View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Ghent University Academic Bibliography

sensors

MDPI

Article

A Robust Sparse Representation Model for Hyperspectral Image Classification †

Shaoguang Huang [1,*], Hongyan Zhang [2] and Aleksandra Pižurica [1]

1 Department of Telecommunications and Information Processing, Ghent University, Sint Pietersnieuwstraat 41, 9000 Gent, Belgium; aleksandra.pizurica@ugent.be

2 The State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Luoyu Road 129, Wuhan 430079, China; zhanghongyan@whu.edu.cn

* Correspondence: shaoguang.huang@ugent.be; Tel.: +32-9-264-3416

† This paper is an extended version of our paper published in ICIP 2017, "Robust Joint Sparsity Model for Hyperspectral Image Classification".

Received: 10 August 2017; Accepted: 7 September 2017; Published: 12 September 2017

Abstract: Sparse representation has been extensively investigated for hyperspectral image (HSI) classification and led to substantial improvements in the performance over the traditional methods, such as support vector machine (SVM). However, the existing sparsity-based classification methods typically assume Gaussian noise, neglecting the fact that HSIs are often corrupted by different types of noise in practice. In this paper, we develop a robust classification model that admits realistic mixed noise, which includes Gaussian noise and sparse noise. We combine a model for mixed noise with a prior on the representation coefficients of input data within a unified framework, which produces three kinds of robust classification methods based on sparse representation classification (SRC), joint SRC and joint SRC on a super-pixels level. Experimental results on simulated and real data demonstrate the effectiveness of the proposed method and clear benefits from the introduced mixed-noise model.

Keywords: robust classification; hyperspectral image; super-pixel segmentation; sparse representation

1. Introduction

Unlike classical multispectral images, hyperspectral images (HSIs) provide richer spectral information about the image objects in hundreds of narrow bands. A HSI is captured as a three-dimensional data cube comprising two-dimensional spatial information and one-dimensional spectral information. The spectral signature of a pixel is a vector whose entries correspond to spectral responses in different bands. Different materials have diverse spectral signatures, thus hyperspectral imaging allows differentiation between materials that are often visually indistinguishable. Numerous application areas include agriculture [1,2], defense and security [3] and environmental monitoring [4,5].

Classification of HSIs currently enjoys huge interest in the remote sensing community. The objective of supervised hyperspectral classification is to group pixels into different classes with the classifiers trained by the given training samples. A large number of HSI classification methods have been proposed, based on artificial neural networks [6], multinomial logistic regression [7,8], spectral-spatial preprocessing with multihypothesis prediction [9], information fusion [10] and support vector machines (SVM) [11], just to name a few. With the target of exploiting spatial information in the classification task, spatial-spectral classification approaches have been developed, including SVM with composite kernels [12], methods based on mathematical morphology [13–17] and image segmentation [18].

In recent years, sparse representation classification (SRC) [19] emerged as another effective classification approach, which became widely adopted for HSI [20–34]. SRC assumes that each test sample can be sparsely represented as a linear combination of atoms from a dictionary, which is constructed or learned from training samples [19]. Chen et al. [20] introduced the joint sparse representation classification (JSRC) in HSI classification by incorporating spatial information. The model was based on the observation that the pixels in a patch share similar spectral characteristics and can be represented by a common set of atoms but with different sparse coefficients. Zhang et al. [21] proposed a nonlocal weighted joint sparse representation (NLW-JSRC) to further improve the classification accuracy. They enforced a weight matrix on the pixels of a patch in order to discard the invalid pixels whose class was different from that of the central pixel. The works in [22,24] extended the JSRC to the kernel versions to address the linearly non-separable problem. In [27], a multi-layer spatial-spectral sparse representation framework was proposed for HSI classification in order to stabilize the sparse codes of the traditional single-layer sparse representation. Related classification methods effectively exploiting spatial information with adaptive neighborhood were reported in [25,26,31] and produced good results. Recent studies in [29–33] indicated that learning a compact and discriminative dictionary from the training samples can reduce the computational burden significantly.

