98.6 ± 0.7	97.8 ± 1.0	97.8 ± 1.2
5	98.4 ± 0.4	98.5 ± 0.5	98.6 ± 0.3	98.8 ± 0.4	98.9 ± 0.5	$\mathbf{99.3 \pm 0.1}$	99.0 ± 0.5	99.2 ± 0.4
6	99.8 ± 0.1	99.8 ± 0.1	99.9 ± 0.1	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
7	99.5 ± 0.2	99.6 ± 0.2	99.8 ± 0.1	99.7 ± 0.2	99.9 ± 0.1	99.9 ± 0.1	99.8 ± 0.2	$\mathbf{100.0 \pm 0.0}$
8	91.4 ± 0.7	91.3 ± 0.7	92.6 ± 0.5	91.5 ± 0.6	92.5 ± 0.5	91.3 ± 0.5	93.7 ± 1.3	$\mathbf{97.7 \pm 0.4}$
9	99.7 ± 0.3	99.8 ± 0.3	99.1 ± 0.5	99.8 ± 0.1	99.9 ± 0.3	99.8 ± 0.5	99.8 ± 0.3	$\mathbf{100.0 \pm 0.0}$
10	94.1 ± 0.7	95.8 ± 0.7	95.4 ± 0.7	99.3 ± 0.2	99.5 ± 0.2	99.5 ± 0.2	99.9 ± 0.2	$\mathbf{99.9 \pm 0.1}$
11	95.3 ± 1.9	97.8 ± 1.6	96.7 ± 2.3	99.3 ± 0.2	99.5 ± 0.3	99.6 ± 0.2	99.8 ± 0.2	$\mathbf{100.0 \pm 0.0}$
12	99.5 ± 0.5	99.8 ± 0.2	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
13	98.1 ± 0.6	98.5 ± 0.7	99.1 ± 0.2	99.8 ± 0.2	99.8 ± 0.2	99.9 ± 0.1	99.8 ± 0.3	$\mathbf{100.0 \pm 0.0}$
14	93.6 ± 1.0	95.5 ± 1.4	96.8 ± 1.0	99.2 ± 0.9	99.3 ± 1.0	99.8 ± 0.4	99.8 ± 0.3	$\mathbf{99.9 \pm 0.2}$
15	55.7 ± 1.6	61.8 ± 1.6	68.8 ± 1.5	75.8 ± 3.8	91.5 ± 2.7	87.3 ± 3.4	91.0 ± 2.1	$\mathbf{98.4 \pm 0.7}$
16	98.5 ± 0.7	98.6 ± 0.7	98.8 ± 0.4	99.9 ± 0.2	99.8 ± 0.2	99.7 ± 0.2	99.9 ± 0.1	$\mathbf{100.0 \pm 0.0}$
Overall	91.3 ± 0.2	92.4 ± 0.2	93.6 ± 0.2	94.8 ± 0.5	97.1 ± 0.3	96.4 ± 0.4	97.4 ± 0.2	$\mathbf{99.2 \pm 0.1}$
Average	95.0 ± 0.2	95.9 ± 0.2	96.4 ± 0.3	97.7 ± 0.3	98.7 ± 0.2	98.5 ± 0.2	98.8 ± 0.1	$\mathbf{99.6 \pm 0.1}$
κ	90.3 ± 0.2	91.5 ± 0.2	92.9 ± 0.2	94.2 ± 0.6	96.8 ± 0.3	96.0 ± 0.4	97.1 ± 0.3	$\mathbf{99.1 \pm 0.1}$

