Intrinsic Representation of Hyperspectral Imagery For Unsupervised Feature Extraction

Linlin Xu, *Member, IEEE*,
Alexander Wong, *Member, IEEE*, Fan Li, David A. Clausi, *Senior Member, IEEE*

***Abstract*—Unsupervised feature extraction from hyperspectral images (HSI) relies on efficient data representation. However, classical data representation techniques, e.g., principal component analysis (PCA) and independent component analysis (ICA), do not reflect the intrinsic characteristics of HSI, and as such are less efficient for producing discriminative features. To address this issue, we have developed an intrinsic representation (IR) approach to support HSI classification. Based on the linear spectral mixture model (LSMM), the IR approach explains the underlying physical factors that are responsible for generating HSI. Moreover, it addresses other important characteristics of HSI, i.e., the noise variance heterogeneity effect in spectral domain and the spatial correlation effect in image domain. The IR model is solved iteratively by alternating the estimation of IR coefficients given IR bases and the update of IR bases given the coefficients. The resulting IR coefficients are discriminative, compact and noise-resistent, thereby constitute powerful features for improved HSI classification. The experiments on both simulated and real HSI demonstrate that the features extracted by the IR model are more capable of boosting the classification performance than the other referenced techniques.**

I. Introduction

By using hundreds of spectral bands of very narrow band width, hyperspectral imagery (HSI) is more capable of discriminating different ground targets than the other remote sensing techniques, and thereby providing a powerful measure for the classification of different land cover types. The existence of various spectral bands greatly increases the data dimensionality, thereby increasing the difficulty to perform classification. According to the Hughes phenomenon [1] [2], the increase in data dimensionality requires an exponential increase in the number of training samples. With limited training samples in remote sensing applications, the high dimensionality causes a huge training burden and eventually insufficient classification accuracy of supervised classification techniques [3] [4]. To address this problem, developing an informative data representation that can reduce the dimensionality while in the meantime preserve useful information for supporting classification is essential.

Different approaches can be used for hyperspectral feature extraction and dimension reduction. A widely used technique is the principal component analysis (PCA) method [5]. PCA assumes that the informative variables have large variance and are statistically uncorrelated with each other. Based on this assumption, PCA intends to find linear orthogonal subspaces where most data variance in HSI can be explained. Another popular feature extraction technique, the independent component analysis (ICA) has been adapted for HSI analysis [6]. Comparing with PCA, ICA seeks statistically independent signals using higher-order statistics, and, as such is more capable of capturing subtle material substances that are not sufficient to constitute reliable second-order statistics [6]. Another technique that improves PCA for remote sensing image analysis is the maximum noise fraction (MNF) method [7]. MNF seeks variables with the largest signal-to-noise ratio (SNR), and can be treated a noise-adjusted version of PCA [8]. The projection pursuit approaches have been designed for unsupervised feature extraction from HSI [9]–[11]. Nonlinear feature extraction technique, i.e., ISOMAP, has been applied to HSI [12]–[15]. The Morphological Transformation approach has been extended for dimensional reduction and classification of HSI using spatial-spectral information [16]. Other techniques include the lower rank tensor approximation [17] and minimum change rate deviation [18], accounting for the spatial correlation among neighboring pixels.

Most of the above-mentioned feature extraction techniques constitute constrained representations of HSI, which rely on some statistical criteria for defining the ”informativeness” of features. However, these criteria do not reflect the data generation mechanism of HSI, and are thereby not tailored to the characteristics of HSI for classification purpose. Recently, obtaining physically-meaningful features has been studied to support HSI classification [19]–[22]. These studies adopt some spectral unmixing techniques to obtain the abundance of endmembers (i.e., the spectra of the pure materials), which is afterward used as features for supervised HSI classification. These studies suggest that the physical features have interpretational advantage and can outperform the standard feature extraction techniques, e.g., PCA and ICA. In this paper, we extend this idea by presenting an intrinsic representation (IR) approach for unsupervised feature extraction. IR is essentially a generative model that seeks physically meaningful features based on the data generation model of HSI. Here, the word “intrinsic” is used to describe the innate endmember-abundance patterns of the spatial processes from a physical perspective, and IR therefore differs essentially from the intrinsic image decomposition (IID) approaches where the same word describes the reflectance, illumination and shading characteristics of the targets from a computer vision perspective [23], [24].

Following the linear spectral mixture model (LSMM) that is commonly used to describe the data generation process of HSI [25], IR expresses a HSI pixel as a nonnegative linear combination of some latent bases plus Gaussian noise. The latent bases in IR correspond to the endmembers, i.e., the spectra of the pure materials, whereas the nonnegative coefficients in IR correspond to the abundances of endmembers, i.e., the fractional contribution of individual endmembers. Based on this latent structure, IR is capable of disentangling the underlying factors that are responsible for HSI generation. Moreover, IR addresses other important characteristics of HSI, i.e., the correlation effects among pixels in spatial domain and the noise variance heterogeneity effect across spectral bands in spectral domain. The IR model is solved iteratively by alternating (1) the estimation of nonnegative coefficients given the latent bases and the (2) updating of latent bases given the coefficients.

Consequently, the nonnegative coefficients of latent bases in IR constitute desirable features for HSI classification. First, different land cover types tend to assume different material compositions. Since the IR features reflect the material composition in HSI pixels, they exploit this fundamental factor that distinguishes different classes in HSI. Second, the IR features maintain the spatial correlation effect in class labels, i.e., adjacent pixels in HSI tend to assume the same class label. Since IR features yield similar values for spatially-close pixels, they encourage close pixels to have the same class membership. Third, IR features can better resist the influence of noise. In IR, the noise of different spectral bands are explicitly estimated and separated to produce "noise-free" features. Last, the IR features can reduce data dimensionality. Since the number of latent bases in IR is generally fewer than the number of spectral bands in HSI, the use of IR features can greatly reduce the data dimensionality, thus reducing the risk of overfitting when performing supervised HSI classifications.

The rest of the paper is organized as follows. Section II describes the proposed IR method and the optimization schemes. Section III conducts experiments to compare the proposed IR method with several other feature extraction methods for supervised HSI classification. Section IV concludes the study.

II. Constrained Representation of HSI

In constrained representation of HSI, the observed spectral pixel stack $\mathbf{X} \in \mathbb{R}^{N\times P}$ is represented by the product of a latent bases matrix $\mathbf{A} \in \mathbb{R}^{P\times K}$ and a feature matrix $\mathbf{S} \in \mathbb{R}^{N\times K}$, plus some Gaussian noise $\mathbf{N} \in \mathbb{R}^{N\times P}$:

$$\underset{N\times P}{\mathbf{X}} = \underset{N\times K}{\mathbf{S}}\underset{K\times P}{\mathbf{A}^T} + \underset{N\times P}{\mathbf{N}} \tag{1}$$

$$\begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{s}_1^T \\ \boldsymbol{s}_2^T \\ \vdots \\ \boldsymbol{s}_N^T \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_K \end{pmatrix}^T + \begin{pmatrix} \boldsymbol{n}^T \\ \boldsymbol{n}^T \\ \vdots \\ \boldsymbol{n}^T \end{pmatrix} \tag{2}$$

where the $P \times 1$ dimensional spectral pixel $\boldsymbol{x}_i$ is associated with $K \times 1$ dimensional feature $\boldsymbol{s}_i$ in the subspace defined by $\mathbf{A}$:

$$\boldsymbol{x}_i = \mathbf{A}\boldsymbol{s}_i + \boldsymbol{n} \ (for\ i = 1, 2, ..., N) \tag{3}$$

and the noise distribution is assumed to satisfy a Gaussian model:

$$p(\boldsymbol{n}) = \frac{1}{\sqrt{(2\pi)^p\,|\boldsymbol{\Lambda}|}} exp\left(-\frac{1}{2}\boldsymbol{n}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{n}))\right) \tag{4}$$

Based on the generative model defined above, the task of feature extraction intends to obtain $\boldsymbol{s}_i$ $(for\ i = 1, 2, ..., N)$ by minimizing the representational error subject to some constraints:

$$\hat{\mathbf{S}} = arg \min_{\mathbf{A},\mathbf{S},\boldsymbol{\Lambda}} \sum_i dist(\boldsymbol{x}_i, \mathbf{A}\boldsymbol{s}_i) \ \ s.t.\ C(\mathbf{S}) \tag{5}$$

where $dist(\boldsymbol{x}_i, \mathbf{A}\boldsymbol{s}_i)$ represents the distance between $\boldsymbol{x}_i$ and $\mathbf{A}\boldsymbol{s}_i$ based on a particular distance measure, and $C(\mathbf{S})$ denotes the constraint imposed on $\mathbf{S}$.

The constraint on $\mathbf{S}$ defines the "informativeness" of the extracted features. Different constraints lead to different types of features. For example, to achieve uncorrelated features, PCA forces variables in $\mathbf{S}$ to satisfy a Gaussian distribution [26]. To achieve statistically independent features, ICA requires variables in $\mathbf{S}$ to satisfy super-Gaussian or sub-Gaussian distributions [27]. However, these constraints define the "informativeness" of features from statistical perspectives, and could not reflect the intrinsic characteristics of HSI. According to LSMM, the physical quantities hidden in HSI are nonnegative. Therefore, in the proposed IR approach, the terms in $\mathbf{S}$ are constrained to be nonnegative to promote physically-meaningful features for HSI classification. Moreover, some other constraints are also adopted in IR to address other important characteristics of HSI, as detailed in section III.