However, all of these sparsity-based methods for HSI classification only take into account Gaussian noise. In real applications, HSIs are inevitably corrupted by different kinds of noise, including Gaussian noise and sparse noise. Here, sparse noise is defined as the noise of arbitrary magnitude that only affects certain bands or pixels, which can be impulse noise, dead lines and strips. It may arise due to the defective pixels and poor imaging conditions such as water vapor and atmospheric effect [35]. With the consideration of sparse noise in the tasks of HSIs denoising [36,37], unmixing [35,38] and robust learning [39], significant improvements have been achieved over the state-of-the-art methods, which indicates the importance of taking the sparse noise into account in those tasks. For the classification task, the sparse noise can hinder the performance undoubtedly. We are not aware of any sparsity-based classification method that takes it explicitly into account. This motivates us to develop a robust classification model that accounts for realistic degradations in the HSIs.

The key idea of our model is to incorporate the presence of sparse noise in HSIs into the classification problem, by combining the appropriate statistical models for the sparse noise and the representation coefficients of test pixel(s) within a unified framework. In particular, we make use of the fact that test pixels can be represented with relatively few atoms from a well constructed dictionary, meaning that the representation coefficients are sparse or jointly sparse within small neighborhoods. This is the main assumption of SRC and JSRC models. In addition, we introduce a statistical model for the sparse noise as an instance of a multivariate Laplacian distribution, which allows us to derive an optimization problem that extends elegantly the previous ones with an additional ℓ_1 norm on the sparse noise term. Following this idea, we extend and generalize the existing SRC [19] and JSRC [20] methods to the robust versions, i.e., robust SRC (R-SRC) and robust JSRC (R-JSRC), respectively. We also derive an optimization algorithm for the corresponding objective function, based on the alternating minimization strategy.

Moreover, in order to further exploit the available spatial information, we extend the R-JSRC model to a classification model on a super-pixel level. In the JSRC model, spatial information is defined by the collection of neighbouring pixels in a square window of fixed size, while super-pixel segmentation can adaptively divide the HSIs into a number of non-overlapping homogenous regions depending on the spatial content, which makes the joint sparse representation more effective and precise. We name this extended method robust super-pixel level joint sparse representation classification (R-SJSRC). The results on simulated and real data demonstrate improved performance in comparison to recent related methods and a clear benefit resulting from the introduced robust model. Parts of this work have been accepted for presentation at a conference [40]. In comparison to the conference version, here we give more elaborate presentation and analysis of the method. Moreover,

extra experiments with both simulated and real HSI data are conducted to investigate the effect of sparse noise and parameters on performance.

The main contributions of the paper can be summarized as follows:

(1) A robust sparsity-based classification model for HSIs is proposed when the data is corrupted by Gaussian noise and sparse noise, by incorporating the appropriate priors for noise-free data and degradations into an optimization framework.
(2) An efficient algorithm is developed to solve the optimization problem by using an alternating minimization strategy.
(3) The robust model is extended to efficiently incorporate spatial information. By jointly processing super-pixels, we strongly improve the performance both in terms of the classification accuracy and processing speed.

The rest of this paper is organized as follows. Section 2 reviews briefly the classical sparsity-based models in HSI classification. Section 3 extends the existing sparsity-based models to the robust versions and designs an effective algorithm to solve corresponding optimization problems. Section 4 presents experimental results with simulated and real data and Section 5 concludes the paper.