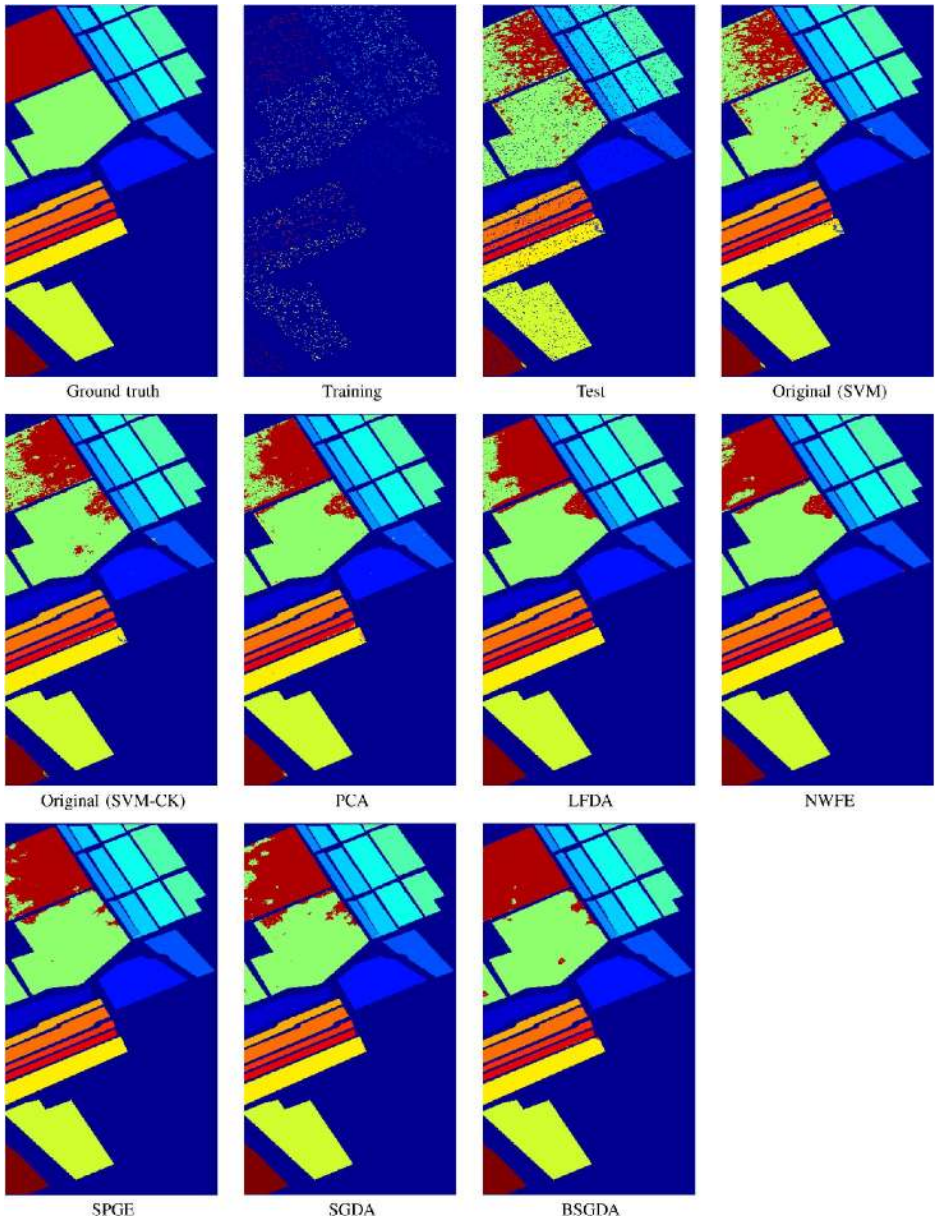

Fig. 9. Ground-truth and SVM-CK classification maps for Salinas ($K = 30$).

Gaussian heat kernel. Specifically, two vertices are connected in the k-NN graph if and only if they are among the k nearest neighbors of each other (we use $k = 5$ in both graphs). Regarding weight assignment, all edge weights are unity for binary weights, while the edge weight between vertices m and m' is

$$\mathbf{W}_{m,m'} = \exp\left(-\frac{\|\mathbf{x}_m - \mathbf{x}_{m'}\|^2}{2}\right) \tag{37}$$

for the Gaussian heat kernel. The parameters are chosen after careful cross-validation. We see that the k-NN graphs result in performance significantly lower than that of the ℓ_1 graph. We note that, for Fig. 6, we test using SPGE; however, similar results have been observed for the SGDA and BSGDA graph constructions.

