Nearest Regularized Subspace for Hyperspectral Classification

Wei Li, *Member, IEEE*, Eric W. Tramel, *Member, IEEE*, Saurabh Prasad, *Member, IEEE*, and James E. Fowler, *Senior Member, IEEE*

***Abstract*—A classifier that couples nearest-subspace classification with a distance-weighted Tikhonov regularization is proposed for hyperspectral imagery. The resulting nearest-regularized-subspace classifier seeks an approximation of each testing sample via a linear combination of training samples within each class. The class label is then derived according to the class which best approximates the test sample. The distance-weighted Tikhonov regularization is then modified by measuring distance within a locality-preserving lower-dimensional subspace. Furthermore, a competitive process among the classes is proposed to simplify parameter tuning. Classification results for several hyperspectral image data sets demonstrate superior performance of the proposed approach when compared to other, more traditional classification techniques.**

***Index Terms*—Classification, hyperspectral data, Tikhonov regularization.**

I. Introduction

OVER the last decade, hyperspectral imagery (HSI) obtained by remote-sensing systems has been investigated at length [1]. HSI provides high-resolution spectral information over a wide range of the electromagnetic spectrum with hundreds of observed spectral bands. Numerous supervised classification techniques for hyperspectral data have been developed (e.g., [2]–[5]) for a variety of application areas, including agricultural monitoring, environment-pollution monitoring, and urban-growth analysis, among others.

The k-nearest-neighbor (k-NN) classifier (e.g., [6], [7]), one of the simplest and oldest classification methods, has been widely used for HSI classification. This nonparametric classifier usually employs a Euclidean distance between the training and testing samples, assigning class labels according to the most frequently occurring class of the k nearest training samples. However, the high-dimensional nature of HSI data creates complications for k-NN classification in terms of both computational complexity and classification accuracy. Many dimensionality-reducing techniques have been proposed to combat this so-called curse of dimensionality, such as the popular linear discriminant analysis (LDA) [8] and its variants (e.g., [9], [10]). Typically, parametric classification is employed after dimensionality reduction, for example the maximum likelihood estimation (MLE) [11] of posterior probabilities. The support vector machine (SVM) [12] is a state-of-the-art classifier which has also been shown to work well for hyperspectral classification tasks. An SVM seeks to separate classes by learning an optimal decision hyperplane which best separates the training samples in a kernel-induced high-dimensional feature space. Nonlinear kernels may also be used within the SVM framework to achieve nonlinear separation in the feature space via linear separation in the kernel-induced space. Variations of the SVM (e.g., [3], [13]) have been proposed to further improve classification performance. For example, in [13], locality Fisher's discriminant analysis (LFDA) was employed to reduce the dimensionality of hyperspectral data for the SVM classifier. The LFDA-SVM technique of [13] was demonstrated to be effective for HSI classification, especially when few training samples are available.

Recently, Wright *et al.* [14] introduced sparse-representation classification (SRC) for face recognition—in essence, SRC represents a testing sample by a sparse linear combination of training samples calculated via ℓ_1 minimization. In [15], the authors applied a sparse framework for HSI classification and subsequently exploited sparsity for the classification task in a graphical model [16], [17] and a kernel space [18], [19]. There are a number of additional works that invoke sparse representation specifically for HSI classification—for example, [20] adopted sparse representation in the special case that very few labeled training samples are available; [21] considered discriminative sparse representation; while [22] introduced sparse representation in semisupervised learning.

An approach similar to SRC was taken by Zhang *et al.* [23] who proposed collaborative-representation classification (CRC) for face recognition. CRC is similar to SRC in that a linear combination of training samples represents a testing sample. However, contrary to the ℓ_1-based sparsity-inducing regularization of SRC, CRC uses an ℓ_2-regularized minimization, providing competitive face-recognition accuracy but at significantly lower computational complexity.

In this paper, we couple nearest-subspace classification with the distance-weighted Tikhonov regularization from [24], [25]. In the resulting system, which can be considered to be a

Manuscript received April 3, 2012; revised August 10, 2012, November 20, 2012, and December 28, 2012; accepted January 15, 2013. Date of publication March 7, 2013; date of current version November 26, 2013. This work was supported by the University of Houston Startup Funding, and the National Science Foundation under Grant CCF-0915307.

W. Li is with the University of California, Davis, Davis, CA 95616 USA (e-mail: liwei089@ieee.org).

E. W. Tramel and J. E. Fowler are with the Department of Electrical and Computer Engineering and the Geosystems Research Institute, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: eric.tramel@gmail.com; fowler@ece.msstate.edu).

S. Prasad is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77204-4005 USA (e-mail: saurabh.prasad@ieee.org).

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.2013.2241773

nearest-regularized-subspace (NRS) classifier, an approximation for each testing sample is created via linear combination of all available training samples within each class. In this manner, an approximation of each test sample is generated from training samples of each class independently. The class label is then derived according to the class of the most accurate representation. In a general sense, this NRS classification is similar to both SRC and CRC in that testing samples are approximated via linear combinations of training samples; however, NRS differs in that, not only does it use a noncollaborative approach to the approximation, but it also employs non-uniform regularization.

We also introduce, as a further extension of the proposed NRS paradigm, a discrimination-enhancing distance measure [26] designed to improve classification accuracy. Furthermore, a competitive strategy is presented for automatically obtaining optimal performance for the proposed system, thus avoiding involved parameter tuning via cross-validation. Classification results are presented for several HSI data sets to demonstrate the superior classification accuracy of the proposed approach when compared to traditional classification techniques. Ultimately, our work is composed of three main contributions: 1) the NRS classification system based on a distance-weighted Tikhonov regularization (an ℓ_2-regularized term) calculating a representation for each testing sample; 2) a discrimination-enhancing distance measure which improves the Tikhonov biasing term; and 3) a competitive strategy that eliminates the need for involved parameter tuning.

The paper is organized as follows: in Section II, we provide a brief review of relevant classification methods, while in Section III, we provide a detailed description of the proposed NRS classifier and its variants. In Section IV, we experimentally compare the performance of the proposed method with several conventional HSI classification techniques. We conclude by summarizing our results in Section V.

II. Background

A. Nearest-Neighbor Classification

The nearest-neighbor (NN) algorithm (e.g., [6], [7]) is perhaps the simplest supervised method to predict a testing-sample label. The NN classifier attempts to find the training sample nearest to the testing sample according to a given distance measure, assigning the former's category to the latter. Consider a data set with training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ in $\mathbb{R}^d$ (d-dimensional feature space) and class labels $\omega_i \in \{1, 2, \ldots, C\}$, where C is the number of classes, and n is the total number of training samples. Let n_l be the number of available training samples for the l^{th} class, $\sum_{l=1}^{C} n_l = n$. Commonly, Euclidean distance is used, such that the distance measure between training sample $\mathbf{x}_i$ and given testing sample $\mathbf{y}$ is

$$d(\mathbf{x}_i, \mathbf{y}) = \|\mathbf{x}_i - \mathbf{y}\|_2^2. \tag{1}$$

The k-NN classifier is a straightforward extension of the original NN classifier. Instead of using only one sample closest to testing point $\mathbf{y}$, the k-NN classifier chooses the k nearest samples from training data $\mathbf{X}$. Typically, k is an odd number, and majority voting is employed to decide the final label.

B. ℓ_1- and ℓ_2-Regularized Collaborative Representation for Classification

Classification based on sparse representation has been recently studied for both for face recognition [14], and HSI analysis [15], [20]. The SRC approach offers classification which is robust to noise and model errors; for more discussion of the geometrical and graphical interpretations of SRC, we refer the reader to [14].

In essence, an SRC method classifies a testing sample $\mathbf{y}$ according to the class which produces the most accurate sparse representation of $\mathbf{y}$, i.e., the class which produces the most parsimonious description using the training data as the "dictionary" for forming the representation. First, an approximation of $\mathbf{y}$ is calculated via a sparse linear combination of all available training samples. That is, for training samples arranged column-wise in the matrix $\mathbf{X}$ of dimensionality $d \times n$, we desire to find an $n \times 1$ vector of sparse coefficients, $\boldsymbol{\alpha}$, such that $\mathbf{X}\boldsymbol{\alpha}$ is near to $\mathbf{y}$. Basis pursuit denoising (BPDN) [27] offers one approach for calculating $\boldsymbol{\alpha}$ by solving the ℓ_1-regularized minimization

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \tag{2}$$

where the regularization parameter, $\lambda > 0$, balances the influence of the residual and sparsity terms. We mention the BPDN formulation in particular here because of its confluence with several regularization techniques we present later. However, other formulations may be equivalently substituted, such as the least absolute shrinkage and selection operator (LASSO) [28] or basis pursuit (BP) [27]. In any event, after $\boldsymbol{\alpha}$ is calculated, a representation for each class, $\widetilde{\mathbf{y}}_l$, is created through a process we term *postpartitioning*.