2. Sparsity-Based Models in HSI Classification

2.1. Sparse Representation Classification

Let $\mathbf{x} \in \mathbb{R}^B$ be a test pixel and $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_C] \in \mathbb{R}^{B \times d}$ a structured dictionary constructed from training samples, where B is the number of bands in the HSI; d is the number of training samples; C is the number of classes, and $\mathbf{D}_i \in \mathbb{R}^{B \times d_i}$ (i = 1, 2, ..., C) is the sub-dictionary in which each column is a training sample of i-th class, and d_i is the number of training samples from class i, such that $\sum_{i=1}^{C} d_i = d$. The goal of sparse representation is to represent each test pixel as

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{n}, \tag{1}$$

where $\mathbf{n} \in \mathbb{R}^B$ is Gaussian noise and $\boldsymbol{\alpha} \in \mathbb{R}^d$ are sparse coefficients, satisfying

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad s.t. \quad \|\boldsymbol{\alpha}\|_0 \leq K. \tag{2}$$

$\|\boldsymbol{\alpha}\|_0$ denotes the number of non-zero elements in $\boldsymbol{\alpha}$ and K is the sparsity level, i.e., the largest number of atoms in dictionary $\mathbf{D}$ needed to represent any input sample $\mathbf{x}$. Problem in Equation (2) is typically solved with a greedy algorithm, such as Orthogonal Matching Pursuit (OMP) [41].

The class of the test sample is identified by calculating the class-specific residuals r_i [19]:

$$\begin{aligned} class(\mathbf{x}) &= \arg\min_{i=1,2,...,C} r_i(\mathbf{x}) \\ &= \arg\min_{i=1,2,...,C} \|\mathbf{x} - \mathbf{D}_i\boldsymbol{\alpha}_i\|_2, \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}_i$ are the sparse coefficients associated with class i.

2.2. Joint Sparse Representation Classification

An effective method to exploit the spatial information of the HSI is using joint sparse representation of neighbouring pixels. The assumption is that the pixels in a small patch are likely to belong to the same class and thus share the same sparsity pattern, meaning that they can be represented by the same set of atoms but with different sets of coefficients [20]. In the JSRC model, the spatial neighbourhood for the central pixel is a square window and all the neighbouring pixels are gathered

into the input matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T] \in \mathbb{R}^{B \times T}$, where $\mathbf{x}_i$ is the spectral signature of the i-th pixel in a patch of size $\sqrt{T} \times \sqrt{T}$. Denoting by $\boldsymbol{\alpha}_i$ the sparse coefficients of $\mathbf{x}_i$ in dictionary $\mathbf{D}$ leads to

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T] \\ &= [\mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{n}_1, \mathbf{D}\boldsymbol{\alpha}_2 + \mathbf{n}_2, ..., \mathbf{D}\boldsymbol{\alpha}_T + \mathbf{n}_T] \\ &= \mathbf{D}\mathbf{A} + \mathbf{N}, \end{aligned} \tag{4}$$

where $\mathbf{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_T] \in \mathbb{R}^{d \times T}$ is the coefficient matrix.

Since all $\mathbf{x}_i$ in a small patch are likely to belong to the same class and thus share the same set of atoms, $\boldsymbol{\alpha}_i$ have non-zero entries at the same positions. Therefore, $\mathbf{A}$ is row-sparse, and can be obtained by solving the following problem with Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm [42]:

$$\begin{aligned} \hat{\mathbf{A}} &= \arg\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 \\ &s.t.\ \|\mathbf{A}\|_{row,0} \leq K_0, \end{aligned} \tag{5}$$

where $\|\mathbf{X}\|_F$ denotes the Frobenius norm of $\mathbf{X}$, $\|\mathbf{A}\|_{row,0}$ denotes the number of non-zero rows of $\mathbf{A}$ and K_0 is the row-sparsity level. In a similar way to SRC, the central test pixel of the patch is labeled by minimizing the class-specific residual:

$$class(\mathbf{x}_{central}) = \arg\min_{i=1,2,...,C} \|\mathbf{X} - \mathbf{D}_i\mathbf{A}_i\|_F, \tag{6}$$

where $\mathbf{A}_i$ is the portion of the sparse matrix $\mathbf{A}$ associated with class i.