III. Intrinsic Representation of HSI

The proposed IR of HSI is based on the same generative model defined in (3), which can be reformulated as:

$$\boldsymbol{x}_i = \sum_{k=1}^{K} \boldsymbol{a}_k s_{ik} + \boldsymbol{n} \tag{6}$$

Accordingly to LSMM, $\{\boldsymbol{a}_k |\ for\ k = 1, 2, ..., K\}$ represents the collection of endmembers that are responsible for generating HSI, and s_{ik} is the abundance of $\boldsymbol{a}_k$ on spectral pixel $\boldsymbol{x}_i$. In LSMM, the abundances are required to be nonnegative and sum-to-one, i.e., $\forall s_{ik} \geqslant 0$ and $\sum_k s_{ik} = 1$. However, we only adopt the nonnegative constraint to simplify the model optimization. Besides the nonnegative constraint, other constraints are also adopted to account for other important characteristics of HSI, leading to the following objective function:

$$\begin{aligned} \hat{\mathbf{S}} = arg \min_{\{\mathbf{A},\mathbf{S},\boldsymbol{\Lambda}\}} \sum_i (\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i)^T \boldsymbol{\Lambda}^{-1} (\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i) \\ s.t.\ \forall s_{ik} \geqslant 0 \ \ and \ \ \mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \mathbf{A}\boldsymbol{s}_j / M = 0 \end{aligned} \tag{7}$$

where $\Psi(i)$ represents the neighborhood centered at site i that involves M neighboring pixels $\{\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, ..., \boldsymbol{x}_{j_M}\}$.

In (7), $(\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i)^T \boldsymbol{\Lambda}^{-1} (\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i)$ measures the representation error, where $\boldsymbol{\Lambda}$ is expressed as a diagonal matrix:

$$\mathbf{\Lambda} = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_P^2 \end{pmatrix} \tag{8}$$

in which σ_i^2 is the noise variance of the ith band. In HSI, σ_i^2 of different bands are not necessarily equal considering the varying physical properties of different spectral bands as well as the existence of junk bands [28]. This phenomenon is called noise heterogeneity effect in this paper. Using $\mathbf{\Lambda}$ defined in (8) allows accommodating such effect.

In (7), $\mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \mathbf{A}\boldsymbol{s}_j/M = 0$ is the spatial smoothness constraint. This constraint encourages adjacent pixels to have, in average, similar values in feature space. Since spatially-close pixels in HSI have higher possibility of belonging to the same class, the spatial smoothness constraint adopted here is supposed to enhance the discriminative capability of features by yielding similar feature values for adjacent pixels.

The unknown parameters in IR includes $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Lambda}$. Since the number of unknown parameters is more than the number of observations, the objective function defined in (7) constitutes an ill-posed optimization problem. In section IV, the model is optimized iteratively by alternating the estimation of $\mathbf{S}$ and $\mathbf{\Lambda}$ given $\mathbf{A}$, as detailed in section IV-A, and the update of $\mathbf{A}$ given $\mathbf{S}$, as detailed in section IV-B. The final estimate of $\mathbf{S}$ will be used to feed classifiers for HSI classification.

IV. Model Optimization

This section details the iterative optimization approach for solving the objective function defined in (7). We first introduce the estimation of $\mathbf{S}$ Given $\mathbf{A}$ in section IV-A, then describe the update of $\mathbf{A}$ based on $\mathbf{S}$ in section IV-B. We finally provide a summary of the proposed algorithm.

A. Estimate $\boldsymbol{s}$ Given $\mathbf{A}$

Based on $\mathbf{A}$, the elements $\boldsymbol{s}_i$ $(for\ i = 1, 2, ..., N)$ in $\mathbf{S}$ are estimated separately. Since the final algorithm is built up simple ones, below, we illustrate the estimation of $\boldsymbol{s}_i$ progressively by addressing objective functions with increasing complexity.

1) Least Squares: Without any constraints, the estimation of $\boldsymbol{s}_i$ given $\mathbf{A}$ can be achieved by minimizing the representational error:

$$\hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \|\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i\|_2^2 \tag{9}$$

which can be solved by the classical least squares approach:

$$\hat{\boldsymbol{s}}_i = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\boldsymbol{x}_i) \tag{10}$$

2) Nonnegative Least Squares: The solution provided by (10) may lead to negative values in $\hat{\boldsymbol{s}}_i$, which, however, is required to contain only nonnegative values according to LSMM. As a result, $\hat{\boldsymbol{s}}_i$ should be estimated by optimizing the following objective function with nonnegative constraint:

$$\hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \|\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i\|_2^2 \ \ s.t.\ \forall s_{ik} \geqslant 0 \tag{11}$$

Due to the nonnegative constraint, there is no known analytical solution to (11). Nevertheless, (11) can be solved by an iterative approach called active-set proposed by Lawson and Hanson in [29] and modified by Bro and Jong for fast computation [30]. The fast active set algorithm is summarized in Algorithm 1.

Algorithm 1 NNLS

Input: $\boldsymbol{x} = \boldsymbol{x}_i$, $\mathbf{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_K)$

Output: $\boldsymbol{s} = (s_{i1}, s_{i2}, ..., s_{iK})^T$

Initialization: two complementary indices sets $P = \varnothing$ and $Z = \{1, 2, ..., K\}$, $\boldsymbol{s} = \mathbf{0}$, $\boldsymbol{w} = \mathbf{A}^T(\boldsymbol{x} - \mathbf{A}\boldsymbol{s})$

```
while $Z \neq \varnothing\ and\ max(\boldsymbol{w}^Z > tol1)$ do
    $t =\ index\ of\ max(\boldsymbol{w}^Z)\ in\ \boldsymbol{w}$
    add $t$ to $P$, and remove $t$ from $Z$
    $\boldsymbol{g}^P = [(\mathbf{A}^T\mathbf{A})^P]^{-1}(\mathbf{A}^T\boldsymbol{b})^P$
    $\boldsymbol{g}^Z = \mathbf{0}$
    while $min(\boldsymbol{g}^P) < tol1$ do
        $\alpha = min(\boldsymbol{s}_k/(\boldsymbol{s}_k - \boldsymbol{g}_k)\ for\ i\ in\ P)$
        $\boldsymbol{s} = \boldsymbol{s} + \alpha(\boldsymbol{g} - \boldsymbol{s})$
        $Q = indices\ in\ \boldsymbol{s}\ where\ abs(\boldsymbol{s}^P) < tol1$
        add $Q$ to $Z$
        remove $Q$ from $P$
        $\boldsymbol{g}^P = [(\mathbf{A}^T\mathbf{A})^P]^{-1}(\mathbf{A}^T\boldsymbol{b})^P$
        $\boldsymbol{g}^Z = \mathbf{0}$
    end while
    $\boldsymbol{s} = \boldsymbol{g}$
    $\boldsymbol{w} = \mathbf{A}^T(\boldsymbol{x} - \mathbf{A}\boldsymbol{s})$
end while
```

Note: $\mathbf{I}^J$ is restricted to the row and column variables of $\mathbf{I}$ that are included in indices set J.

3) Weighted Least Squares: However, (9) and (11) assume that noise variables in $\boldsymbol{n}$ are Gaussian distributed. To adjust the noise heterogeneity effect across spectral bands, we adopt the diagonal covariance matrix in (8). Accordingly, (9) is reformulated to accommodate the noise heterogeneity effect:

$$\hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} (\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i)^T \mathbf{\Lambda}^{-1} (\boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i) \tag{12}$$

which can be reformulated as:

$$\hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \left\|\mathbf{\Lambda}^{-0.5}\boldsymbol{x}_i - \mathbf{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_i\right\|_2^2 \tag{13}$$

and (13) can be solved by a weighted least squares approach:

$$\hat{\boldsymbol{s}}_i = (\mathbf{A}^T\mathbf{\Lambda}^{-1}\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{\Lambda}^{-1}\boldsymbol{x}_i) \tag{14}$$

4) Weighted Nonnegative Least Squares: Similarly, the objective function in (11) is reformulated using $\mathbf{\Lambda}$ to accommodate the noise heterogeneity effect:

$$\hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \left\|\mathbf{\Lambda}^{-0.5}\boldsymbol{x}_i - \mathbf{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_i\right\|_2^2 \ \ s.t.\ \forall s_{ik} \geqslant 0 \tag{15}$$

Note that NNLS algorithm defined in Algorithm 1 assumes i.i.d. noise, and thereby does not apply to (15). To adapt NNLS for addressing the noise heterogeneity effect, we present a weighted NNLS (WNNLS) algorithm, defined in Algorithm 2, where the $\mathbf{A}$ and $\boldsymbol{x}_i$ are adjusted using $\boldsymbol{\Lambda}$ before being used as input to the NNLS algorithm.

```
Algorithm 2 WNNLS
Input: $\boldsymbol{x} = \boldsymbol{x}_i$, $\mathbf{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_K)$, $\boldsymbol{\Lambda}^{-0.5}$
Output: $\boldsymbol{s} = (s_{i1}, s_{i2}, ..., s_{iK})^T$
Initialization: $\boldsymbol{x}_w = \boldsymbol{\Lambda}^{-0.5}\boldsymbol{x}$ and $\mathbf{A}_w = \boldsymbol{\Lambda}^{-0.5}\mathbf{A}$,
$\boldsymbol{s} = NNLS(\boldsymbol{x}_w, \mathbf{A}_w)$
```

5) Weighted Nonnegative Least Squares with Spatial Smoothness Constraint: Although the objective function defined by (15) can provide a sound estimation of abundances $\boldsymbol{s}_i$ by addressing the nonnegative underlying factors and the noise heterogeneity effect, it does not account for the spatial correlation effect, due to which spatially-close pixels tend to have similar abundance patterns. To further improve IR for addressing the spatial correlation effect, we add to (15) the following spatial smoothness constraint:

$$\mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \mathbf{A}\boldsymbol{s}_j/M = 0 \tag{16}$$

where $\Psi(i)$ represents the neighborhood centered at site i that involves M neighboring pixels $\{\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, ..., \boldsymbol{x}_{j_M}\}$. This constraint encourages adjacent pixels to have similar abundance values.