Finally, Fig. 7 demonstrates the effect of added noise, in terms of signal-to-noise ratio (SNR), on the overall classification accuracy of the various algorithms. We test both Gaussian and uniform noise and average over 100 runs as was done in [29]. We observe that BSGDA gives the best accuracy at every SNR, emphasizing the robustness to noise that BSGDA provides as compared to other DR techniques. We note that low SNR ($\leq$ 30 dB) is commonly encountered in real-world situations, which is where BSGDA provides the most gain over other techniques.

C. AVIRIS Data Set: Salinas

The second data set evaluated was collected over the Valley of Salinas, Central Coast of California, in 1998. It contains 217 $\times$ 512 pixels and 224 spectral bands over 0.4 to 2.5 μm with spatial resolution of 3.7 m. Table III tabulates the 16 classes of interest; we randomly choose 5% of the labeled pixels in each class as training data and the remainder for test data. We note that this image has greater spatial homogeneity than is present in the Indian Pines image.

The parameters of SVM and SVM-CK are set to $(\varsigma, \gamma, \mu) = (512, 0.1768, 0.8)$ for the original data, and slightly varied for DR according to cross validation. We investigate performance as the reduction of dimensionality varies in Fig. 8; again, BSGDA consistently outperforms the others. The classification accuracy for each class, overall accuracy, average accuracy, and the κ coefficient are given in Table IV when $K = 30$. Classification maps on labeled pixels obtained are shown in Fig. 9; we see that the BSGDA map is smoother and in greater accord with the ground truth.

D. ROSIS Data Set: Pavia

The final image is the University of Pavia, an urban scene acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) [30]. The ROSIS sensor generates 115 spectral bands over 0.43 to 0.86 μm. The University of Pavia image has a spatial resolution of 1.3 m and 610 $\times$ 340 pixels, each having 103 bands after bad-band removal. There are nine ground-truth classes, as shown in Table V. For this image, we randomly select 8% of all labeled data as training and the remaining 92% as testing. The University of Pavia image is collected from an urban area and consequently consists of small buildings, materials, and trees—this image is substantially different from the homogeneous Indian Pines and Salinas data sets. Classifying this image therefore is expected to be a challenging task for many algorithms.

TABLE V
GROUND-TRUTH CLASSES AND CORRESPONDING TRAINING- AND TESTING-SET SIZES FOR UNIVERSITY OF PAVIA

	Classes		
No.	*Name*	*train*	*test*
1	Asphalt	530	6101
2	Meadows	1492	17157
3	Gravel	168	1931
4	Trees	245	2819
5	Metal sheets	108	1237
6	Bare soil	402	4627
7	Bitumen	106	1224
8	Bricks	295	3387
9	Shadows	76	871
	Total	3422	39354