The postpartitioning approach separates $\mathbf{X}$ into l different class-specific sub-dictionaries according to the given class labels of the training data, $\mathbf{X}_l = \{\mathbf{x}_i |\ \forall i\ s.t.\ \omega_i = l\}$; additionally, the coefficient vector $\boldsymbol{\alpha}$ is also "partitioned" similarly into $\boldsymbol{\alpha}_l = \{\alpha_i |\ \forall i\ s.t.\ \omega_i = l\}$. After this partitioning, class-specific representations, $\widetilde{\mathbf{y}}_l$, are calculated as

$$\widetilde{\mathbf{y}}_l = \mathbf{X}_l\boldsymbol{\alpha}_l. \tag{3}$$

We note that this use of all the training data concurrently, as in postpartitioning, stands in contrast to the traditional approach used in nearest-subspace (NS) classifiers [29], [30] which use what we call *prepartitioning*. In such prepartitioning, the training data is first partitioned into $\mathbf{X}_l$, and these partitions are instead used to calculate each $\widetilde{\mathbf{y}}_l$ independently, via, e.g., BPDN applied independently for each partition.

In SRC, after calculating each $\widetilde{\mathbf{y}}_l$ via (3), the class label of $\mathbf{y}$ is then determined according to the class which minimizes the residual, i.e.,

$$\text{class}(\mathbf{y}) = \arg\min_{l=1,\ldots,C}(r_l) \tag{4}$$

where $r_l = \|\widetilde{\mathbf{y}}_l - \mathbf{y}\|_2^2$ is the residual between the approximation and corresponding testing sample. A detailed description of the SRC algorithm is given as Algorithm 1.

Algorithm 1 The SRC Algorithm
Input: Training data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, class labels ω_i, testing sample $\mathbf{y} \in \mathbb{R}^d$, λ
Calculate $\boldsymbol{\alpha}$ via ℓ_1-minimization of (2)
for all $l \in \{1, 2, \ldots, C\}$ **do**
Partition $\mathbf{X}_l$, $\boldsymbol{\alpha}_l$
Calculate $\widetilde{\mathbf{y}}_l = \mathbf{X}_l \boldsymbol{\alpha}_l$
end for
Decide class$(\mathbf{y})$ via (4)
Output: class$(\mathbf{y})$

In [14], [15], it was posited that the sparse representation alone led to the observed improvements in classification accuracy. However, both [31] and [23] raise concerns over the SRC framework. In [31], it was shown via analysis of singular values that face data sets are, generally, not a suitable fit for SRC. To show that a sparse approach is unwarranted for face recognition, a QR decomposition was used to calculate each $\widetilde{\mathbf{y}}_l$ instead of sparse approximation; the resulting performance of this technique was competitive with that of SRC.

Additionally, in [23], it was suggested that the improvement in classification accuracy was not due to sparsity, but rather due to the "collaborative" nature of the approximation. Specifically, it was argued that using the entire training data set to form approximations via postpartitioning rather than using prepartitioning as in NS allows for acceptable classification accuracy when signal dimensionality is high or when the number of available training samples are few. To support this argument, [23] proposed the CRC approach which swapped the ℓ_1 penalty of SRC for an ℓ_2 penalty in the style of Tikhonov regularization [32], i.e.,

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2. \tag{5}$$

Rather than enforcing a strong assumption about the nature of the data set's geometry, the ℓ_2-regularization (or *shrinkage*) term instead serves only to overcome the potential for ill-conditioning and ill-posedness in the inverse problem.

One particular advantage of CRC is that (5) may be solved with a simple and closed form. To do this, we take the derivative of the cost function of (5)

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T\boldsymbol{\theta} \tag{6}$$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta}. \tag{7}$$

By setting the derivative to zero, we find the value of $\boldsymbol{\theta}$ which minimizes the cost function

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{8}$$

where $\mathbf{I}$ is an identity matrix of appropriate size. From (8), we may calculate the approximation of $\mathbf{y}$

$$\widetilde{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}_{\mathrm{CRC}}\mathbf{y}. \tag{9}$$

After calculating $\widetilde{\mathbf{y}}$, the postpartitioning and classification is carried out in a manner identical to the SRC via (3) and (4). It is noted in [23] that $\mathbf{H}_{\mathrm{CRC}}$ is dependent upon only the available training data. Thus, the projector $\mathbf{H}_{\mathrm{CRC}}$ may be precomputed to reduce classification time for large-volume tasks. CRC was shown to provide face-recognition accuracy comparable to SRC with much lower computational cost. A detailed description of CRC is given as Algorithm 2.

Algorithm 2 The CRC Algorithm
Input: Training data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, class labels ω_i, testing sample $\mathbf{y} \in \mathbb{R}^d$, λ
Calculate $\boldsymbol{\alpha}$ according to (9)
for all $l \in \{1, 2, \ldots, C\}$ **do**
Partition $\mathbf{X}_l$, $\boldsymbol{\alpha}_l$
Calculate $\widetilde{\mathbf{y}}_l = \mathbf{X}_l \boldsymbol{\alpha}_l$
end for
Decide class$(\mathbf{y})$ according to (4)
Output: class$(\mathbf{y})$

Algorithm 3 Proposed NRS Classifier
Input: Training data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, class labels ω_i, testing sample $\mathbf{y} \in \mathbb{R}^d$, λ
Partition $\mathbf{X}_l$
for all $l \in \{1, 2, \ldots, C\}$ **do**
Calculate $\boldsymbol{\Gamma}_{l,y}$ according to (11)
Calculate $\widetilde{\mathbf{y}}_l$ according to (12)
end for
Decide class$(\mathbf{y})$ according to (4)
Output: class$(\mathbf{y})$

The common element between these works and the sparse approaches of [14], [15] is the assumption of a collaborative, postpartitioning framework for calculating class representations, $\widetilde{\mathbf{y}}_l$. However, this general approach is only loosely justified in previous literature with few significant details given for the departure from the NS approach of prepartitioning.

We investigate the effects of pre- and postpartitioning empirically for hyperspectral data in Fig. 1 using the Indian Pines data set with 1496 training samples (see Section IV-A for a detailed description of this data set). The classification accuracy is calculated over a range of possible values for the free regularization parameter, λ. We denote the prepartitioning technique here as CRC-Pre. The only difference between CRC-Pre and the postpartitioning-based CRC is that each $\widetilde{\mathbf{y}}_l$ is calculated in the former using only the training samples from class l, $\mathbf{X}_l$. Even though HSI data resides in the context proposed for collaborative techniques—namely high-dimensionality data with few training samples—Fig. 1 shows collaborative postpartitioning may actually do more harm than good. From these results, it is evident that advances in face recognition using collaborative approximations cannot be applied wholesale to HSI classification. We argue that a different approach is required.

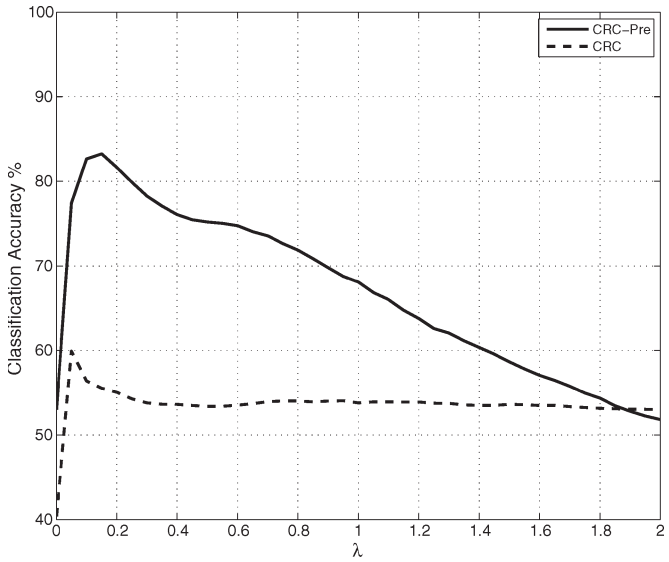

Fig. 1. Classification accuracy of pre- and postpartitioning (CRC-Pre and CRC, respectively) for the Indian Pines HSI data set over a range of values for the regularization parameter λ.

III. Proposed NRS Classifier

A. Basic NRS Algorithm

In this section, we propose the NRS classifier which couples prepartitioning as in NS with non-uniform Tikhonov regularization for the classification of hyperspectral data when few training samples are available. Like CRC, NRS makes use of Tikhonov regularization [32] to generate each $\widetilde{\mathbf{y}}_l$. However, instead of using uniform regularization as CRC does, we adopt a technique proposed in [24], [25], therein termed multihypothesis (MH) prediction, which biases atoms of $\mathbf{X}_l$ according to their Euclidean distance from $\mathbf{y}$. In [24], [25], MH prediction was used to recover video macroblocks from a small set of random linear measurements taken on the encoder side when a set of high-quality keyframe macroblocks was available on the decoder side via a linear combination of these keyframe macroblocks. The non-uniform nature of the regularization was used to penalize potentially inaccurate macroblocks from being assigned large contributions in the final recovery.