The constraint in (16) is combined with the objective function defined in (15), leading to a complete representational framework that considers both noise heterogeneity effect in spectral domain and the spatial correlation effects in spatial domain:

$$\begin{aligned} \hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \left\|\boldsymbol{\Lambda}^{-0.5}\boldsymbol{x}_i - \boldsymbol{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_i\right\|_2^2 \\ s.t.\ \forall s_{ik} \geqslant 0\ and\ \mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \mathbf{A}\boldsymbol{s}_j/M = 0 \end{aligned} \tag{17}$$

The optimization problem defined by (17) can be reformulated as:

$$\begin{aligned} \hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i} \left\|\boldsymbol{\Lambda}^{-0.5}\boldsymbol{x}_i - \boldsymbol{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_i\right\|_2^2 \\ +\lambda \left\|\boldsymbol{\Omega}^{0.5}\mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \boldsymbol{\Omega}^{0.5}\mathbf{A}\boldsymbol{s}_j/M\right\|_2^2 \quad s.t.\ \forall s_{ik} \geqslant 0 \end{aligned} \tag{18}$$

where λ determines the overall weight of smoothness constraint relative to the representational error, and $\boldsymbol{\Omega}$ is a diagonal matrix, assigning different smoothness weights to different spectral bands.

Note that in (18), $\{\boldsymbol{s}_j | j \in \Psi(i)\}$ are also unknown. Therefore, solving (18) requires simultaneously estimating both $\boldsymbol{s}_i$ and $\{\boldsymbol{s}_j | j \in \Psi(i)\}$. Note that the estimated values of $\{\boldsymbol{s}_j\}$ are not used. It is achieved based on the following objective function.

$$\begin{aligned} \hat{\boldsymbol{s}}_i = arg \min_{\boldsymbol{s}_i, \{\boldsymbol{s}_j\}} \left\|\boldsymbol{\Lambda}^{-0.5}\boldsymbol{x}_i - \boldsymbol{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_i\right\|_2^2 \\ + \sum_{j\in\Psi(i)} \left\|\boldsymbol{\Lambda}^{-0.5}\boldsymbol{x}_j - \boldsymbol{\Lambda}^{-0.5}\mathbf{A}\boldsymbol{s}_j\right\|_2^2 \\ +\lambda \left\|\boldsymbol{\Omega}^{0.5}\mathbf{A}\boldsymbol{s}_i - \sum_{j\in\Psi(i)} \boldsymbol{\Omega}^{0.5}\mathbf{A}\boldsymbol{s}_j/M\right\|_2^2 \quad s.t.\ \forall s \geqslant 0 \end{aligned} \tag{19}$$

which can be reformulated as:

$$\hat{\bar{\boldsymbol{s}}}_i = arg \min_{\boldsymbol{s}_i, \{\boldsymbol{s}_j\}} \left\|\bar{\boldsymbol{\Lambda}}^{-0.5}\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{\Lambda}}^{-0.5}\bar{\mathbf{A}}\bar{\boldsymbol{s}}_i)^T\right\|_2^2 \quad s.t.\ \forall\bar{s} \geqslant 0 \tag{20}$$

$$where\ \bar{\boldsymbol{x}}_i = \underbrace{\begin{pmatrix} \mathbf{0} \\ \boldsymbol{x}_i \\ \boldsymbol{x}_{j_1} \\ \vdots \\ \boldsymbol{x}_{j_M} \end{pmatrix}}_{P(M+2)\times 1},\quad \bar{\boldsymbol{s}}_i = \underbrace{\begin{pmatrix} \boldsymbol{s}_i \\ \boldsymbol{s}_{j_1} \\ \vdots \\ \boldsymbol{s}_{j_M} \end{pmatrix}}_{K(M+1)\times 1}, \tag{21}$$

$$\bar{\mathbf{A}} = \underbrace{\begin{pmatrix} M\lambda\mathbf{A} & -\lambda\mathbf{A} & \cdots & -\lambda\mathbf{A} \\ \mathbf{A} & & & \\ & \mathbf{A} & & \\ & & \ddots & \\ & & & \mathbf{A} \end{pmatrix}}_{P(M+2)\times K(M+1)} \tag{22}$$

$$and\ \bar{\boldsymbol{\Lambda}}^{-0.5} = \underbrace{\begin{pmatrix} \boldsymbol{\Omega}^{0.5} & & & & \\ & \boldsymbol{\Lambda}^{-0.5} & & & \\ & & \boldsymbol{\Lambda}^{-0.5} & & \\ & & & \ddots & \\ & & & & \boldsymbol{\Lambda}^{-0.5} \end{pmatrix}}_{P(M+2)\times P(M+2)} \tag{23}$$

$$with\ \boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & & & \\ & \omega_2 & & \\ & & \ddots & \\ & & & \omega_P \end{pmatrix} \tag{24}$$

As we can see, (20) shares a similar form with the objective function defined in (15), and can therefore be solved by the WNNLS algorithm in Algorithm 2. Accordingly, we provide in Algorithm 3 a WNNLS with spatial smoothness constraint (WNNLS-SSC) algorithm for solving (20) and (17).

```
Algorithm 3 WNNLS-SSC
Input: $\mathbf{X}$, $\mathbf{A}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Omega}$, $\lambda$
Output: $\mathbf{S}$
Initialization: obtain $\bar{\mathbf{A}}$ using (22) and $\bar{\boldsymbol{\Lambda}}^{-0.5}$ using (23)
for $i = 1 : N$ do
    obtain $\bar{\boldsymbol{x}}_i$ using (21)
    $\bar{\boldsymbol{s}}_i = WNNLS(\bar{\boldsymbol{x}}_i, \bar{\mathbf{A}}, \bar{\boldsymbol{\Lambda}}^{-0.5})$
end for
```

There are two types of weighting parameters, i.e., λ and ω_p $(for\ p = 1, 2, ..., P)$, which together determine the final degree of spatial smoothness constraints. The parameter λ determines the relative weights between representational error and overall smoothness constraints on all bands in HSI. Bigger λ values promote more homogeneous results, leading to smoother abundances values $\mathbf{S}$ in spatial domain. In contrast, smaller λ values favor solutions with smaller representational error, resulting in noisy appearance of abundances values in image space. Therefore, the value of λ determines the trade-off between signal preservation and noise removal.

Although λ can determine the overall weight of smoothness constraint relative to the representational error, it can not determine the relative weights among different spectral bands. The band-wise spatial smoothness weights ω_p $(for\ p = 1, 2, ..., P)$ are adopted for this purpose. Since different bands tend to assume different noise levels, we would like to impose larger ω value to bands with higher noise level to better resist the influence of noise, but give smaller ω value to bands with lower noise levels to more efficiently capture the signal in HSI.

Therefore, ω_p should be adjusted based on the noise level of the pth band, being proportional to noise variance σ_p^2 defined in (8). In this paper, we adopt the following assignment of ω_p:

$$\omega_p = \frac{P\sigma_p^2}{\sum_{h=1}^{P} \sigma_h^2} \tag{25}$$

Another important parameter is M, the size of neighborhood defined in (16), which determines the scale of spatial correlation effect. Generally speaking, bigger neighborhood with larger M should be adopted if the noise level is high, or the scene complexity is low. In practice, we found that small neighborhood of 3-by-3 or 5-by-5 is sufficient for resisting the noise influence without compromising significantly the capability to capture scene signals.

6) Iterative Weighted Nonnegative Least Squares with Spatial Smoothness Constraint: Although the WNNLS-SSC algorithm given in Algorithm 3 can be employed to solve the objective function with spatial smoothness constraint defined in (20), it relies on known noise variances of different bands in $\mathbf{\Lambda}$, which however are generally unknown in practice. To provide a solution to (20) in the scenarios where we have no prior knowledge concerning the noise characteristics of different bands, we introduce an iterative weighted NNLS with spatial smoothness constraint (IWNNLS-SSC) as detailed in Algorithm 4, which can adaptively estimate the band-dependent noise variances in HSI, and use them for solving (20).

Based on IWNNLS-SSC, the IR features $\mathbf{S}$ can be extracted by solving a well-defined generative model that captures key characteristics of HSI, i.e., physical explanatory variables, the spatial correlation effect, as well as the noise heterogeneity effect. Since IWNNLS-SSC is robust to noise influence, and is able to estimated noise variances, it may also be appropriate for HSI denoising that aims to separate signal from noise.

Algorithm 4 IWNNLS-SSC

```
Input: $\mathbf{X}$, $\mathbf{A}$, $\mathbf{\Omega}$, $\lambda$
Output: $\mathbf{S}$
Initialization: $\mathbf{\Lambda} = \mathbf{I}$, obtain $\bar{\mathbf{A}}$ using (22) and $\bar{\mathbf{\Lambda}}^{-0.5}$ using (23)
while $iter < maxIter1$ do
    for $i = 1 : N$ do
        obtain $\bar{\boldsymbol{x}}_i$ using (21)
        $\bar{\boldsymbol{s}}_i = WNNLS(\bar{\boldsymbol{x}}_i, \bar{\mathbf{A}}, \bar{\mathbf{\Lambda}}^{-0.5})$
        $\boldsymbol{r}_i = \boldsymbol{x}_i - \mathbf{A}\boldsymbol{s}_i\ for\ i = 1, 2, ..., N$
    end for
    update $\mathbf{\Lambda}$ using $\{\boldsymbol{r}_i\}$, by assigning $\mathbf{\Lambda} = VAR(\{\boldsymbol{r}_i\})$
    update $\mathbf{\Omega}$ using $\mathbf{\Lambda}$, based on (25)
    update $\bar{\mathbf{\Lambda}}^{-0.5}$ using $\mathbf{\Lambda}$ and $\mathbf{\Omega}$, based on (23)
end while
```

B. Update $\mathbf{A}$ *based on* $\mathbf{S}$

In section IV-A, $\mathbf{A}$ is assumed to be known in advance to estimate $\mathbf{S}$ using IWNNLS-SSC. Once $\mathbf{S}$ is obtained, it can be used to improve $\mathbf{A}$. The key issue is how to efficiently use the information in $\mathbf{S}$ to update $\mathbf{A}$.