3. Proposed Method

3.1. Robust SRC Model

Here, we develop a more general classification method, which takes into account not only the Gaussian noise (as described above) but also sparse noise, which affects real HSIs. The motivation is as follows. In practice, HSIs are often contaminated by horizontal and vertical strips, impulse noise and dead lines. This type of degradation is called sparse noise as it affects only relatively few pixels. Sparse noise typically arises in situations with poor imaging conditions due to sensor artifacts. In the real HSIs, different bands can be corrupted by different kinds of noise [35,38]. In some bands, sparse noise is a dominant degradation, while others may be corrupted by mixed noise. An example of the noise in real HSI (Hyperspectral Digital Image Collection Experiment (HYDICE) Urban data set [35]) can be found in Figure 1, where Figure 1a shows a band affected with stripe noise, and Figure 1b shows a band affected by a mixture of sparse noise and Gaussian noise. We model the observed pixel in HSI as:

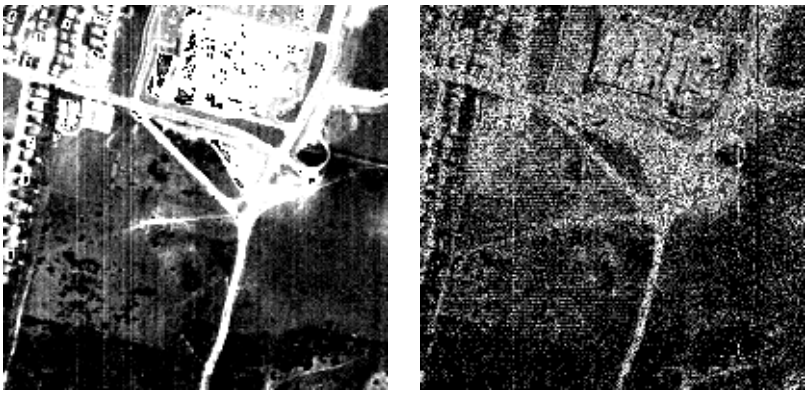
$$\mathbf{x} = \mathbf{y} + \mathbf{s} + \mathbf{n}, \tag{7}$$

where $\mathbf{y} \in \mathbb{R}^B$ is an error-free sample, $\mathbf{s} \in \mathbb{R}^B$ sparse noise and $\mathbf{n} \in \mathbb{R}^B$ Gaussian noise.

As the error-free samples are not available in practice, we have to express $\mathbf{y}$ in terms of the observed samples. To this end, we will employ in our derivation a hypothetic, ideal dictionary $\mathbf{D}^y$. Let $\mathbf{D}^y \in \mathbb{R}^{B \times d} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_d]$ denote an ideal, error-free dictionary and $\mathbf{y}_j \in \mathbb{R}^B$ the j-th error-free training sample. The main assumption of SRC is that any $\mathbf{y}$ can be represented by a few atoms in $\mathbf{D}^y$ as follows:

$$\mathbf{y} = \mathbf{D}^y\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{\alpha}$ is a sparse vector and $\boldsymbol{\varepsilon}$ is arbitrarily small.



(**a**) (**b**)

Figure 1. An illustration of noise in a real HSI. Two bands from HYDICE Urban data set are shown. (**a**) sparse noise (vertical lines) in band 2 (contrast enhanced); (**b**) mixed noise in band 206.

The model (7) holds for any observed sample: $\mathbf{x}_i = \mathbf{y}_i + \mathbf{s}_i + \mathbf{n}_i$. Equivalently, we can write

$$\mathbf{D} = \mathbf{D}^y + \mathbf{D}^s + \mathbf{D}^n, \tag{9}$$

where $\mathbf{D} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d]$, $\mathbf{D}^s = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_d]$ and $\mathbf{D}^n = [\mathbf{n}_1, \mathbf{n}_2, ..., \mathbf{n}_d]$ are collections, or dictionaries, composed of the observed data $\mathbf{x}_i$, sparse noise components $\mathbf{s}_i$ and Gaussian noise components $\mathbf{n}_i$, respectively.