Likewise, in supervised classification, we are given a set of training, or hypothesis, data from which we desire to create approximations via linear combination. Namely, we seek an approximation of $\mathbf{y}$ for each class, $\widetilde{\mathbf{y}}_l$, calculated from only the training samples particular to class l, $\mathbf{X}_l$. We calculate the per-class coefficients, $\boldsymbol{\alpha}_l$, according to

$$\boldsymbol{\alpha}_l = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}_l\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\Gamma}_{l,y}\boldsymbol{\theta}\|_2^2 \tag{10}$$

where $\boldsymbol{\Gamma}_{l,y}$ is a biasing Tikhonov matrix specific to each class l and test sample $\mathbf{y}$, and λ is a global regularization parameter which balances the minimization between the residual and regularization terms. Specifically, we use a diagonal $\boldsymbol{\Gamma}_l$ in the form of

$$\boldsymbol{\Gamma}_{l,y} = \begin{bmatrix} \|\mathbf{y} - \mathbf{x}_{l,1}\|_2 & & 0 \\ & \ddots & \\ 0 & & \|\mathbf{y} - \mathbf{x}_{l,n_l}\|_2 \end{bmatrix} \tag{11}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_l}$ are the columns of $\mathbf{X}_l$ for the l^{th} class. According to the minimization defined in (10) and the structure of $\boldsymbol{\Gamma}_{l,y}$ given in (11), hypotheses which are the most dissimilar to $\mathbf{y}$, in terms of Euclidean distance, should be given much less contribution toward the linear combination than those which are most similar. Using this distance-weighting measure for $\boldsymbol{\Gamma}_{l,y}$ enforces a structural meaning to calculated weights without making as stringent of an assumption as true sparsity. Each testing sample $\widetilde{\mathbf{y}}_l$ can then be calculated in closed form in a similar fashion to (5), resulting in

$$\widetilde{\mathbf{y}}_l = \mathbf{X}_l\left(\mathbf{X}_l^T\mathbf{X}_l + \lambda\boldsymbol{\Gamma}_{l,y}^T\boldsymbol{\Gamma}_{l,y}\right)^{-1}\mathbf{X}_l^T\mathbf{y} = \mathbf{H}_{\text{NRS}}\mathbf{y}. \tag{12}$$

After calculating $\widetilde{\mathbf{y}}_l$ for each class, the class assignment for $\mathbf{y}$ is calculated according to (4).

The effect of the ℓ_2-regularization term is twofold. First, if the training samples are sufficiently similar in each class, or if a large set of training samples is used ($n_l \gg d$), the matrix $\mathbf{X}_l^T\mathbf{X}_l$ will either have poor conditioning or be near-singular. The consequence is that the calculation of its inverse will be inaccurate (or impossible), creating a lack of backwards stability in the inverse problem, leading to the calculated weights being of high variance and conveying little to no meaning. Enforcing the regularization term enforces stability on the problem by effectively inflating the singular values of $\mathbf{X}_l$, improving the conditioning of the problem. Second, the form of the biasing matrix $\boldsymbol{\Gamma}_{l,y}$ used in the regularization term allows for discrimination between classes. Without this term, it is possible, in certain conditions, for each $\mathbf{X}_l$ to approximate $\mathbf{y}$ with arbitrary accuracy, thus removing any discriminative power from r_l. This situation can be effected by setting $\lambda = 0$, causing (10) to become a least-squares (LSQ) problem. As illustrated in Fig. 1, a near-zero regularization term destroys the accuracy of the classifier.

To further investigate the relationship between NRS and CRC-Pre and the effects of the non-uniform Tikhonov matrix of (11), we compare the properties of both $\mathbf{H}_{\text{CRC}}$ and $\mathbf{H}_{\text{NRS}}$ in terms of their eigendecompositions as they relate to the singular values of the data matrix $\mathbf{X}_l$; the regularization parameter λ; and, in the case of NRS, the generalized singular values between $\mathbf{X}_l$ and the Tikhonov matrix $\boldsymbol{\Gamma}_{l,y}$. We show how NRS achieves varying degrees of shrinkage according to each particular test sample and contrast this with the uniform shrinkage applied by CRC-Pre. We also demonstrate how this variability across test samples gives NRS more flexibility in determining complex decision boundaries.

To do this, we will first decompose $\mathbf{H}_{\text{CRC}}$ according to the singular value decomposition (SVD) of $\mathbf{X}_l$, namely

$$\mathbf{X}_l = \widetilde{\mathbf{U}}\boldsymbol{\Sigma}\widetilde{\mathbf{V}}^T \tag{13}$$

where $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are both orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix of the singular values, $\{\sigma_1, \sigma_2, \ldots, \sigma_d\}$, of the data matrix $\mathbf{X}_l$. We substitute this decomposition into the equation for $\mathbf{H}_{\text{CRC}}$ given in (9)

$$\mathbf{H}_{\text{CRC}} = \widetilde{\mathbf{U}}\boldsymbol{\Sigma}\widetilde{\mathbf{V}}^T(\widetilde{\mathbf{V}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\widetilde{\mathbf{V}}^T + \lambda\mathbf{I})^{-1}\widetilde{\mathbf{V}}\boldsymbol{\Sigma}\widetilde{\mathbf{U}}^T \tag{14}$$

$$= \widetilde{\mathbf{U}}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}\widetilde{\mathbf{U}}^T \tag{15}$$

$$= \widetilde{\mathbf{U}}\widetilde{\mathbf{M}}\widetilde{\mathbf{U}}^T \tag{16}$$

where $\widetilde{\mathbf{M}}$ is a diagonal matrix consisting of the values $\{\widetilde{\mu}_1, \widetilde{\mu}_2, \ldots, \widetilde{\mu}_d\} \in [0, 1]$

$$\widetilde{\mu}_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda}. \tag{17}$$

To decompose $\mathbf{H}_{\text{NRS}}$, we instead employ generalized SVD of both $\mathbf{X}_l$ and $\mathbf{\Gamma}_{l,y}$

$$\mathbf{X}_l = \mathbf{U}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{Z}^T \tag{18}$$

$$\mathbf{\Gamma}_{l,y} = \mathbf{V}\mathbf{\Sigma}_{\mathbf{\Gamma}}\mathbf{Z}^T \tag{19}$$

where $\mathbf{Z} = \mathbf{Q}[\mathbf{0}, \mathbf{R}]^T$; $\mathbf{Q}$, $\mathbf{U}$, and $\mathbf{V}$ are orthogonal; and $\mathbf{R}$ is upper-triangular. The diagonal matrices $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{\Gamma}}$ contain the singular values $\{\sigma_{\mathbf{X},1}, \sigma_{\mathbf{X},2}, \ldots, \sigma_{\mathbf{X},p}\}$ and $\{\sigma_{\mathbf{\Gamma},1}, \sigma_{\mathbf{\Gamma},2}, \ldots, \sigma_{\mathbf{\Gamma},p}\}$ where $p = \min(d, n_l)$, since the generalized singular values of the two matrices cannot be calculated beyond the smallest column rank of the two. We note that $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{\Gamma}}$ are both zero-padded such that the dimensions are appropriate. The singular values provided by the generalized SVD decomposition have two unique properties, first, $\sigma_{\mathbf{X},i} \in [0, 1]$, and $\sigma_{\mathbf{\Gamma},i} \in [0, 1]$; and, second, $\sigma_{\mathbf{X},i}^2 + \sigma_{\mathbf{\Gamma},i}^2 = 1$.

Next, we substitute these decompositions into (12), and, using a procedure similar that used to calculate (16), we find

$$\mathbf{H}_{\text{NRS}} = \mathbf{U}\mathbf{\Sigma}_{\mathbf{X}}(\mathbf{\Sigma}_{\mathbf{X}}\mathbf{\Sigma}_{\mathbf{X}} + \lambda\mathbf{\Sigma}_{\mathbf{\Gamma}}\mathbf{\Sigma}_{\mathbf{\Gamma}})^{-1}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{U}^T \tag{20}$$

$$= \mathbf{U}\mathbf{M}\mathbf{U}^T \tag{21}$$

where $\mathbf{M}$ is a diagonal matrix containing the values in the range [0, 1]

$$\mu_i = \frac{\sigma_{\mathbf{X},i}^2}{\sigma_{\mathbf{X},i}^2 + \lambda\sigma_{\mathbf{\Gamma},i}^2}. \tag{22}$$

Additionally, since $\mathbf{M}$ must be of dimensionality $d \times d$, in the event that $n_l < d$, the entries $\{\mu_{n_l+1}, \ldots \mu_d\}$ are set to zero. The same is true for the values of $\widetilde{\mathbf{M}}$. Since the matrices $\widetilde{\mathbf{U}}$ and $\mathbf{U}$ are orthogonal, the two decompositions of (16) and (21) represent the eigendecompositions of the projection matrices $\mathbf{H}_{\text{CRC}}$ and $\mathbf{H}_{\text{NRS}}$.