Since $\mathbf{A}$ corresponds to hyperspectral endmembers, some hyperspectral endmember extraction approaches, e.g., VCA [31] and MVC-NMF [32]), may be applicable. However, these approaches rely on extra criteria for regulating the endmember estimation.

In our previous publication [33], we estimate $\boldsymbol{a}_k$ as the mean value of the "purified" pixels that are due to the sole contribution of the kth endmember. This approach utilizes the information in $\mathbf{S}$, and can be better integrated into the current iterative optimization framework. Nevertheless, the K-P-Means algorithm [33] estimates $\boldsymbol{a}_k$ using only pixels in the kth class, and relies on the estimation of hard/discrete class membership by identifying the dominant endmembers for each pixel. It therefore could not identify the endmember that does not admit presence on any pixels in HSI.

Here, we estimate the endmembers using a "soft" version of the K-P-Means approach, in which $\boldsymbol{a}_k$ is estimated as the weighted means of all pixels in HSI, with the weight determined by the abundance values. There are two steps in estimating $\boldsymbol{a}_k$: (1) purify all pixels in HSI by removing the contribution of other endmembers $\{\boldsymbol{a}_t |\ t \neq k\}$; (2) estimate $\boldsymbol{a}_k$ as the weighted means of purified pixels.

1) Purify pixels in HSI: The first step is to purify all the pixels to obtain the sole contribution of $\boldsymbol{a}_k$ in each pixel. Comparing with mixed pixel $\boldsymbol{x}_i$ that involves the linear contribution of multiple endmembers, the purified pixel denoted by $\boldsymbol{y}_i^k$ contains only the contribution of the kth endmember $\boldsymbol{a}_k$. According to the LSMM, $\boldsymbol{y}_i^k$ can be achieved by removing from $\boldsymbol{x}_i$ the contribution of all other endmembers $\{\boldsymbol{a}_t |\ t \neq k\}$:

$$\boldsymbol{y}_i^k = \boldsymbol{x}_i - \sum_{t \neq k}^{K} \boldsymbol{a}_t s_{it} \tag{26}$$

2) Estimate endmember using purified pixels: Given $\{\boldsymbol{y}_i^k\}$ associated with $\boldsymbol{a}_k$, the final step is to estimate $\boldsymbol{a}_k$ using these purified pixels. Since $\{\boldsymbol{y}_i^k\}$ are only due to the contribution of $\boldsymbol{a}_k$ subject to different weights and random noise, they can be used to achieve a weighted mean estimate of $\boldsymbol{a}_k$:

$$\hat{\boldsymbol{a}}_k = \frac{\sum_{i=1}^{N} \boldsymbol{y}_i^k s_{ik}}{\sum_{i=1}^{N} s_{ik}} \tag{27}$$

where the weight s_{ik} is the abundance of $\boldsymbol{a}_k$ on the ith pixel. The update of $\mathbf{A}$ is achieved by updating $\{\boldsymbol{a}_k | k = 1, 2, ..., K\}$ sequentially according to the above steps.

C. Summary of Complete Algorithm

Based on the two key optimization steps described respectively in sections IV-A and IV-B, a complete algorithm used for solving IR can be achieved by iteratively alternating these two steps. Accordingly, we provide in Algorithm 5 a summary of the IR optimization steps.

Algorithm 5 IR

```
Input: $\mathbf{X}$, $\mathbf{A}$, $\mathbf{\Omega}$, $\lambda$
Output: $\mathbf{S}$
Initialization: $\mathbf{A} \leftarrow$ randomly pick up $K$ pixels from $\{\boldsymbol{x}_i\}$
while $iter < maxIter2$ do
    update $\mathbf{S}$ =IWNNLS-SSC($\mathbf{X}, \mathbf{A}, \mathbf{\Omega}, \lambda$)
    for $k = 1, 2, ..., K$ do
        update $\boldsymbol{a}_k$ using $\mathbf{S}$, $\mathbf{X}$ and $\{\boldsymbol{a}_t |\ t \neq k\}$, as described in section IV-B
    end for
end while
```

In Algorithm 5, the update of $\boldsymbol{a}_k$ relies on the other endmembers $\{\boldsymbol{a}_t |\ t \neq k\}$, which might have been updated before $\boldsymbol{a}_k$. To speed up the convergence, we use the updated values of $\{\boldsymbol{a}_t |\ t < k\}$ for updating $\boldsymbol{a}_k$.

V. Experiments and Discussion

In this section, the proposed IR model is tested on both simulated and real HSI, in comparison with several other data representation methods for supervised HSI classification. We start with the introduction to the experiment design, followed by the discussion of results achieved by different methods using both simulated and real HSI.

A. Experimental Setup

The overall strategy in this comparative study is to adopt different data representation methods to extract from HSI separate sets of features, which will then be used to feed classifiers for supervised classification of HSI. The classification performance measured by some numerical statistics can be used as criteria for evaluating the compared techniques.

1) Methods Compared: Two widely used data representation techniques, i.e., PCA [5] and ICA [6], are used for comparison with the proposed method. In addition, two benchmark hyperspectral unmixing techniques, i.e., VCA [31] and MVC-NMF [32], are employed to learn features of physical meanings. For both methods, we use as features the abundance of the endmembers. Because MVC-NMF innately produces abundances, we use the output abundances as features. However, for VCA that only yields endmembers, we employ the NNLS algorithm to calculate the abundances based on the endmembers obtained by VCA.

Three variants of the the proposed IR model, i.e., IR0, IR1 and IR2, are implemented by adopting different algorithms to estimate $\mathbf{A}$ and $\mathbf{S}$.

- IR0 is implemented by using the centroids of K-Means as estimates of $\mathbf{A}$ and IWNNLS-SSC to estimate $\mathbf{S}$.
- IR1 is implemented by adopting the purified means approach in section IV-B to estimate $\mathbf{A}$ and the WNNLS algorithm in Algorithm 2 to estimate $\mathbf{S}$.
- IR2 is implemented by exactly following Algorithm 5, using purified means approach in section IV-B to estimate $\mathbf{A}$ and IWNNLS-SSC to estimate $\mathbf{S}$.

As a result, the performance difference between IR2 and IR1 reveals the role of spatial smoothness constraint, while that between IR2 and IR0 reveals the effectiveness of using physically meaningful features. To serve as a baseline method, we also use all the original bands in hyperspectral data for supervised classification, and the results are referred as the "original" classification results. The popular nonlinear feature extraction technique, i.e., Isomap [34] is also tested in the experiments with real hyperspectral images.

2) Parameter Setting: For PCA, we use the first K PCs that explains the majority of the data variance as features for classification. For ICA, we use fastICA algorithm that estimate K ICs by minimizing the mutual information of the transformed components. The MVC-NMF method uses the endmembers estimated by VCA for initialization. The parameter $tol1$ in IR-based methods is set to be a very small value that is close to zero: $tol1 = 10^{-10}$. IR0 and IR2 use 4-nearest neighbors, and set the weight of smoothness constraint by $\lambda = 10 * mean(diag(\mathbf{\Lambda}))/\mu$, with μ being the mean value of the image. The $maxIter2$ in IR1 and IR2 is set to be 20, and $maxIter1$ in IR2 is set to be 1. The $maxIter2$ in IR0 is set to be 1. In IR1, for the estimation of the noise variances $\{\sigma_j^2\}$ in $\mathbf{\Lambda}$, we use a software interface to identify a homogeneous subimage (HS) in the hyperspectral image. For each band in the HS, the empirical variance is computed and used as the estimate of the variance in the corresponding band: $\sigma_j^2 = var(\{x_{ij} | for\ i\ in\ HS\})$. In IR-based approaches, to deal with the sensitivity to random initializations, like the K-means algorithm, in each run, we used 10 replicates with different random initializations to adopt the output of the one with the smallest representational error. In IR1 and IR2, $\mathbf{A}$ is initialized by using random samples. Initializing $\mathbf{A}$ using other methods such as VCA may potentially improve performance. However, to allow for a fair comparison with the other methods, we decided not to rely on external algorithms for initialization, but instead use random samples. For Isomap, we use the 30 nearest neighbors for building the graph.

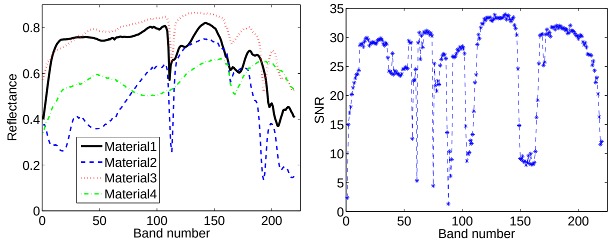
3) Classification Techniques: Each method described above will be used to transform HSI into feature space of reduced dimensionality, where HSI classification is performed using two classifiers, i.e., K-NN and SVM. The reason we adopt K-NN is because it is sensitive to the discriminative capability of features, and thereby can better reveal the difference among techniques used for feature learning. To further enhance such sensitivity, we use only one neighbor in K-NN to perform classification. SVM represents the benchmark classification technique in both remote sensing and machine learning communities. We use the radial basis kernel and select the two hyperparameters, i.e., sigma and the regularization constant, by performing grid search using 10-fold cross-validation based on the training samples.

4) Numerical Measures: A relatively small percentage of the labeled samples in HSI will be used for training the classifiers, while the rest are used for testing. Three numerical measures, i.e., the overall accuracy (OA), averaged accuracy (AA) and the Kappa coefficient, are used for evaluating the classification performance. To reduce the bias caused by randomness in choosing training samples, the numerical measures are averaged over ten independent runs using different training samples.