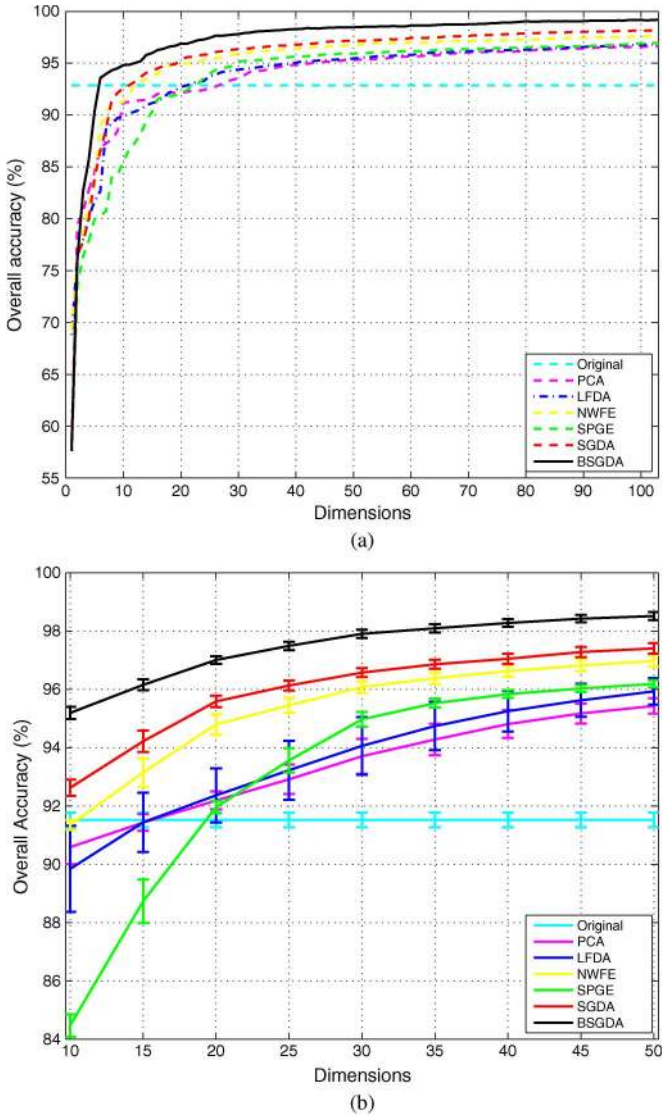

Fig. 10. SVM-CK classification performance (overall accuracy) under DR for University of Pavia. Error bars indicate standard deviation. (a) Mean accuracy. (b) Mean accuracy with standard deviation.

We set $(\varsigma, \gamma, \mu) = (128, 0.3536, 0.45)$ for SVM-CK of the original data and modify appropriately for DR based on cross validation. The classification performance as the reduction of

TABLE VI
SVM-CK CLASSIFICATION ACCURACY (%) FOR UNIVERSITY OF PAVIA ($K = 30$)

Class	Original (SVM)	Original (SVM-CK)	PCA	LFDA	NWFE	SPGE	SGDA	BSGDA
1	93.6 ± 0.8	93.6 ± 0.9	95.4 ± 0.8	96.0 ± 0.8	96.2 ± 0.7	96.1 ± 0.7	96.0 ± 0.7	$\mathbf{98.3 \pm 0.4}$
2	98.8 ± 0.2	99.0 ± 0.2	99.2 ± 0.2	99.8 ± 0.2	99.9 ± 0.0	99.6 ± 0.2	99.9 ± 0.1	$\mathbf{100.0 \pm 0.0}$
3	66.5 ± 2.5	66.6 ± 2.6	70.2 ± 2.0	67.1 ± 1.9	70.7 ± 2.4	66.1 ± 1.5	70.6 ± 2.7	$\mathbf{84.0 \pm 2.2}$
4	91.9 ± 1.0	93.6 ± 1.0	96.9 ± 0.5	96.4 ± 0.5	$\mathbf{97.7 \pm 0.4}$	92.2 ± 0.7	96.2 ± 0.6	96.4 ± 0.7
5	99.4 ± 0.3	99.4 ± 0.3	99.6 ± 0.1	99.7 ± 0.2	99.8 ± 0.1	99.9 ± 0.5	99.9 ± 0.1	$\mathbf{100.0 \pm 0.0}$
6	72.7 ± 1.3	73.6 ± 1.3	92.3 ± 1.4	91.2 ± 4.5	94.3 ± 1.1	97.2 ± 0.5	97.7 ± 0.3	$\mathbf{99.4 \pm 0.3}$
7	64.0 ± 7.5	64.6 ± 7.4	34.0 ± 17.7	44.2 ± 25.1	86.0 ± 2.2	69.0 ± 7.1	92.1 ± 1.4	$\mathbf{96.9 \pm 1.5}$
8	91.4 ± 1.0	91.6 ± 1.0	93.3 ± 1.0	93.5 ± 1.0	93.4 ± 1.1	91.4 ± 1.3	93.6 ± 1.2	$\mathbf{93.8 \pm 0.9}$
9	99.8 ± 0.1	99.9 ± 0.3	99.9 ± 0.5	99.6 ± 0.3	98.9 ± 0.6	$\mathbf{99.9 \pm 0.1}$	99.9 ± 0.2	96.5 ± 1.5
Overall	91.2 ± 0.2	91.5 ± 0.2	93.7 ± 0.6	94.1 ± 1.0	96.1 ± 0.2	95.0 ± 0.3	96.6 ± 0.2	$\mathbf{97.9 \pm 0.1}$
Average	86.5 ± 0.8	86.9 ± 0.8	86.8 ± 2.0	87.5 ± 2.9	93.0 ± 0.4	90.2 ± 0.8	94.0 ± 0.3	$\mathbf{96.1 \pm 0.3}$
κ	88.1 ± 0.3	88.6 ± 0.3	91.6 ± 0.8	92.1 ± 1.3	94.8 ± 0.3	93.3 ± 0.3	95.4 ± 0.2	$\mathbf{97.2 \pm 0.2}$