We can observe that the values of $\widetilde{\mu}_i$ are dependent on the structure of the training samples given for each class and the value of the regularization parameter λ. This means that the same amount of shrinkage is applied to all test samples for a given class, creating a more general decision boundary (as evidenced by Figs. 3 and 4 which we describe shortly). The values of μ_i, however, are additionally dependent on the distance relationships between the columns of $\mathbf{X}_l$ and $\mathbf{y}$ according to the distance metric used to construct $\mathbf{\Gamma}_{l,y}$. In fact, from the properties of the singular values provided by the generalized SVD (i.e., $\sigma_{\mathbf{X},i}^2 + \sigma_{\mathbf{\Gamma},i}^2 = 1$), we can describe μ_i entirely by the singular values of $\mathbf{\Gamma}_{l,y}$

$$\mu_i = \frac{1 - \sigma_{\mathbf{\Gamma},i}^2}{1 - (1 - \lambda)\sigma_{\mathbf{\Gamma},i}^2}. \tag{23}$$

We plot the value of μ_i as a function of $\sigma_{\mathbf{\Gamma},i}$ for several different values of the regularization parameter, λ, in Fig. 2. Here, we see the inverse relationship between the size of the singular values of $\mathbf{\Gamma}_{l,y}$ and the resulting eigenvalues of $\mathbf{H}_{\text{NRS}}$. As $\mathbf{y}$ becomes more distant from the class training samples $\mathbf{X}_l$, the entries of $\mathbf{\Gamma}_{l,y}$ increase, and its singular values $\mathbf{\Sigma}_{\mathbf{\Gamma}}$ increase accordingly in proportion to $\mathbf{\Sigma}_{\mathbf{X}}$. This increase in $\mathbf{\Sigma}_{\mathbf{\Gamma}}$ forces the eigenvalues of $\mathbf{H}_{\text{NRS}}$ to decrease. Essentially, classes whose training samples are distant from $\mathbf{y}$ incur a stiffer shrinkage penalty than classes which contain training samples in close proximity to $\mathbf{y}$. By increasing the shrinkage penalty on such class's approximations, $\|\widetilde{\mathbf{y}}_l - \mathbf{y}\|_2^2$ grows, making these class assignments unlikely for $\mathbf{y}$. By including proximity information into the regularization, we see that the NRS classifier is a blend between distance, or exemplar, based classifiers such as k-NN, and NS-style classifiers such as CRC-Pre.

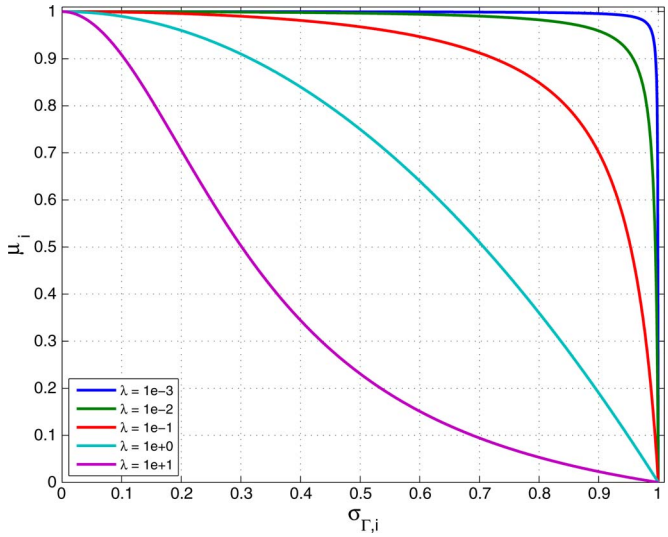

Fig. 2. Relationship between the singular values, $\sigma_{\mathbf{\Gamma},i}$, and the eigenvalues of $\mathbf{H}_{\text{NRS}}$, μ_i, for different values of the regularization parameter, λ.

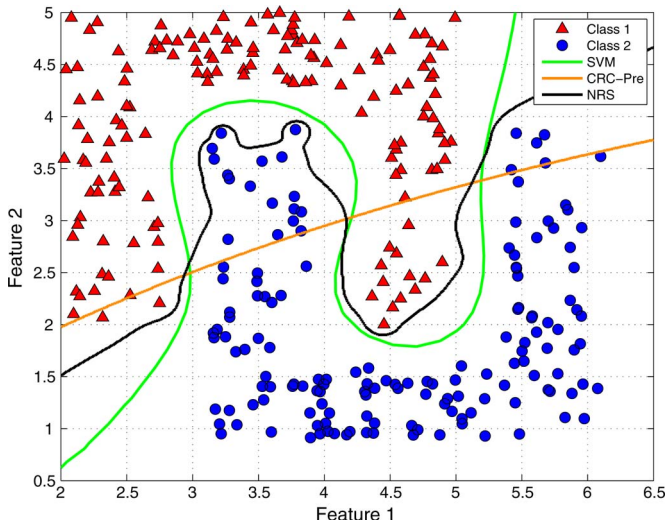

Fig. 3. Decision boundary determined using the NRS classifier on a two-class synthetic data set, calculated for $\lambda = 1$. The decision boundary for the SVM classifier using the radial-basis kernel is also shown.

Figs. 3 and 4 show the decision boundaries produced for two synthetic 2-D data sets using the proposed NRS as well

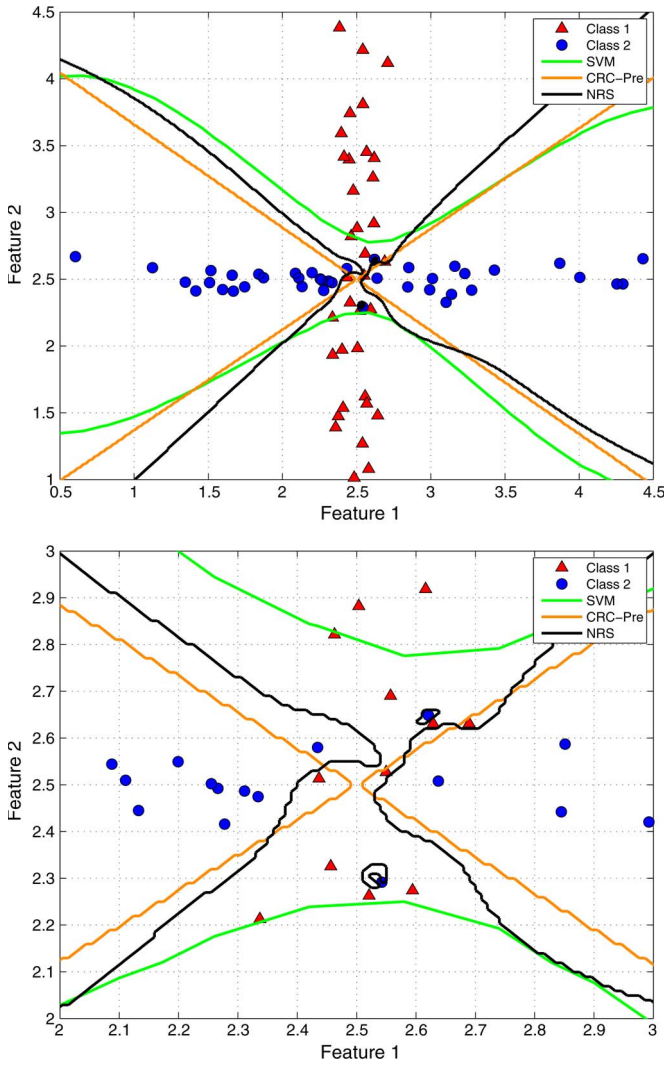

Fig. 4. (Top) Decision boundaries determined using the NRS classifier for $\lambda = 1$ and the SVM classifier using the radial-basis kernel for two synthetic normally distributed intersecting classes with common mean. (Bottom) Closer inspection of the class intersection.

as the SVM classifier using a radial-basis kernel and the CRC-Pre classifier. In both experiments, the data sets are not linearly separable and require complex boundaries for accurate classification. In Fig. 3, both the SVM and NRS classifiers produce a flexible boundary which accurately cuts between the two classes; however, the SVM boundary appears to be a more general fit, with the NRS boundary being more data dependent. The CRC-Pre classifier, however, cannot accurately distinguish between the two classes. In Fig. 4, we see two overlapping classes with shared means. Here, the NRS boundary performs better by cutting closer to the mean, reducing incorrect classification for samples generated from Class 1 near the mean, as compared to the SVM. The CRC-Pre classifier performs well in this environment, creating an almost linear set of decision boundaries between the two classes.

There are several differences between the proposed method and the previously discussed k-NN, SRC, and CRC techniques. First, the NRS classifier, unlike the k-NN classifier, does not limit its classification to the correspondence between testing samples and the provided training data alone. Instead, by forming an approximation from each class, NRS compares the testing sample with what can be considered to be an imaginary training sample which could have conceivably been drawn from the same process that produced the class training data provided. Secondly, the NRS classifier does not rely on time-consuming iterative sparse-recovery algorithms, as is the case with SRC and other such sparse techniques for classification. While recent investigations of sparse regularization have been of wide interest in signal processing in general, in this area at least, they do not seem to provide significant performance gains to outweigh their computationally expensive implementations. Lastly, while both NRS and CRC employ Tikhonov regularization to calculate class approximations, NRS cleaves to the traditional NS approach of prepartitioning and calculating class approximations independently while additionally employing a non-uniform shrinkage on the coefficients of $\boldsymbol{\alpha}_l$.

When constructing the biasing matrix $\boldsymbol{\Gamma}_{l,y}$ as in (11), we see that only the Euclidean distance between training and test samples is considered. In Section IV, it is demonstrated that this approach to biasing provides gains in classification accuracy for HSI data sets; however, it is well known that using Euclidean distances for very high-dimensional data can be an exercise in futility for certain data distributions. In the next section, we propose a method to alter the construction of $\boldsymbol{\Gamma}_{l,y}$ by using a generalized distance measure chosen to maximize class discrimination.

B. Dynamic Regularization for Classification

In Section III-A, we see that the proposed NRS classifier does not estimate or explicitly account for class probability distributions—instead it measures only the ability of each class to approximate a given target sample given a regularization parameter, λ. This regularization parameter is a significant factor in our proposed system, and, in fact, in all regularization-based techniques which make use of weighted-sum penalty functions. From Fig. 1, we can see that the setting of this parameter can also greatly affect classification accuracy. Both the SRC and CRC approaches offer little information on how this parameter should be set other than to suggest that cross-validation (CV) approaches could be used—splitting the training set into two parts and testing for a value which maximizes classification accuracy. However, the CV approach might not give an accurate estimation of the optimal λ when very few training samples are available, or might even be infeasible for extremely small training sets.