The strategy described above is tested on both simulated and real HSI. Since the simulated HSI has known data generation model and relatively small number of classes, the simulated study provides us a controlled environment for better comparing different methods and testing the parameter sensitivities.

B. Experiment with Simulated Hyperspectral Image

In this experiment, we simulate a 64-by-64 sized HSI with four classes, with each being dominated by a different material, whose signature with 224 bands is shown in Fig. 1(a). Each pixel in the simulated HSI is a mix of the four endmembers. Using the four endmembers, mixed pixels are created by first dividing the entire image into 8-by-8 sized homogeneous blocks of one of the 4 endmembers, then degrading the blocks by applying a spatial low pass filter of 17-by-17 [32]. To further increase mixing degree, the remaining relatively pure pixels with 80% or larger single abundance are forced to take equal abundances over all endmembers.



(a)

(b)

Fig. 1. (a) Four endmembers randomly selected from the USGS spectral library for simulating HSI. (b) The simulated SNR values of different bands when $\alpha = 7$ and $c = 20$. Different bands tend to have different SNR values.

The resulting HSI is further degraded by zero-mean Gaussian noise with different noise variances in different bands. To simulate this noise heterogeneity effect, we assign different signal-to-noise ratio (SNR) values to different bands. The band-dependent SNR values used for simulation are estimated from the benchmark Indian Pines image introduced in section V-C1. Suppose that the estimated SNR vector $\boldsymbol{q}$ has been centralized and normalized, then the simulated SNR $\boldsymbol{r}$ can be obtained according to the following rule:

$$\boldsymbol{r} = \alpha \boldsymbol{q} + c \tag{28}$$

where α is the amplitude that determines the magnitude of fluctuation of band-dependent SNR values, and c is the center value that defines the overall SNR of all bands. Therefore, by changing c, images with different overall noise levels can be simulated. Larger c value produces an image with lower noise level. The SNR with $\alpha = 7$ and $c = 5$ is shown in Fig. 1(b).

To test the sensitivity to noise level variation, we simulate six HSIs with different overall noise levels by fixing $\alpha = 7$ but varying c, i.e., $c = 5, 10, 15, 20, 25, 30$. Each data representation method is applied to the six HSIs to learn four features for each HSI, and the resulting features are used to feed the K-NN classifier using 10 pixels in each class as training set and the rest pixels as testing set. Fig. 2(a) shows the OA over different c values. Overall, the performance of all methods increase with the decrease of noise levels. The proposed approaches, i.e., IR0, IR1 and IR2, achieve higher OA than the other methods over all noise levels. IR2 performs better than IR1 and IR0, indicating the benefits of adopting the spatial smoothness constraints for estimating $\mathbf{S}$ and purified means approach for estimating $\mathbf{A}$. IR1 performs better than IR0, and approaches IR2 with the decrease of noise levels. PCA and ICA tend to perform better than the original image when noise levels are high. VCA and MVC-NMF fail to achieve higher OA than the original image.

To test the sensitivity to the number of features, we apply each method to the HSI with $\alpha = 7$ and $c = 20$ to learn varying number of features. Fig. 2(b) shows the OA achieved by K-NN as a function of the number of features. Generally speaking, IR2 consistently outperforms the other methods, and achieves the best performance when using four features. IR1 is close to IR2 when using small number of features. IR1 and IR0 achieve higher OA than all the other referenced methods, followed by PCA and ICA. The observation that learning 4 features in IR1 and IR2 gives better statistics than more features is likely due to the fact that in the simulated data there are actually 4 endmembers, which when captured by the IR approaches, provide the maximum discriminative capability. PCA and ICA show the best performance when using 2 feature, and slightly decreased performance when features increases. It is probably because PCA and ICA explore the statistical information, rather than the physical information. So, using more features will not necessarily give PCA and ICA more information, but endanger them with the pollution of noise.

Fig. 3 shows the scatter plots of 2 out of 4 features learnt by different methods from the HSI with $\alpha = 7$ and $c = 20$. Different symbols represent the true class labels of pixels. In VCA and MVC-NMF, the four classes are highly mixed, indicating very low separability among classes. PCA and ICA demonstrate better separation of material 2 and 3, but highly dispersion of the other two classes. IR0 indicate fair separation of material 1, 2 and 4, but highly dispersion of material 3. IR1 tends to cluster the four classes better than IR0. However, material 3 is still highly mixed with the other classes. IR2 demonstrates the best class separability over all methods. In IR2, pixels belonging to the same class tend to cluster together, and different classes are less mixed.

Fig. 4 shows the classification maps of K-NN using a number of 4 features learnt by different methods from the HSI with $\alpha = 7$ and $c = 20$. The maps indicate consistent results with the scatter plots in Fig. 3. Classes in maps of PCA, ICA, VCA and MVC-NMF are greatly misclassified. In IR0, material 3 tends to be wrong classified into other classes. Classes in IR1 are identified very well, but with some within-class artifacts and fluffy boundaries. IR2 enables better delineation of the boundaries, and produces a map that is the most similar to the ground truth. The fact that IR1 produced less within-class artifacts than the standard methods may be explainable by the fact that IR1 addresses the noise heterogeneous effect which contributes largely to the class signatures uncertainty and thereby the label variability in spatial domain.

C. Experiment with Real Hyperspectral Image

1) Indian Pines Scene: The Indian pines scene was standard HSI dataset captured by airborne visible/infrared imaging spectrometer (AVIRIS) over a vegetation area in northwestern Indiana, USA, with a spatial resolution of 20 m, consisting of 145-by-145 pixels and 220 spectral reflectance bands.

Each data representation method is applied to the Indian pines scene to learn 20 features, which is then used to feed SVM and K-NN for classification. There are a total of 16 classes in Indian pines scene. For each class, 10% of the labelled samples are used for training the classifiers, leaving the rest samples for testing.

Table I shows the statistics of different methods. IR2 ranks first in terms of all measures on both classifiers, outperforming the second best, IR1, by 9.4% according to the OA of 1-NN and 11.2% according to the OA of SVM. Comparing with the original image, using IR2 features boosts the OA of 1-NN by 15.7% and the OA of SVM by 13.1%. IR1 achieves higher statistics than all the other referenced methods, i.e., PCA, ICA, Isomap and MVC-NMF, on both classifiers. According to the OA of SVM and 1-NN, only IR1 and IR2 perform better than the original image.

Fig. 5 shows the classification maps achieved by 1-NN using different features. Generally speaking, the classification map associated with IR2 is the most similar one to the ground truth map. It contains less within-class artifacts than those of the other methods, demonstrating the effectiveness of using spatial smoothness constraint for feature learning. Nevertheless, the intense artifacts in the map of IR0 that also adopts the spatial smoothness constraint implies the benefit and necessity of employing simultaneously the purified means approach and the smoothness constraint for model optimization, as conducted in IR2. The IR1 features also leads to significantly less misclassification. However, maps produced by using features of the other methods, i.e., PCA, ICA, Isomap and MVC-NMF, do not demonstrate any visual advantage over map produced using the original image.

Fig. 6 shows the plots of OA and AA achieved by 1-NN using different number of training samples. All feature extraction methods improve the performance of 1-NN, with the increase of training samples. Nevertheless, irrespective of the number of training samples, IR2 leads to much higher OA and AA values than the other methods. Comparing with using the original image, using IR2 features improves the OA by about 20% and AA by about 15% on average over all numbers of training samples. Using IR1 also improves the performance of 1-NN than using the original image, and the improvement seems to grow with the increase of training samples. PCA achieves comparable statistics with the original image in most cases. Features extracted by Isomap, ICA and MVC-NMF fail to improve the classification performance over the original image.

Table II shows the statistics of CMTMF$_{unsup}$ reported in [22]. CMTMF$_{unsup}$ represents the state-of-the-art unmixing-based feature extraction technique. As we can see, IR2 achieves comparable results with CMTMF$_{unsup}$ on the Indian Pines image.

TABLE II
OVERALL ACCURACY (OA %) AND AVERAGED ACCURACY (AA %) ACHIEVED BY SVM USING RESPECTIVELY THE CMTMF$_{unsup}$ FEATURES [22] AND THE IR2 FEATURES ON THE INDIAN PINES IMAGE (WITH 15% TRAINING SAMPLES PER CLASS) AND THE PAVIA UNIVERSITY IMAGE (WITH 50 TRAINING SAMPLES PER CLASS).

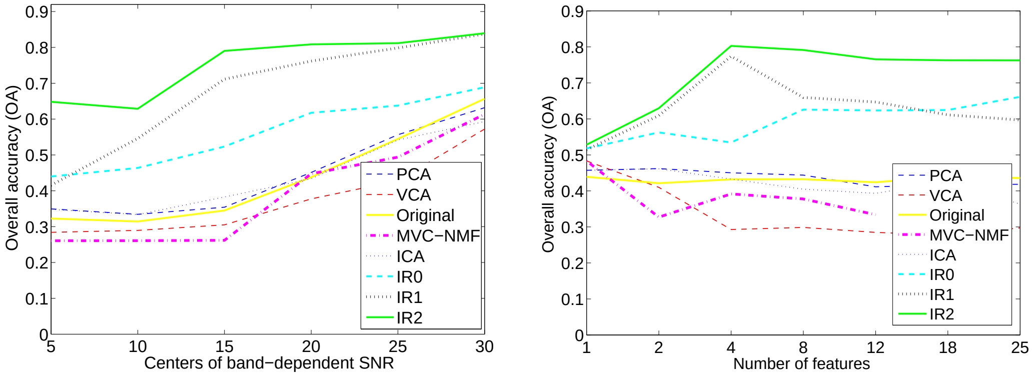
	Indian Pines		Pavia University	
	OA	AA	OA	AA
CMTMF$_{unsup}$ [22]	91.6	89.6	86.8	88.1
IR2	91.5	90.7	91.8	92.5

2) Pavia University Scene: The Pavia University scene was acquired by the reflective optics system imaging spectrometer over the University of Pavia, with a spatial resolution of 1.3 m, consisting of 610-by-340 pixels and 115 spectral bands.