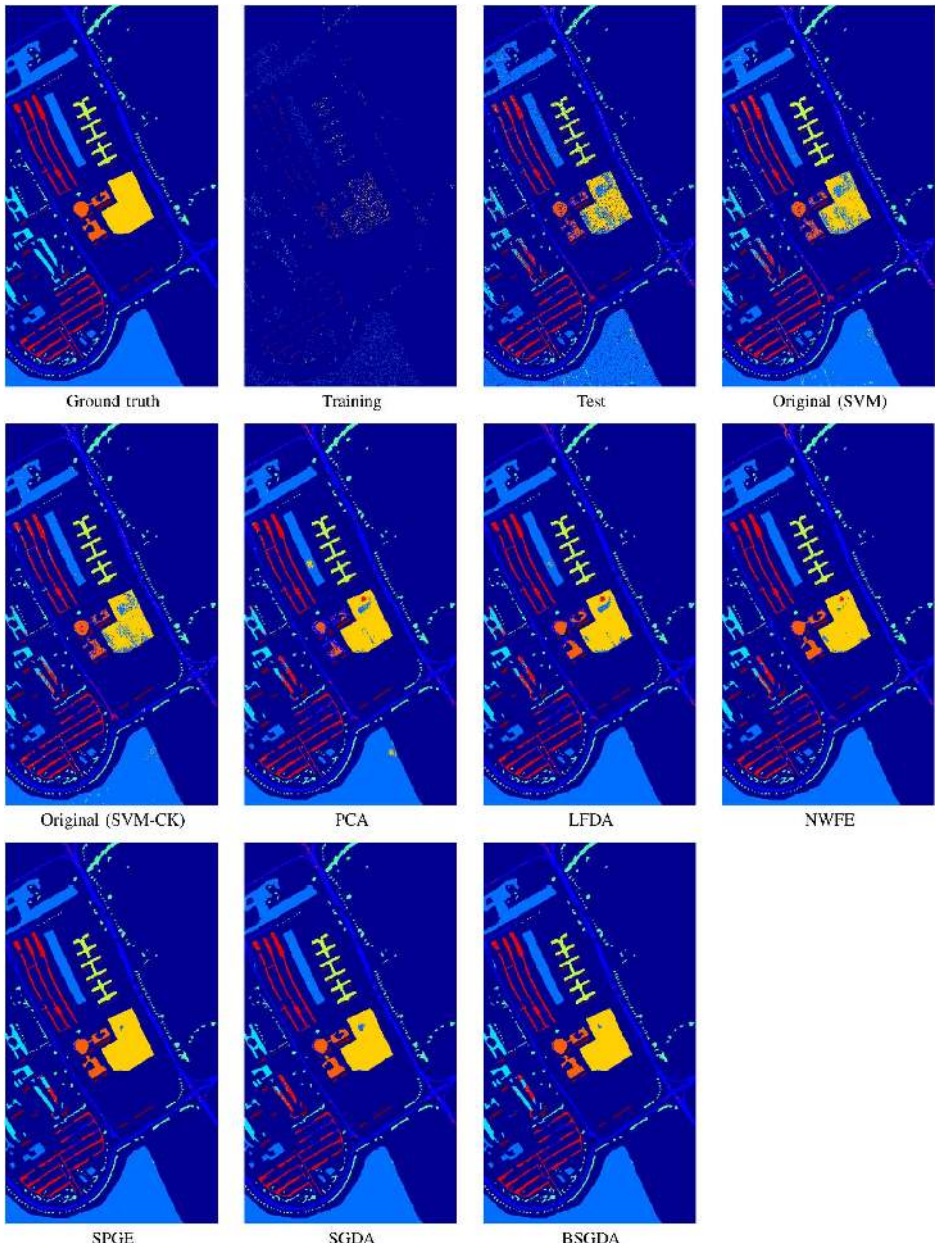

Fig. 11. Ground-truth and SVM-CK classification maps for University of Pavia ($K = 30$).

dimensionality varies is shown in Fig. 10. The classification accuracy for each class, overall accuracy, average accuracy, and the Cohen-κ coefficient are shown in Table VI using a reduced dimensionality of $K = 30$. For this image, BSGDA yields better overall performance compared to SGDA as well as other methods for DR plus SVM-CK. Classification maps

TABLE VII
COMPUTATION TIME FOR "UNIVERSITY OF PAVIA" ($\mathbf{W}_s$ HAS SIZE 3422×3422)

Algorithm	Time (sec.)
NWFE	34539
LFDA	0.15
PCA	0.06
SPGE	735
SGDA	738
BSGDA	18

are shown in Fig. 11. There is a significant difference between BSGDA and the other DR methods. Visually, BSGDA produces the smoothest map.

To illustrate the computational efficiency of the proposed algorithms as compared to other DR approaches, Table VII shows the computing times[6] for the various methods for the University of Pavia data set in which the proposed algorithms are slower than PCA/LFDA but much faster than NWFE. Specifically, we note:

1) Computation for NWFE includes calculations of a pairwise distance matrix, weighted means, scatter-matrix weight, as well as a generalized eigenvalue decomposition [28].
2) PCA is calculated via SVD, while LFDA constructs *local* between-class and within-class scatter matrices based on an *affinity* matrix with the projection matrix finally computed by a generalized eigenvalue problem [1], [4].