We propose to eliminate the need for CV estimations of λ by constructing a classifier which does not require fine-tuning of many side variables (for which classifiers such as SVM are notorious) at the cost of somewhat increased computation. We do this by making the observation that, in the case of classification, we are actually unconcerned with the accuracy of the approximations $\tilde{\mathbf{y}}_l$; rather, we want just that their proximities to $\mathbf{y}$ are such that they allow us to discriminate the class of $\mathbf{y}$ accurately.

To observe the behavior of the NRS classifier with respect to λ, a two-feature synthetic testing environment is considered in

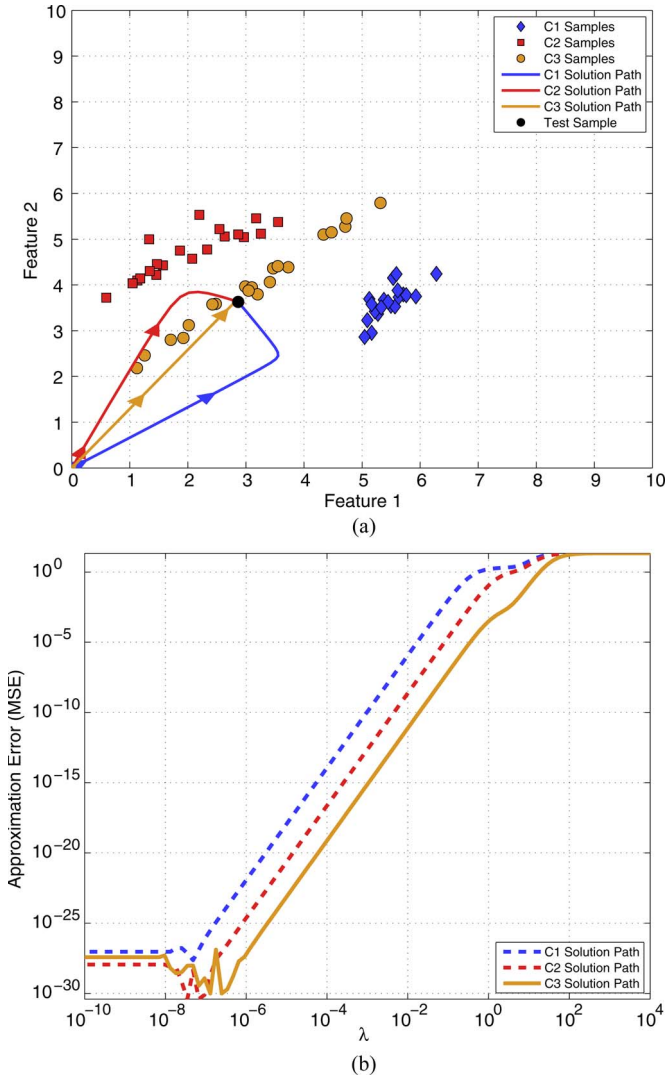

Fig. 5. Behavior of the NRS classifier for a synthetic three-class problem in two dimensions for a test sample drawn from class C3. (a) The 20 training samples per class and the solution paths for each class as λ decreases from 10^4 to 10^{-5}. (b) Per-class NRS-classifier approximation accuracy. Approximations generated by the true class (C3) are more accurate for $\lambda > 10^{-6}$.

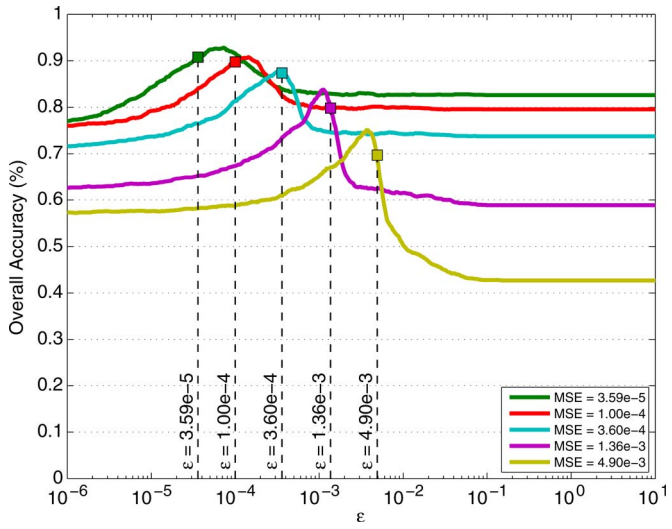

Fig. 6. Effect of noise on the optimal setting for ϵ in terms of overall classification accuracy of the NRS classifier for the Pavia Centre data set. Curves correspond to differing levels of noise corruption applied to the data set. The horizontal lines correspond to the value of ϵ which matches the level of noise in the data set.

Fig. 5. For this data set, all samples exist in only two dimensions, facilitating the visualization of the classifier behavior. Three classes of synthetic data randomly drawn from Gaussian distributions are created with a single test sample drawn from one of these three classes. By treating each approximation as a function of λ for a fixed training set and test sample, $\widetilde{\mathbf{y}}_l(\lambda)$, and by varying λ over a range of values (in this case 10^4 to 10^{-10}), a set of approximations over the domain of λ tested, which we term a solution path, is generated for each class.

Looking at the approximation accuracy of the solution paths in Fig. 5(b), an interesting phenomenon becomes apparent. For large values of λ, the regularization term $\|\mathbf{\Gamma}_{l,y}\boldsymbol{\alpha}_l\|_2^2$ becomes the dominant term in the cost function of (10), and the representations approach the zero vector to minimize this biased norm. However, for small λ, the representations approach the test sample, $\mathbf{y}$. Between these two modes, an inflection point occurs wherein the solution path rapidly changes direction. This feature is common to any minimization problems which utilize a weighted sum of cost functions as Tikhonov regularization does. For classes whose members best represent $\mathbf{y}$, this saddle point is less pronounced. For classes whose members are most dissimilar, the inflection point is quite pronounced, as the "initial" trajectories of these classes are oriented away from $\mathbf{y}$. However, the solution path created by the correct class tends to approach $\mathbf{y}$ more rapidly, i.e., the approximations for the third class, $\widetilde{\mathbf{y}}_3$, are more accurate for larger values of λ than the approximations generated by the other classes. The rapidity of convergence can be seen in Fig. 5.

We propose to use this feature to eliminate the need for setting a fixed value of λ prior to classification. We do this by setting a threshold, ϵ, on the approximation accuracy in terms of mean square error (MSE), $(1/d)\|\widetilde{\mathbf{y}}_l(\lambda) - \mathbf{y}\|_2^2$, and determining the classification based upon the first class to pass this threshold as λ is stepped from large to small values, causing the proposed method to resemble a "race" between the classes. From Fig. 5(b), we can see that, for this small scale demonstration, ϵ is a more robust parameter, as any choice within the range of $[10^{-25}, 10^0]$ would leave the classification unchanged. This is in contrast to the parameter λ, for which, in different test environments, small deviations from the optimal setting may degrade classification performance significantly.

Also, the addition of noise to the data set can cause the optimal choice for λ to shift away from *a priori* expected values. Instead of indirectly accounting for noise by adjusting λ, an approximation of the noise energy can be used to set ϵ directly. We demonstrate this in Fig. 6 wherein five different levels of zero-mean iid Gaussian noise were applied to the Pavia Centre data set. A wide range of possible values for ϵ were tested under these varying-noise conditions, and the overall classification accuracy is shown as a function of ϵ. The horizontal lines shown in Fig. 6 represent the values of ϵ which match the true noise levels tested. We can see from this chart that the peak classification accuracy for the range of tested ϵ at

each noise level corresponds closely with the overall accuracy achieved by setting ϵ to match the true noise level.

Additionally, if only a small number of training samples are available to drive the classification, the effectiveness of using CV approaches to estimate an optimal fixed setting for λ can be greatly diminished. Also, it is reasonable to assume that not every test sample requires the same value of λ to ensure correct classification. The proposed method accounts for the individuality of each test sample by sidestepping the need for a fixed λ at all, testing each sample's classification across a range of λ. Together, these features make dynamic regularization more robust than using a fixed λ and ensure stable classifier performance for the practitioner.

C. Enhancing Discrimination Power

One popular method of enhancing discrimination for hyperspectral classification is through LDA [8]. LDA projects from its natural, perhaps high-dimensional, space into a lower-dimensional subspace via a transform procedure aimed at maximizing between-class scatter while minimizing within-class scatter. Recently, an extension of LDA, locality Fisher's discriminant analysis (LFDA) [33], was proposed. LFDA combines the separability-enhancing power of LDA with locality-preserving projections (LPP) [34] to form a transformation, $\mathbf{L}$, which can handle multimodal non-Gaussian class distributions while preserving the local structure of the class distributions in the projected subspace.