Each method is applied to this image to extract 15 features, which are then used to feed the 1-NN and SVM classifiers. There are nine ground-truth classes in this image. In each class, 100 labeled samples are used for training, while the rest are used for testing.

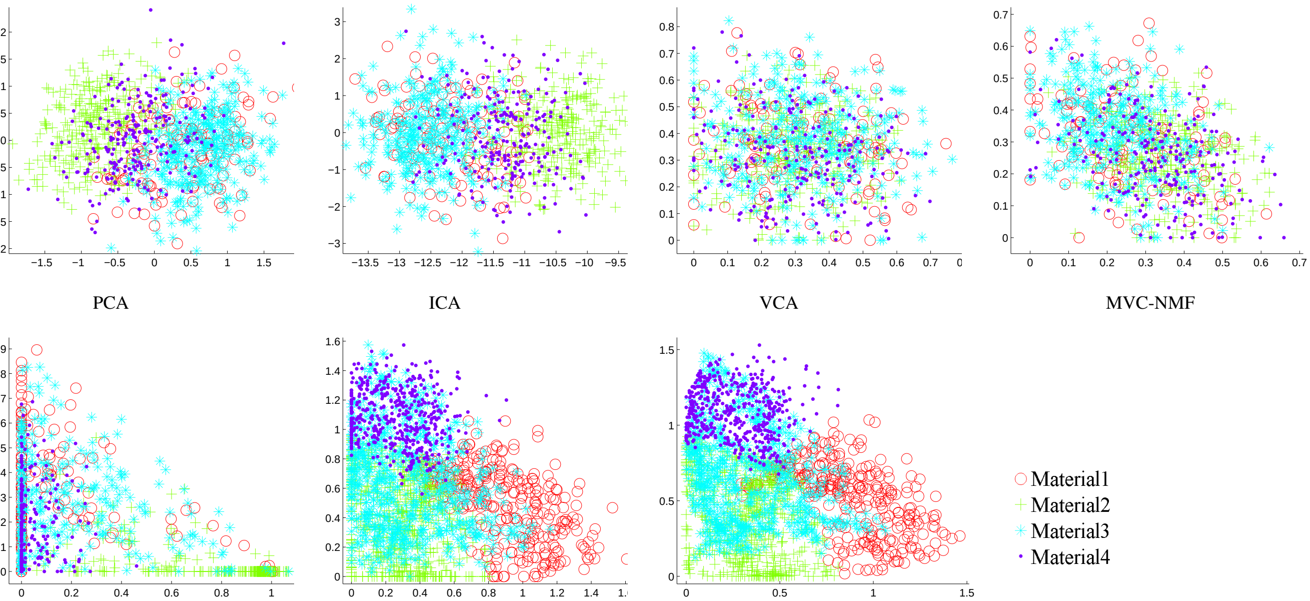
Table III shows the statistics obtained by 1-NN and SVM using features extracted by different methods. It indicates consistent results with Table I. IR2 significantly outperforms the other methods in terms of all measures achieved by both classifiers. IR2 produces OA that is 11.7 higher than the original image according to 1-NN, and 4.3% higher according to SVM. IR1 also achieves higher statistics values than the original image. However, IR0, PCA, ICA, Isomap and MVC-NMF fail to improve the classification performance over the original image.

Fig. 7 displays the classification maps obtained by 1-NN using different types of features. Due to the low class separability of MVC-NMF features, the resulting map tend to have



(a) (b)

Fig. 2. (a) The plot of OA achieved by K-NN using different features over different noise levels defined by the center of band-dependent SNR. Overall, the performance of all methods increase with the increase of SNR. The proposed approaches, i.e., IR0, IR1 and IR2, achieve higher OA than the other methods over all noise levels. IR2 performs better than IR1 and IR0, indicating the benefits of adopting the spatial smoothness constraints for estimating **S** and purified means approach for estimating **A**. (b) The plot of OA achieved by K-NN using different features as a function of the number of features. Generally speaking, IR2 consistently outperforms the other methods, and achieves its best performance when using four features. IR1 is close to IR2 when using a small number of features.



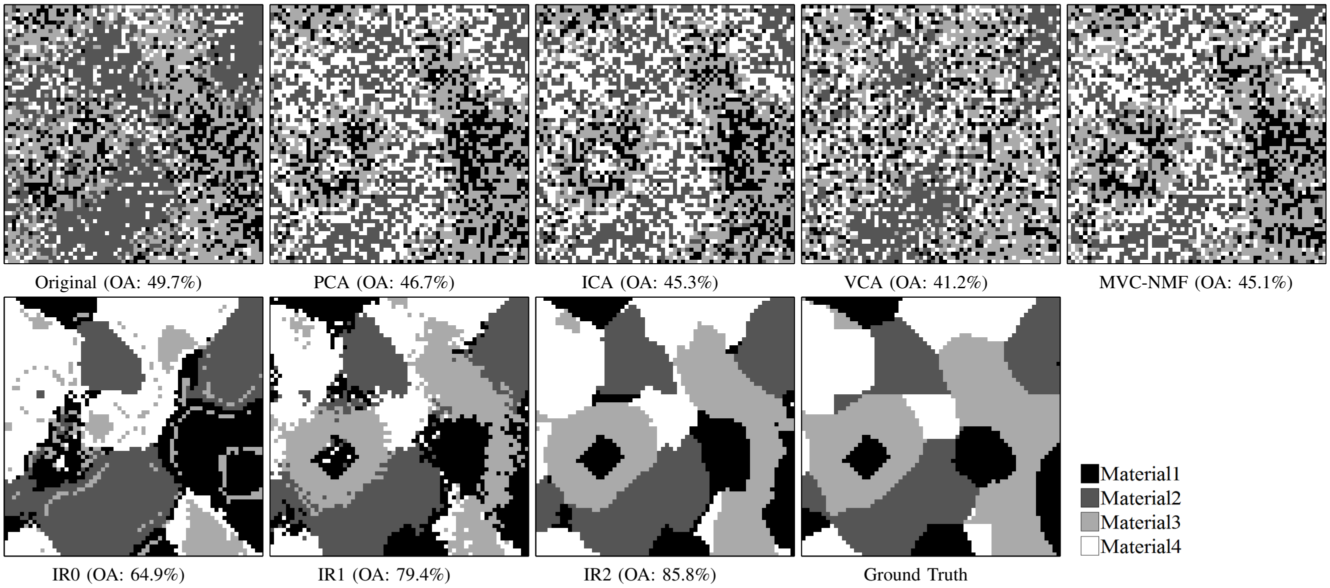
PCA ICA VCA MVC-NMF

IR0 IR1 IR2

Fig. 3. Scatter plots of 2 out of 4 features extracted by different methods from the HSI with $\alpha = 7$ and $c = 20$. Different symbols represent the true class labels of pixels. In VCA and MVC-NMF, the four classes are highly mixed, indicating very low separability among classes. PCA and ICA demonstrate better separation of material 2 and 3, but highly dispersion of the other two classes. IR0 indicate fair separation of material 1, 2 and 4, but highly dispersion of material 3. IR1 tends to cluster the four classes better than IR0. However, material 3 is still highly mixed with the other classes. IR2 demonstrates the best class separability over all methods. In IR2, pixels belonging to the same class tend to cluster together, and different classes are less mixed.

TABLE I
OVERALL ACCURACY (OA), AVERAGED ACCURACY (AA) AND KAPPA COEFFICIENT (KAPPA) ACHIEVED BY DIFFERENT METHODS ON THE INDIAN PINES IMAGE (BEST RESULTS ARE HIGHLIGHTED IN BOLD)

	1-NN			SVM		
	OA (%)	AA (%)	Kappa	OA (%)	AA (%)	Kappa
Original	70.8	67.5	0.667	75.4	69.0	0.712
PCA	70.8	67.7	0.667	73.7	69.2	0.699
ICA	68.3	64.0	0.638	72.3	68.1	0.683
Isomap	65.5	62.0	0.606	68.1	62.8	0.634
MVC-NMF	66.9	63.6	0.622	69.8	62.0	0.654
IR0	66.9	64.7	0.623	68.6	66.6	0.640
IR1	77.1	70.3	0.739	79.7	74.5	0.768
IR2	**86.5**	**84.6**	**0.846**	**88.5**	**88.1**	**0.869**



Original (OA: 49.7%) PCA (OA: 46.7%) ICA (OA: 45.3%) VCA (OA: 41.2%) MVC-NMF (OA: 45.1%)

IR0 (OA: 64.9%) IR1 (OA: 79.4%) IR2 (OA: 85.8%) Ground Truth

Fig. 4. The classification maps achieved K-NN using four features extracted by different methods from the HSI with $\alpha = 7$ and $c = 20$. The maps indicate consistent results with the scatter plots in Fig. 3. Classes in maps of PCA, ICA, VCA and MVC-NMF are greatly misclassified. In IR0, material 3 tends to be wrongly classified into other classes. Classes in IR1 are identified very well, but with some within-class artifacts and inconsistent boundaries. IR2 enables better delineation of the boundaries, and produces a map that is the most similar to the ground truth with the highest accuracy.

large misclassification. PCA, ICA, Isomap and IR0 features enable 1-NN to perform better than MVC-NMF. However, the resulting maps still contain many within-class artifacts. IR1 produces less misclassification than the aforementioned methods, especially in the classes of Meadow and Bare soil. IR2 leads to a map that has the least within-class artifacts and accurate identification of small classes.