3) The heaviest computational burden in SPGE, SGDA, and BSGDA is the construction of an ℓ_1 graph while the projection matrices are calculated the same as in NWFE and LFDA. BSGDA is fast due to the use of within-class samples only for sparse representation of each sample.

V. CONCLUSION

DR has been widely used as a preprocessing step for HSI analysis. In this paper, we proposed a new supervised DR using a sparse graph. The resulting SGDA exploited the discriminant capability from sparse representation, and the discriminant power was reinforced when labeled information was explicitly utilized. In particular, the proposed BSGDA constrained the sparse representation to using only samples with the same class label to ensure that within-class samples are further clustered together in the low-dimensional space, thereby maintaining or even enhancing class separability. Experimental results demonstrated that the transformed data from SGDA- and BSGDA-based DR yielded classification performance for HSI superior to that of other widely used DR methods. The performance improvement is significant even when the number of training samples is too limited to apply traditional LDA.

The proposed algorithms inherit the advantage of graph-based data analysis in terms of capturing geometric structures of the original data (here, it is about maintaining sparse presentation among classes). However, it has the same disadvantage of high computational cost associated with graph construction, which is related to determining a large weight matrix (equal to the number of samples) and solving the corresponding eigendecomposition problem.

[6]All of the experiments are carried out using MATLAB on a quad-core 3.2-GHz machine with 5.8 GB of RAM.