In LFDA [13], the *affinity* between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $A_{i,j} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\gamma_i\gamma_j)$, where $\gamma_i = \|\mathbf{x}_i - \mathbf{x}_i^{(k_{\rm nn})}\|$ denotes the local scaling of data samples in the neighborhood of $\mathbf{x}_i$, and $\mathbf{x}_i^{(k_{\rm nn})}$ is the $k_{\rm nn}$-nearest neighbor of $\mathbf{x}_i$. The resulting *affinity matrix*, $\mathbf{A}$, is a symmetric matrix of size $n \times n$, which measures the distance among data samples. In LFDA, the *local* between-class, $\mathbf{S}^{(\rm lb)}$, and within-class, $\mathbf{S}^{(\rm lw)}$, scatter matrices are defined as

$$\mathbf{S}^{(\rm lb)} = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^{(\rm lb)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{24}$$

$$\mathbf{S}^{(\rm lw)} = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}^{(\rm lw)}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{25}$$

where $\mathbf{W}^{(\rm lb)}$ and $\mathbf{W}^{(\rm lw)}$ are $n \times n$ matrices defined as

$$\mathbf{W}_{i,j}^{(\rm lb)} = \begin{cases} A_{i,j}(1/n - 1/n_l), & \text{if } y_i = y_j = l \\ \frac{1}{n}, & \text{if } y_i \neq y_j \end{cases} \tag{26}$$

$$\mathbf{W}_{i,j}^{(\rm lw)} = \begin{cases} \frac{A_{i,j}}{n_l}, & \text{if } y_i = y_j = l \\ 0, & \text{if } y_i \neq y_j \end{cases} \tag{27}$$

where n_l is the number of available training samples for the $l^{\rm th}$ class, $\sum_{l=1}^{C} n_l = n$. This weight assignment provides an important benefit to the traditional LDA formulation—if a class-conditional probability distribution function is multi-modal, different modes will contribute to the scatter independently, thereby resulting in a more accurate representation of multi-modal data. This important neighborhood-preserving property ensures that local neighborhood relationships in the original space are retained in the projected subspace. LFDA obtains good between-class separation while preserving the within-class local structure simultaneously. The modified Fisher's ratio in LFDA employs these local scatter matrices to estimate the dimensionality-reduction projection as the solution, $\mathbf{L}$, to generalized eigenvalue problem, $\mathbf{S}^{(\rm lb)}\mathbf{L} = \mathbf{\Lambda}\mathbf{S}^{(\rm lw)}\mathbf{L}$. The reader is referred to [13], [33] for more details on LFDA.

In this paper,we define a generalized distance measure by comparing the distances between points within the projection space of $\mathbf{L}$, namely

$$\begin{aligned} D_{\rm LFDA}(\mathbf{x},\mathbf{y}) &= \|\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y}\|_2, \\ &= \sqrt{(\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y})^T(\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y})} \\ &= \sqrt{(\mathbf{x} - \mathbf{y})^T\mathbf{P}(\mathbf{x} - \mathbf{y})} \end{aligned} \tag{28}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors of $d \times 1$, $\mathbf{L}$ is a projection matrix of size $d' \times d$ (d' is the reduced dimensionality), $\mathbf{P} = \mathbf{L}^T\mathbf{L}$ is a symmetric positive matrix, and $D_{\rm LFDA}(\mathbf{x},\mathbf{y})$ is a single scalar. Using (28), we modify the construction of the biasing Tikhonov matrix of (11) to become

$$\mathbf{\Gamma}_{l,y} = \begin{bmatrix} D_{\rm LFDA}(\mathbf{y},\mathbf{x}_{l,1}) & & 0 \\ & \ddots & \\ 0 & & D_{\rm LFDA}(\mathbf{y},\mathbf{x}_{l,n_l}) \end{bmatrix}. \tag{29}$$

We refer to the classifier using this construction of $\mathbf{\Gamma}_{l,y}$ as NRS-LFDA. By comparing distance relationships within the LFDA-projected space, we gain two distinct advantages when biasing our Tikhonov regularization of (10). First, by reducing the dimensionality of the space in which distances are calculated, distances become more meaningful to the classification task, rather than having all distances be large. Second, the space is chosen in such a manner that inter-class separability is increased, further penalizing classes whose memberships lie mostly distant from the target point. Additionally, the LPP of LFDA means that samples which are truly neighbors of $\mathbf{y}$ are also seen as neighbors within the projected space. Without such locality preservation, calculating distances within a lower-dimensional space (such as that produced by LDA) might not give any information on within-class distance relationships with $\mathbf{y}$ and might offer little benefit in terms of classification accuracy. In the next section, we present results which demonstrate that the NRS-LFDA technique presented here does indeed improve classification accuracy as compared to the original NRS which uses Euclidean distances in the original space.

IV. Experimental Results

A. Experimental Hyperspectral Data

In this section, we demonstrate the effectiveness of the proposed NRS and NRS-LFDA classifiers on HSI data sets. The first HSI data set in our tests was acquired using NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)

TABLE I
PER-CLASS SAMPLES FOR TRAINING AND TESTING DATA FOR THE INDIAN PINES DATA SET

Class	*Training Samples*	*Testing Samples*
1	187	1247
2	187	647
3	187	310
4	187	302
5	187	781
6	187	2281
7	187	427
8	187	1107

Fig. 7. False-color image of the Indian Pines data set.

sensor and was collected over northwest Indiana's Indian Pines test site in June 1992.[1] The image represents a vegetation-classification scenario with 145 $\times$ 145 pixels and 220 spectral bands, post water-band removal, in the 0.4- to 2.45-μm region of the visible and infrared spectrum with a spatial resolution of 20 m. The two main crops, soybean and corn, shown in the HSI are in their early-growth stage. The notation *no till, min till*, and *clean till* indicate the amount of previous crop residue remaining. There are 16 different land-cover classes in the original ground truth; however, we conduct our experiments with eight classes, allowing for more training samples from a statistical viewpoint [35]. The eight classes used in our experiments are *Corn-no-till, Corn-min-till, Soybean-no-till, Soybean-min-till, Soybean-clean-till, Grass/Pasture, Hay-windowed*, and *Woods*. Approximately 8600 labeled pixels are employed to train and validate the efficacy of the proposed classification methods. The pixels chosen for validation were drawn randomly from the ground truth. This data is partitioned into approximately 1496 training pixels and 7102 testing pixels, with the training pixels randomly selected from the 8600 chosen validation samples. The class-specific number of training and testing samples are given in Table I. Additionally, a false-color representation of the HSI is given in Fig. 7.

[1] ftp://ftp.ecn.purdue.edu/biehl/MultiSpec

TABLE II
PER-CLASS SAMPLES FOR TRAINING AND TESTING DATA FOR THE UNIVERSITY OF PAVIA DATA SET

Class	*Training Samples*	*Testing Samples*
1	185	925
2	180	900
3	168	840
4	176	880
5	132	660
6	181	905
7	163	815
8	175	875
9	116	580

TABLE III
PER-CLASS SAMPLES FOR TRAINING AND TESTING DATA FOR THE PAVIA CENTRE DATA SET

Class	*Training Samples*	*Testing Samples*
1	165	990
2	164	984
3	165	990
4	162	972
5	164	984
6	163	978
7	162	972
8	189	1134
9	143	858

The other two HSI data sets used in this work were collected by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. The images, covering the city of Pavia, Italy, were collected under the HySens project managed by DLR (the German Aerospace Agency). The image has 115 spectral bands prior to water-band removal, with a spectral coverage from 0.43- to 0.86-μm and a spatial resolution of 1.3 m. Two scenes are used in our experiment. The first is the university area which has 103 spectral bands with a spatial coverage of 610 $\times$ 340 pixels. The second one is the Pavia city center which has 102 spectral bands with 1096 $\times$ 715 pixels formed by combining two separate images representing different areas of the Pavia city. The labeled ground truth of each data set is comprised of nine classes. The numbers of training and testing samples used for the University of Pavia data set are 1476 and 7380, respectively. The numbers of training and testing samples used for the Pavia Centre data set are 1477 and 8862, respectively. The selection of the validation samples for both data sets were chosen in the same manner as the Indian Pines data set. The numbers of training and testing samples for University of Pavia and Pavia Centre data sets are given in Table II and Table III, respectively. Also, false-color representations of the two data sets are given in Figs. 8 and 9.

B. Experiments

We compare our proposed methods with k-NN, SRC, CRC-Pre, SVM, and the recently proposed LFDA-SVM [13] classifiers. For the k-NN classifier, we find that $k = 3$ usually provides better classification performance compared to other values (such as 1, 5, 7, etc.). For SRC, we chose the parameter $\lambda = 0.01$ in our experiments. Additionally, we use the `l1_ls`[2]

[2] http://www.stanford.edu/boyd/l1_ls

Fig. 8. False-color image of the Pavia University data set.

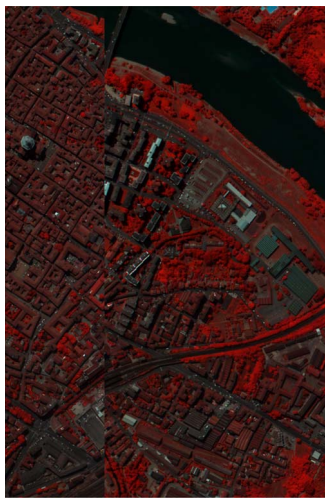

Fig. 9. False-color image of the Pavia Centre data set.