Fig. 8 plots the OA and AA of 1-NN over the number of samples used for training. Similar to Fig. 6, IR2 greatly outperforms the other methods regardless of the number of training samples. IR1 also produced higher OA values but slightly lower AA values than the original image. Statistics that are comparable with the original image are achieved by PCA, and lower statistics are achieved by MVC-NMF, ICA, Isomap and IR1.

Table II indicates that IR2 achieves higher OA and AA values than CMTMF$_{unsup}$ [22] on the Pavia University image.

VI. Conclusions

In this paper, we presented an IR model for unsupervised feature extraction from HSI. Different to the other representation-based feature extraction techniques, such as PCA and ICA, which define the ”informativeness” of features from a statistical perspective based on the domain-independent knowledge, IR aims to capture the discriminative information in HSI by explicitly modeling the physical quantities that are responsible for HSI generation. Moveover, IR accounts for the other key characteristics of HSI, i.e., the spatial correlation effect in spatial domain and the noise heterogeneity effect in spectral domain. IR is solved iteratively by alternating the the estimation of the IR features given the endmembers and the update of the endmembers given the IR features. IR is tested on both real and simulated HSI, in comparison with several other popular unsupervised features extraction methods. The results indicate that IR features are more capable of boosting the classification performance than the referenced methods.

VII. Acknowledgements

This work is funded by the National Natural Science Foundation of China (Grant No. 41501410), Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Space Agency (CSA). This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. This authors would like to thank the anonymous reviewers for their comments and suggestions.

References

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri *et al.*, “Recent advances in techniques for hyperspectral image processing,” *Remote sensing of environment*, vol. 113, pp. S110–S122, 2009.

[2] G. P. Hughes, “On the mean accuracy of statistical pattern recognizers,” *Information Theory, IEEE Transactions on*, vol. 14, no. 1, pp. 55–63, 1968.

[3] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2009, vol. 383.

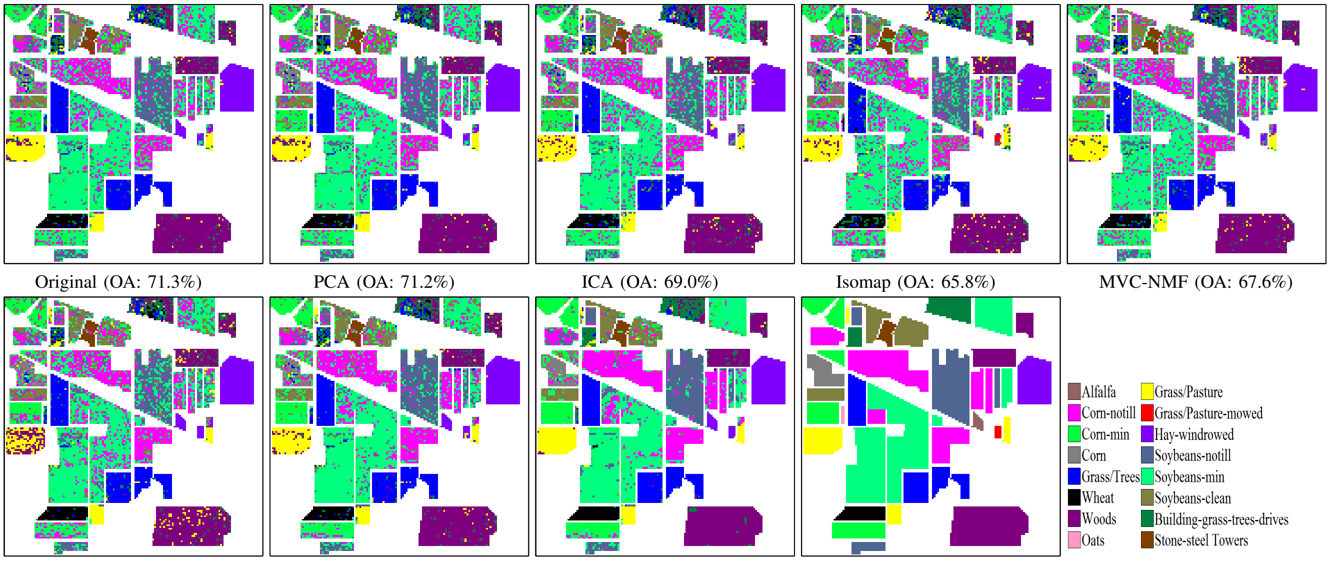
[4] D. L. Donoho *et al.*, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS Math Challenges Lecture*, pp. 1–32, 2000.

[5] H. Hotelling, “Analysis of a complex of statistical variables into principal components.” *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[6] J. Wang and C.-I. Chang, “Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 6, pp. 1586–1600, 2006.

[7] 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,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 26, no. 1, pp. 65–74, 1988.

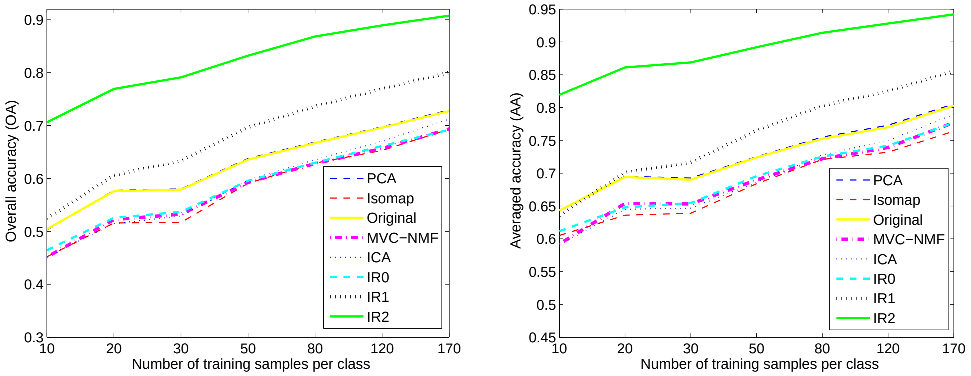
[8] J. B. Lee, A. S. Woodyatt, and M. Berman, “Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 28, no. 3, pp. 295–304, 1990.



Original (OA: 71.3%) PCA (OA: 71.2%) ICA (OA: 69.0%) Isomap (OA: 65.8%) MVC-NMF (OA: 67.6%)

IR0 (OA: 67.1%) IR1 (OA: 78.6%) IR2 (OA: 86.7%) Ground Truth

Fig. 5. The classification maps for Indian pines scene achieved by 1-NN using features extracted by different methods. Generally speaking, the classification map associated with IR2 is the most similar one to the ground truth map. It contains less within-class artifacts than those of the other methods, demonstrating the effectiveness of using spatial smoothness constraint for feature learning. Nevertheless, the intense artifacts in the map of IR0 that also adopts the spatial smoothness constraint implies the benefit and necessity of employing simultaneously the purified means approach and the smoothness constraint for model optimization, as conducted in IR2. The IR1 features also leads to significantly less misclassification. However, maps produced by using features of the other methods, i.e., PCA, ICA, Isomap and MVC-NMF, do not demonstrate any visual advantage over map produced using the original image.

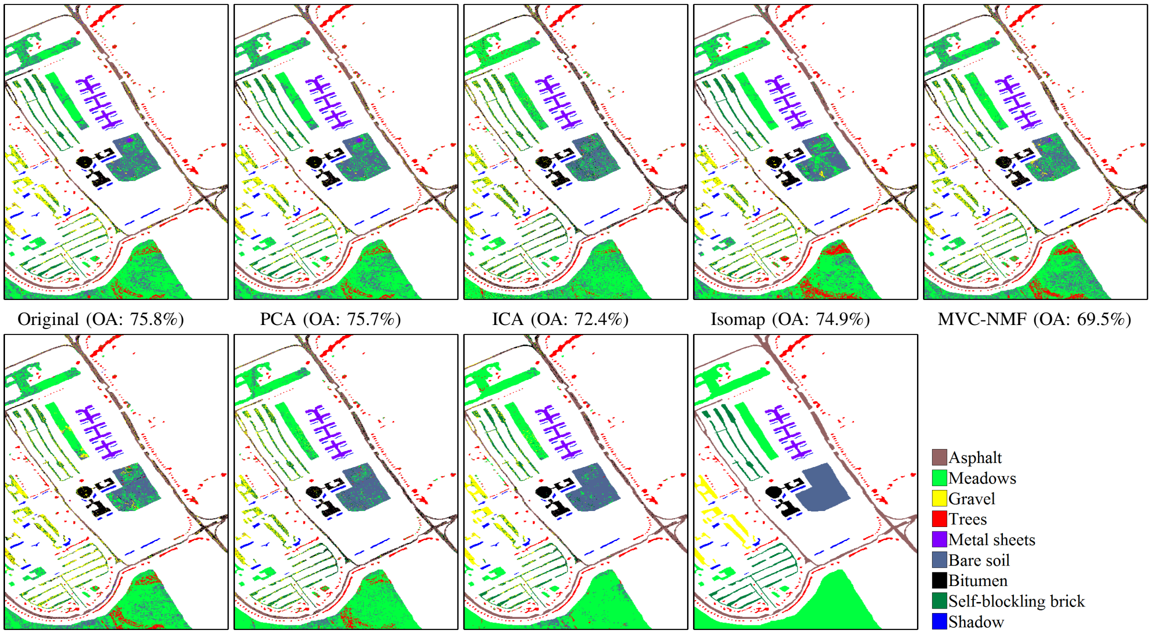


(a) (b)

Fig. 6. The plots of OA and AA achieved by 1-NN on Indian pines scene using different number of training samples. All feature extraction methods improve the performance of K-NN, with the increase of training samples. Nevertheless, irrespective of the number of training samples, IR2 obtains much higher OA and AA values than the other methods. Comparing with using the original image, using IR2 features leads to about 20% higher OA and 15% higher AA on average over all numbers of training samples. Using IR1 also improves the performance of 1-NN than using the original image, and the improvement seems to grow with the increase of training samples. PCA achieves comparable statistics with the original image in most cases.