Substituting the Equations (7) and (9) into (8), we derive the representation of $\mathbf{x}$ as follows:

$$\begin{aligned} \mathbf{x} &= (\mathbf{D} - \mathbf{D}^s - \mathbf{D}^n)\boldsymbol{\alpha} + \boldsymbol{\varepsilon} + \mathbf{s} + \mathbf{n} \\ &= \mathbf{D}\boldsymbol{\alpha} + \mathbf{s}' + \mathbf{n}', \end{aligned} \tag{10}$$

where $\mathbf{s}' = \mathbf{s} - \mathbf{D}^s\boldsymbol{\alpha}$ and $\mathbf{n}' = \mathbf{n} - \mathbf{D}^n\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$.

A linear combination of two (or more) sparse vectors is not necessarily sparse. However, the sparse noise in HSIs is typically detected only in certain and relatively few bands, which means the non-zero elements of $\mathbf{s}$ and $\mathbf{s}_i$ are located at the relatively few positions. Therefore $\mathbf{s}'$, being a linear combination of $\mathbf{s}$ and elements of $\{\mathbf{s}_i\}_{i=1}^d$, is sparse as well. The expression in Equation (10) tells us that the observed pixel contaminated by sparse noise and Gaussian noise can be represented by relatively few atoms from the noisy dictionary with the addition of a sparse term $\mathbf{s}'$ and an error term $\mathbf{n}'$. Note that here $\mathbf{s}'$ in Equation (10) is not exactly the sparse noise of $\mathbf{x}$ but a mixture of the sparse noise in $\mathbf{x}$ and $\mathbf{D}$, which is the reason why this model can not be directly used in the denoising task.

Now, we are ready to define an optimization problem that generalizes the one in Equation (2) as a result of our mixed-noise model. Observe first that the problem in Equation (2) can equivalently be written as

$$\arg\max_{\boldsymbol{\alpha}} p(\mathbf{x};\boldsymbol{\alpha}) \quad s.t. \quad \|\boldsymbol{\alpha}\|_0 < K, \tag{11}$$

where $p(\mathbf{x};\boldsymbol{\alpha})$ is the probability distribution of $\mathbf{x}$ with parameter $\boldsymbol{\alpha}$, which is according to the model in Equation (1): $p(\mathbf{x};\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{D}\boldsymbol{\alpha}; \sigma_n\mathbf{I}) \propto exp(-\frac{1}{2\sigma_n^2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2)$. We formulate a similar problem taking into account the sparse noise $\mathbf{s}'$:

$$\arg\max_{\boldsymbol{\alpha},\mathbf{s}'} p(\mathbf{x},\mathbf{s}';\boldsymbol{\alpha}) \quad s.t. \quad \|\boldsymbol{\alpha}\|_0 < K. \tag{12}$$

Making use of the fact that $p(\mathbf{x}, \mathbf{s}') = p(\mathbf{x}|\mathbf{s}')p(\mathbf{s}')$ and that the parameter $\boldsymbol{\alpha}$ appears only in the first term, we can rewrite the objective function in Equation (12) as

$$\arg\max_{\boldsymbol{\alpha},\mathbf{s}'} p(\mathbf{x}|\mathbf{s}';\boldsymbol{\alpha})p(\mathbf{s}') \qquad s.t. \quad \|\boldsymbol{\alpha}\|_0 < K. \tag{13}$$