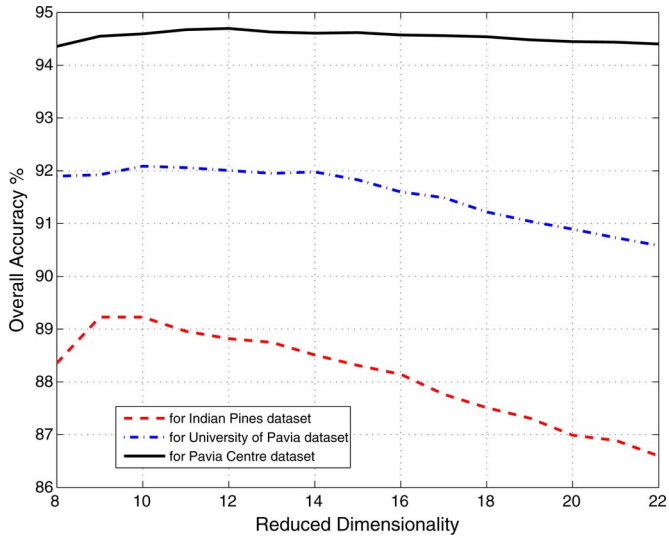

Fig. 10. Overall classification accuracy of three HSI data sets as a function of reduced dimensionality for the NRS-LFDA classifier.

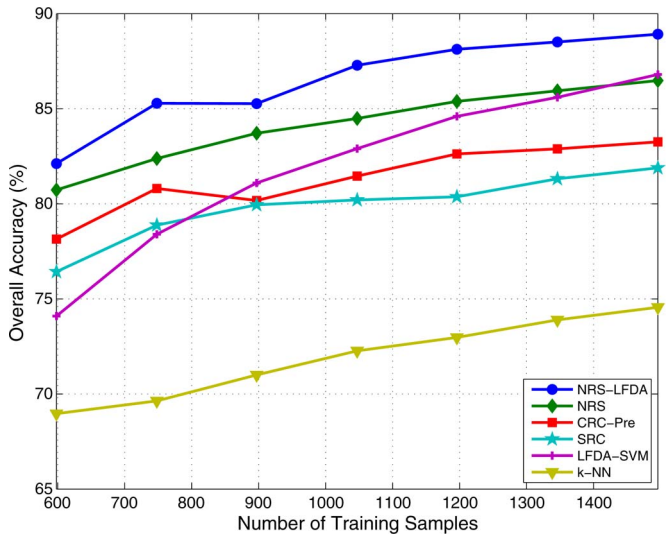

Fig. 11. Classification accuracy versus the number of training samples for the Indian Pines data set.

solver to calculate sparse approximations. We note that, while there exist a large number of sparse solvers suitable for SRC implementation, some of which are optimized for speed and others for representational accuracy, the classification accuracy of the SRC, in relation to the other methods tested, is only nominally affected. For CRC-Pre, the optimal parameter λ is 0.2 for the Indian Pines data set, 0.25 for the University of Pavia data set, and 0.6 for the Pavia Centre data set. The optimal parameters for SVM and LFDA-SVM can be found in [13]. To find a proper setting for the LFDA-projection dimension parameter, d', described in the previous section, the available training data was used to empirically gauge an effective range for d', as shown in Fig. 10. For NRS-LFDA, the dimensionality of LFDA is around 10 for the experimental data sets, and we found that it is not sensitive to sample size. Additionally, for both NRS and NRS-LFDA, a threshold of $\epsilon = 10^{-3}$ was used. In practical situations, the number of available training samples is often insufficient for each class. We illustrate the sensitivity of each classifier to the number of available training samples by testing over different percentages of the data set used for training while retaining the prior probability of each class. To avoid any bias, we randomly choose a subset of training samples for each sample-size value and repeat the experiment 10 times, reporting the average classification accuracy.

It is obvious from Fig. 11 that the proposed methods—the NRS and NRS-LFDA classifiers—outperform other approaches, especially under the small training-size classification scenario. The k-NN classifier has the worst classification accuracy, while SVM does not perform as well as either CRC-Pre or SRC for the cases of small training size. It is worthwhile mentioning that the NRS-LFDA classifier has, on average, 3% better accuracy than the NRS classifier and even greater improvements in accuracy over the other tested classifiers, which verifies that the discriminant-enhancing LFDA distance

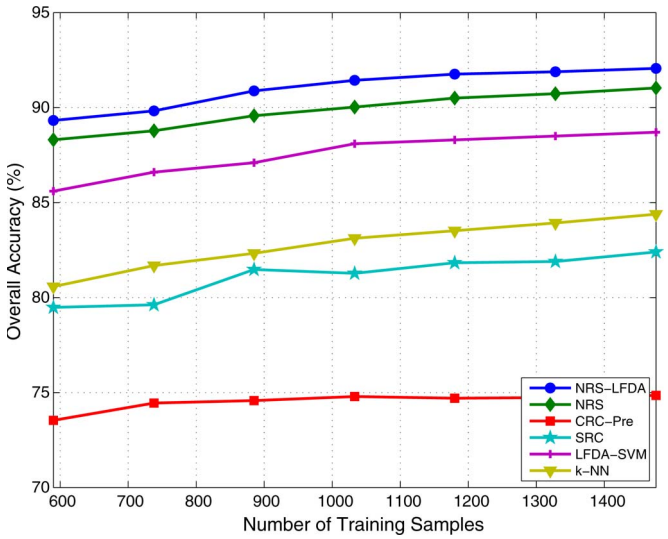

Fig. 12. Classification accuracy versus the number of training samples for the University of Pavia data set.

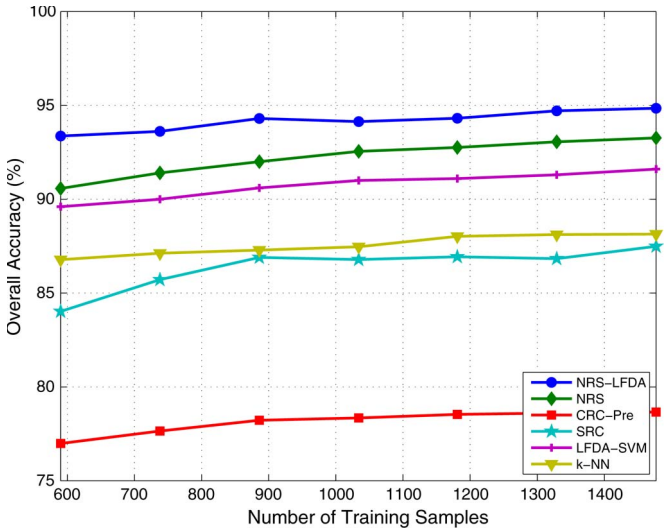

Fig. 13. Classification accuracy versus the number of training samples for the Pavia Centre data set.

metric works well for hyperspectral data. Figs. 12 and 13 show the overall accuracy as a function of number of training samples for the University of Pavia and Pavia Centre data sets, respectively. For these two Pavia data sets, SRC and CRC-Pre have unfavorable classification accuracies, even lower than k-NN. The proposed NRS-LFDA and NRS classifiers still provide the best classification accuracy of the tested classifiers for these data sets.

Fig. 14 provides a visual inspection of the classification maps generated using the whole HSI scene for the Indian Pines data set (145×145, including unlabeled pixels). To facilitate comparison between classification methods, only areas for which we have ground truth are shown in these maps. In Fig. 14, our proposed techniques show the best spatial homogeneity of the tested approaches. This homogeneity is most pronounced within the *Soybean-min till* and *Soybean-clean till* areas.