TABLE III
Overall accuracy (OA), averaged accuracy (AA) and Kappa coefficient (Kappa) achieved by different methods on the Pavia University image (Best results are highlighted in bold)

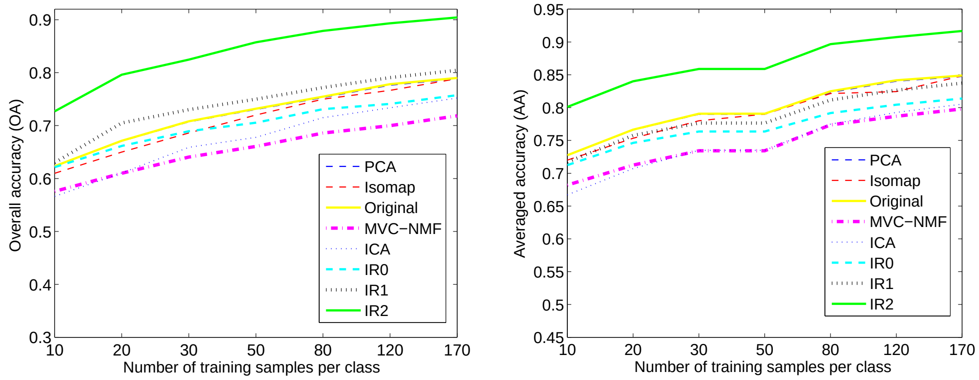
	1-NN			SVM		
	OA (%)	AA (%)	Kappa	OA (%)	AA (%)	Kappa
Original	76.8	83.2	0.704	88.9	90.6	0.854
PCA	76.7	83.1	0.703	87.0	88.6	0.829
ICA	72.5	78.2	0.652	85.9	88.0	0.815
Isomap	75.4	82.6	0.691	79.8	86.8	0.731
MVC-NMF	69.5	78.1	0.619	76.9	84.0	0.706
IR0	73.1	79.9	0.662	78.5	84.3	0.725
IR1	78.2	81.9	0.721	89.2	91.0	0.858
IR2	**88.5**	**90.2**	**0.851**	**93.1**	**94.3**	**0.909**



Original (OA: 75.8%) PCA (OA: 75.7%) ICA (OA: 72.4%) Isomap (OA: 74.9%) MVC-NMF (OA: 69.5%)

IR0 (OA: 73.9%) IR1 (OA: 79.5%) IR2 (OA: 88.8%) Ground Truth

Fig. 7. The classification maps for Pavia University scene obtained by 1-NN using different types of features. Due to the low class separability of MVC-NMF features, the resulting map tend to have large misclassification. PCA, ICA, Isomap and IR0 features enable 1-NN to perform better than MVC-NMF. However, the resulting maps still contain many within-class artifacts. IR1 produces less misclassification than the aforementioned methods, especially in the classes of Meadow and Bare soil. IR2 leads to a map that has the least within-class artifacts and accurate identification of small classes.



(a) (b)

Fig. 8. OA in (a) and AA in (b) achieved by 1-NN on Pavia University scene over different number of samples used for training. Similar to Fig. 6, IR2 greatly outperforms the other methods regardless of the number of training samples. IR1 also produced higher OA values but slightly lower AA values than the original image. Statistics that are comparable with the original image are achieved by PCA, and lower statistics are achieved by MVC-NMF, ICA, Isomap and IR1.

[9] A. Ifarraguerri and C.-I. Chang, "Unsupervised hyperspectral image analysis with projection pursuit," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 38, no. 6, pp. 2529–2538, 2000.

[10] S.-S. Chiang, C.-I. Chang, and I. W. Ginsberg, "Unsupervised target detection in hyperspectral images using projection pursuit," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 7, pp. 1380–1391, 2001.

[11] L. O. Jiménez-Rodríguez, E. Arzuaga-Cruz, and M. Vélez-Reyes, "Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 2, pp. 469–483, 2007.

[12] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 3, pp. 441–454, 2005.

[13] C. M. Bachmann, T. L. Ainsworth, and R. Fusina, "Improved manifold coordinate representations of large-scale hyperspectral scenes," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 10, pp. 2786–2803, 2006.

[14] C. M. Bachmann, T. L. Ainsworth, R. Fusina, M. J. Montes, J. H. Bowles, D. R. Korwan, and D. B. Gillis, "Bathymetric retrieval from hyperspectral imagery using manifold coordinate representations," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 3, pp. 884–897, 2009.

[15] Y. Chen, M. M. Crawford, and J. Ghosh, "Improved nonlinear manifold learning for land cover classification via intelligent landmark selection," in *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*. IEEE, 2006, pp. 545–548.

[16] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 3, pp. 466–479, 2005.

[17] N. Renard, S. Bourennane, and J. Blanc-Talon, "Denoising and dimensionality reduction using multilinear tools for hyperspectral images," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 2, pp. 138–142, 2008.

[18] R. Dianat and S. Kasaei, "Dimension reduction of optical remote sensing images via minimum change rate deviation method," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 1, pp. 198–206,

2010.

[19] B. Luo and J. Chanussot, “Unsupervised classification of hyperspectral images by using linear unmixing algorithm,” in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 2877–2880.

[20] I. Dópido, A. Villa, A. Plaza, and P. Gamba, “A quantitative and comparative assessment of unmixing-based feature extraction techniques for hyperspectral image classification,” *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 2, pp. 421–435, 2012.

[21] I. Dopido, M. Zortea, A. Villa, A. Plaza, and P. Gamba, “Unmixing prior to supervised classification of remotely sensed hyperspectral images,” *Geoscience and Remote Sensing Letters, IEEE*, vol. 8, no. 4, pp. 760–764, 2011.

[22] A. Villa, J. Chanussot, J. A. Benediktsson, and C. Jutten, “Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 3, pp. 521–533, 2011.

[23] M. F. Tappen, W. T. Freeman, and E. H. Adelson, “Recovering intrinsic images from a single image,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 9, pp. 1459–1472, 2005.

[24] X. Kang, S. Li, L. Fang, and J. A. Benediktsson, “Intrinsic image decomposition for feature extraction of hyperspectral images,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 53, no. 4, pp. 2241–2253, 2015.

[25] D. C. Heinz and C.-I. Chang, “Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 3, pp. 529–545, 2001.

[26] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[27] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[28] A. C. Zelinski and V. K. Goyal, “Denoising hyperspectral imagery and recovering junk bands using wavelets and sparse approximation,” in *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*. IEEE, 2006, pp. 387–390.

[29] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1974, vol. 161.

[30] R. Bro and S. De Jong, “A fast non-negativity-constrained least squares algorithm,” *Journal of chemometrics*, vol. 11, no. 5, pp. 393–401, 1997.

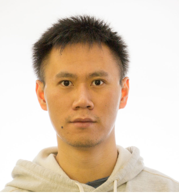
[31] J. M. Nascimento and J. M. Bioucas Dias, “Vertex component analysis: A fast algorithm to unmix hyperspectral data,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, 2005.

[32] L. Miao and H. Qi, “Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 3, pp. 765–777, 2007.

[33] L. Xu, J. Li, A. Wong, and J. Peng, “K-p-means: A clustering algorithm of k purified means for hyperspectral endmember estimation,” *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 10, pp. 1787–1791, 2014.

[34] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

Linlin Xu (M14) received his B.Eng. and M.Sc. degrees in geomatics engineering from China University of Geosciences, Beijing, China, in 2007 and 2010, respectively. He obtained his Ph.D. degree in geography from the University of Waterloo, Waterloo, ON, Canada. He is now working as a post-doc fellow in Vision and Image Processing Lab, Systems Design Engineering from the University of Waterlo. His current research interests are in the areas of hyperspectral and SAR image processing.



Fan Li received his B.S. degree of geographical information system from Sun Yat-sen University, China and M.E. degree in pattern recognition and intelligent system from Wuhan University, China. He is currently working towards his Ph.D degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada. His research interest is the application of machine learning and computer vision techniques to the remote sensing filed.

Alexander Wong (M05) received the B.A.Sc. degree in computer engineering, the M.A.Sc. degree in electrical and computer engineering, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 2005, 2007, and 2010, respectively. He is currently an Assistant Professor with the Department of Systems Design Engineering, University of Waterloo. He has authored refereed journal and conference papers, as well as patents, in various fields, such as computer vision, graphics, image processing, multimedia systems, and wireless communications. His current research interests include image processing, computer vision, pattern recognition, and cognitive radio networks, with a focus on biomedical and remote sensing image processing and analysis such as image registration, image denoising and reconstruction, image super-resolution, image segmentation, tracking, and image and video coding and transmission. Dr. Wong was the recipient of an Outstanding Performance Award, an Engineering Research Excellence Award, an Early Researcher Award from the Ministry of Economic Development and Innovation, a Best Paper Award by the Canadian Image Processing and Pattern Recognition Society, and the Alumni Gold Medal.

David Clausi (S93M96SM03) received the BASc, MASc, and PhD degrees in systems design engineering from the University of Waterloo, Canada, in 1990, 1992, and 1996, respectively. After completing his PhD, he worked in the medical imaging field at Mitra Imaging Inc., Waterloo. He started his academic career as an assistant professor in geomatics engineering at the University of Calgary, Canada, in 1997. He returned to his alma mater in 1999 and was awarded tenure and promotion to associate professor in 2003. He is an active interdisciplinary and multidisciplinary researcher. He has an extensive publication record, publishing refereed journal and conference papers on remote sensing, computer vision, algorithm design, and biomechanics. His primary research interest is the automated interpretation of synthetic aperture radar (SAR) sea ice imagery, in support of operational activities of the Canadian Ice Service. The research results have led to successful commercial implementations. He has received numerous scholarships, conference paper awards, and two Teaching Excellence Awards. He is a senior member of the IEEE.