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.
[2] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification,* 2nd ed. New York, NY, USA: Wiley, 2001.
[4] M. Sugiyama, "Dimensionaltity reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, no. 5, pp. 1027–1061, May 2007.
[5] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 7, pp. 793–802, Jul. 2004.
[6] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 564–568, Oct. 2008.
[7] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, Jan. 2011.
[8] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[9] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with ℓ_1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.
[10] 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.
[11] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1873–1879.
[12] S. Siddiqui, S. Robila, J. Peng, and D. Wang, "Sparse representations for hyperspectral data classification," in *Proc. Int. Geosci. Remote Sens. Symp.*, Boston, MA, USA, Jul. 2008, vol. 2, pp. II-577–II-580.
[13] M. H. Rohban and H. R. Rabiee, "Supervised neighborhood graph construction for semi-supervised classification," *Pattern Recognit.*, vol. 45, no. 4, pp. 1363–1372, Apr. 2012.
[14] 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.
[15] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[16] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.
[17] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving discriminant analysis for single training image face recognition," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 422–429, Apr. 2010.
[18] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, Jan. 2010.
[19] L. Zhang, S. Chen, and L. Qiao, "Graph optimization for dimensionality reduction with sparsity constraints," *Pattern Recognit.*, vol. 45, no. 3, pp. 1205–1210, Mar. 2012.
[20] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[21] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.
[22] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 629–640, Jun. 2011.
[23] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification via kernel sparse representation," in *Proc. Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 1233–1236.

[24] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[25] 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, no. Suppl. 1, pp. S110–S122, Sep. 2009.
[26] E. van den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Scientific Comput.*, vol. 31, no. 2, pp. 890–912, 2008.
[27] 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.
[28] B.-C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.
[29] 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.
[30] P. Gamba, "A collection of data for urban area characterization," in *Proc. Int. Geosci. Remote Sens. Symp.*, Anchorage, AK, USA, Sep. 2004, vol. 1, pp. 69–72.

Nam Hoai Ly (S'11) received the B.E. degree in electronics and telecommunications from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2008. Currently, he is working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Mississippi State University, Starkville.

From 2009 to 2013, he was a Research Assistant with the Geosystems Research Institute and Teaching Assistant with the Department of Electrical and Computer Engineering, Mississippi State University. Currently, he is with Schweitzer Engineering Laboratories, Pullman, WA. His research interests include hyperspectral image processing, machine learning, and signal processing.

Qian Du (S'98–M'00–SM'05) received the Ph.D. degree in electrical engineering from the University of Maryland Baltimore County, Baltimore, in 2000.

She is the Bobby Shackouls Professor in the Department of Electrical and Computer Engineering at Mississippi State University, Starkville. Her research interests include hyperspectral remote sensing image analysis, pattern classification, data compression, and neural networks.

Dr. Du served as Co-chair for the Data Fusion Technical Committee of IEEE Geoscience and Remote Sensing Society (2009–2013) and currently is the chair for Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition (IAPR). She also serves as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and an Associate Editor for IEEE SIGNAL PROCESSING LETTERS. She received the 2010 Best Reviewer award from IEEE Geoscience and Remote Sensing Society. She is the General Chair for the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) in Shanghai, China in 2012. She is a Member of SPIE, ASPRS, and ASEE.

James E. Fowler (S'91–M'96–SM'02) received the B.S. degree in computer and information science engineering and the M.S. and Ph.D. degrees in electrical engineering in 1990, 1992, and 1996, respectively, all from The Ohio State University, Columbus, OH, USA.

In 1995, he was an Intern Researcher at AT&T Labs, Holmdel, NJ, USA, and, in 1997, he held a National Science Foundation-sponsored postdoctoral assignment at the Université de Nice-Sophia Antipolis, Nice, France. In 2004, he was a Visiting Professor with the Département Traitement du Signal et des Images, École Nationale Supérieure des Télécommunications, Paris, France. Currently, he is the Billie J. Ball Professor and Graduate Program Director of the Department of Electrical and Computer Engineering at Mississippi State University, Starkville; he is also a researcher in the Geosystems Research Institute (GRI) at Mississippi State.

Dr. Fowler is an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and the *EURASIP Journal on Image and Video Processing*; he formerly served as an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA and IEEE SIGNAL PROCESSING LETTERS. He is the Chair of the Image, Video, and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society and a Member of the Strategic Planning Committee of the IEEE Publication Services and Products Board. He is the General Co-chair of the 2014 IEEE International Conference on Image Processing, Paris, France, as well as the Publicity Chair of the program committee for the Data Compression Conference.