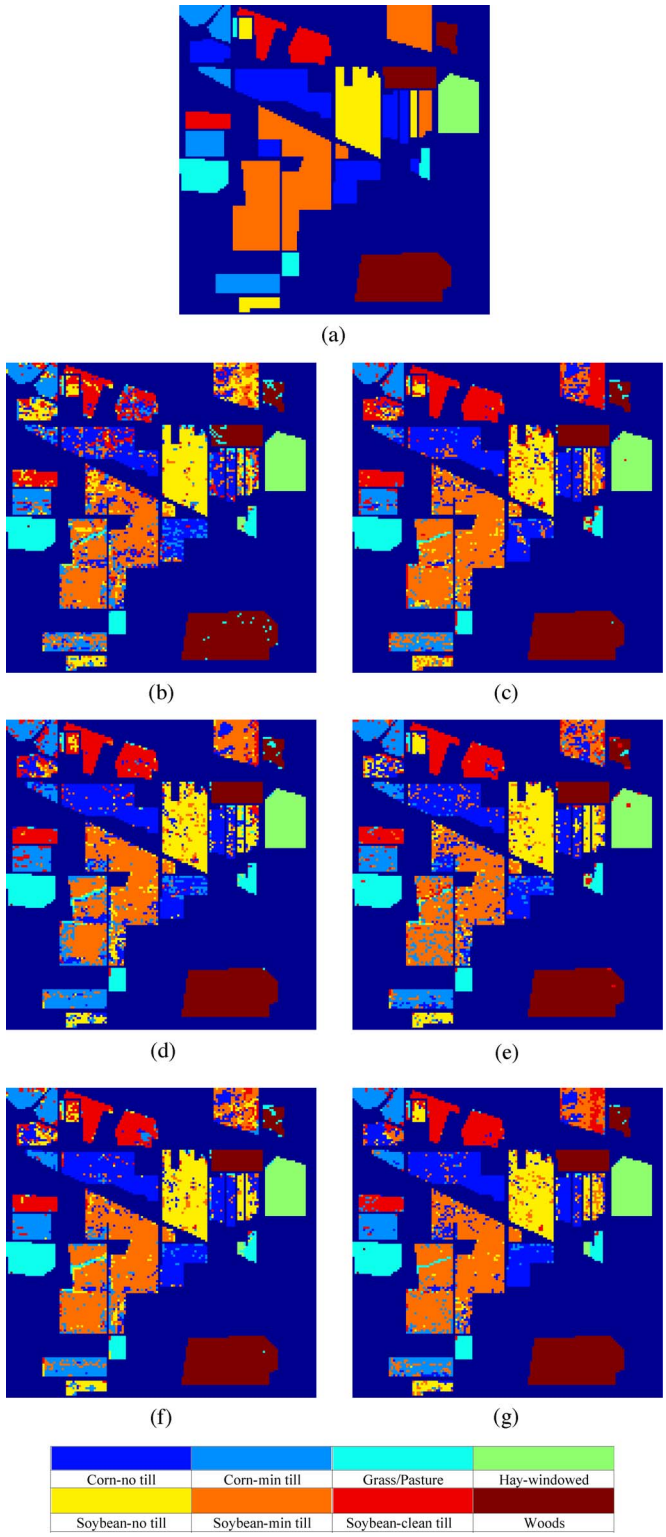

Fig. 14. Thematic maps resulting from classification using 748 training samples for the Indian Pines HSI data set. (a) Ground truth; (b) k-NN; (c) CRC-Pre; (d) SRC; (e) LFDA-SVM; (f) NRS; (g) NRS-LFDA.

Finally, we compare the computational complexity of the classification methods. All the experiments are carried out using MATLAB on a 3.2-GHz machine with 5.8 GB of RAM. As an example, the execution times (in seconds) to train and validate with the Indian Pine data set is shown in Table IV. We find

TABLE IV
EXECUTION TIME (IN SECONDS) TO TRAIN AND VALIDATE WITH THE INDIAN PINES DATA SET (748 SAMPLES FOR TRAINING, THE WHOLE SCENE FOR TESTING)

Algorithm	*Time (s)*
k-NN	24
CRC-Pre	132
NRS	2210
SVM	5364
LFDA-SVM	5367
NRS-LFDA	9633
SRC	23245

that the NRS classifier generally runs around 15 times slower than CRC-Pre, but around 10 times faster than SRC. Notice that both CRC-Pre and SRC require either prior information on the optimal parameter λ, or for a CV approach to be used to estimate this parameter. However, the NRS and NRS-LFDA classifiers do not require such fine tuning. If we were to provide the optimal λ for them, the execution time decreases accordingly (NRS: 135 s, NRS-LFDA: 346 s).

V. CONCLUSION

In this paper, we have presented a classification framework for hyperspectral data using a regularized nearest-subspace approach. For each class, an approximation of the testing sample was calculated via a linear combination of all training samples within the class. A distance-weighted Tikhonov regularization was used to calculate the linear combination of hypotheses in a stable manner. Furthermore, a discrimination-enhancing distance measure based on LFDA was proposed to improve the classification accuracy of the proposed NRS classifier. Additionally, a competitive strategy was introduced to avoid extensive parameter tuning via cross validation. Through our experiments on hyperspectral image data sets, the proposed NRS classifier and its variants provided superior classification performance with fewer training samples than traditional classification methods.

ACKNOWLEDGMENT

The authors thank P. Gamba for providing the ROSIS University of Pavia and Pavia Centre data sets. The authors would also like to thank the reviewers of this work for their valuable comments and suggestions.

REFERENCES

[1] D. A. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.
[2] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.
[3] F. A. Mainji and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2100–2112, Jun. 2011.
[4] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classifications," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3780–3791, Oct. 2010.
[5] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.
[6] L. Samaniego, A. Bárdossy, and K. Schulz, "Supervised classification of recovery sensed image using a modified k-nn technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2112–2125, Jul. 2008.
[7] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.
[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, USA: Wiley, 2001.
[9] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving discriminant analysis in kernel-induced feature spaces for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 894–898, Sep. 2011.
[10] S. Prasad and L. M. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448–1456, May 2008.
[11] S. Di Zenzo, R. Bernstein, S. D. Degloria, and H. C. Kolsky, "Gaussian maximum likelihood and contextual classification algorithms for multicrop classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 25, no. 6, pp. 805–814, Nov. 1987.
[12] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–677, Oct. 2007.
[13] W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce, "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, 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] 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.
[16] U. Srinivas, Y. Chen, V. Monga, N. M. Nasrabadi, and T. D. Tran, "Discriminative graphical models for sparsity-based hyperspectral target detection," in *Proc. Int. Geosci. Remote Sens. Symp.*, Munich, Germany, Jul. 2012, pp. 1489–1492.
[17] U. Srinivas, Y. Chen, V. Monga, N. M. Nasrabadi, and T. D. Tran, "Exploiting sparsity in hyperspectral image classification via graphical models," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 505–509, May 2013.
[18] 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.
[19] 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.
[20] Q. Sami ul Haq, L. Tao, F. Sun, and S. 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.
[21] A. Castrodad, Z. Xing, J. Greer, E. Bosch, L. Carin, and G. Sapiro, "Discriminative sparse representations in hyperspectral imagery," in *Proc. Int. Conf. Image Process.*, Hong Kong, China, Sep. 2010, pp. 1313–1316.
[22] Y. Gu and K. Feng, "L1-graph semisupervised learning for hyperspectral image classification," in *Proc. Int. Geosci. Remote Sens. Symp.*, Munich, Germany, Jul. 2012, pp. 1401–1404.
[23] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 471–478.
[24] E. W. Tramel and J. E. Fowler, "Video compressed sensing with multihypothesis," in *Proc. Data Compress. Conf.*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, USA, Mar. 2011, pp. 193–202.
[25] J. E. Fowler, S. Mun, and E. W. Tramel, "Block-based compressed sensing of images and video," *Found. Trends Signal Process.*, vol. 4, no. 4, pp. 297–416, Mar. 2012.
[26] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Aug. 1998.
[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 73, no. 3, pp. 273–282, Jun. 2011.
[29] J. Laaksonen, "Subspace classifiers in recognition of handwritten digits," Ph.D. dissertation, Helsinki Univ. Technol., Espoo, Finland, May 1997.

[30] Y. Liu, S. Ge, C. Li, and Z. You, " k-ns: A classifier by the distance to the nearest subspace," *IEEE Trans. Neural Netw.*, vol. 22, no. 8, pp. 1256–1268, Aug. 2011.
[31] R. Rigamonti, M. A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1545–1552.
[32] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC, USA: V. H. Winston & Sons, 1977.
[33] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, no. 5, pp. 1027–1061, May 2007.
[34] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2004.
[35] 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.

Wei Li (S'11–M'13) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-Sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent one year as a postdoctoral researcher at the University of California, Davis, CA, USA, and will soon join the College of Information Science and Technology at Beijing University of Chemical Technology, Beijing, China. His research interests include statistical pattern recognition, hyperspectral image analysis, and data compression.

Dr. Li is an active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE REMOTE SENSING LETTERS, and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Eric W. Tramel (S'08–M'13) received the B.S. and Ph.D. degrees in computer engineering from Mississippi State University, Mississippi State, MS, USA, in 2007 and 2012, respectively.

In 2011, he served as a Research Intern at Canon USA, Inc. From 2009 to 2012, he served as a Research Associate within the Geosystems Research Institute at Mississippi State. His research interests include compressed sensing, image and video coding, image and video multiview systems, data-compression, and pattern recognition.

Dr. Tramel is an active Reviewer for IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS.

Saurabh Prasad (S'05–M'09) received the B.S. degree in electrical engineering from Jamia Millia Islamia, New Delhi, India, in 2003, the M.S. degree in electrical engineering from Old Dominion University, Norfolk, VA, in 2005, and the Ph.D. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2008.

He is currently an Assistant Professor in the Electrical and Computer Engineering Department at the University of Houston (UH), Houston, TX, USA, and is also affiliated with UH's Geosensing Systems Engineering Research Center and the National Science Foundation-funded National Center for Airborne Laser Mapping. He is the Principal Investigator/Technicallead on projects funded by the National Geospatial-Intelligence Agency, National Aeronautics and Space Administration, and Department of Homeland Security. He was the Lead Editor of the book entitled *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, published in March 2011. His research interests include statistical pattern recognition, adaptive signal processing, and kernel methods for medical imaging, optical, and synthetic aperture radar remote sensing. In particular, his current research work involves the use of information fusion techniques for designing robust statistical pattern classification algorithms for hyperspectral remote sensing systems operating under low-signal-to-noise-ratio, mixed pixel, and small-training-sample-size conditions.

Dr. Prasad is an active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE REMOTE SENSING LETTERS and the Elsevier *Pattern Recognition Letters*. He was awarded the Geosystems Research Institutes Graduate Research Assistant of the Year award in May 2007, and the Office-of-Research Outstanding Graduate Student Research Award in April 2008 at Mississippi State University, Starkville. In July 2008, he received the Best Student Paper Award at IEEE International Geoscience and Remote Sensing Symposium 2008 held in Boston, MA, USA. In October 2010, he received the State Pride Faculty Award at Mississippi State University for his academic and research contributions.

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 from The Ohio State University, Columbus, OH, USA, in 1990, 1992, and 1996, respectively.

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