From our model (10), it follows that $p(\mathbf{x}|\mathbf{s}';\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{s}'; \sigma_{n'}\mathbf{I}) \propto exp(-\frac{1}{2\sigma_{n'}^2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha} - \mathbf{s}'\|_2^2)$. By imposing a Laplacian prior on $\mathbf{s}'$ of the form: $p(\mathbf{s}') \propto exp(-\frac{1}{2\tau}\|\mathbf{s}'\|_1)$ with $\tau > 0$ and $\|\mathbf{s}'\|_1 = \sum_{i=1}^{B} |s_i'|$, the left-hand term in Equation (13) can be written as

$$\arg\min_{\boldsymbol{\alpha},\mathbf{s}'} \frac{1}{2\sigma_{n'}}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha} - \mathbf{s}'\|_2^2 + \frac{1}{2\tau}\|\mathbf{s}'\|_1. \tag{14}$$

With this, we can rewrite the Equation (13) as

$$\arg\min_{\boldsymbol{\alpha},\mathbf{s}'} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha} - \mathbf{s}'\|_2^2 + \lambda\|\mathbf{s}'\|_1 \qquad s.t. \quad \|\boldsymbol{\alpha}\|_0 \leq K, \tag{15}$$

where $\lambda = \sigma_{n'}/\tau$ is a positive parameter that controls the tradeoff between data fidelity and the constraint on the sparse noise.

The resulting optimization problem in Equation (15) combines a prior knowledge about the representation coefficients $\boldsymbol{\alpha}$ (meaning that $\boldsymbol{\alpha}$ is sparse), a statistical model for the observation $\mathbf{x}$ in Equation (7) expressed as $\mathbf{x} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\alpha} + \mathbf{s}'; \sigma_{n'}\mathbf{I})$, and a prior model for the sparse noise $\mathbf{s}' \propto exp(-\frac{1}{2\tau}\|\mathbf{s}'\|_1)$. We solve this problem by an alternating minimization algorithm described later (Section 3.4).

Once the sparse coefficients are obtained, we can calculate the class of $\mathbf{x}$ by

$$class(\mathbf{x}) = \arg\min_{i=1,2,\ldots,C} \|\mathbf{x} - \mathbf{D}_i\boldsymbol{\alpha}_i - \mathbf{s}'\|_2, \tag{16}$$

where $\boldsymbol{\alpha}_i$ is a sparse vector associated with class i.

3.2. Robust JSRC Model

Similar to Equation (4), by gathering all the neighbouring pixels around a central test pixel into a matrix $\mathbf{X}$, we can rewrite the Equation (10) in matrix form as follows:

$$\mathbf{X} = \mathbf{D}\mathbf{A} + \mathbf{S} + \mathbf{N}, \tag{17}$$

where $\mathbf{S} \in \mathbb{R}^{B\times T}$ and $\mathbf{N} \in \mathbb{R}^{B\times T}$ are the corresponding matrices representing sparse noise and Gaussian noise, respectively. With the assumption as in the JSRC model that the pixels in a small patch share the same set of training samples, the proposed optimization problem with respect to $\mathbf{A}$ and $\mathbf{S}$ can be formulated as:

$$\begin{aligned} \min f(\mathbf{A},\mathbf{S}) &= \min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{A} - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1 \\ &s.t.\ \|\mathbf{A}\|_{row,0} \leq K_0, \end{aligned} \tag{18}$$

where $\|\mathbf{S}\|_1$ is a norm defined as $\|\mathbf{S}\|_1 = \sum_{i,j} |S_{i,j}|$.

After finding the sparse coefficient matrix $\mathbf{A}$ and the sparse noise matrix $\mathbf{S}$, we can label the class of the central pixel by

$$class(\mathbf{x}_{central}) = \arg\min_{i=1,2,\ldots,C} \|\mathbf{X} - \mathbf{D}_i\mathbf{A}_i - \mathbf{S}\|_F, \tag{19}$$

where $\mathbf{A}_i$ denotes the sparse matrix of $\mathbf{A}$ corresponding to class i.

3.3. Robust Super-Pixel Level JSRC

Imposing that pixels within a fixed-size rectangular neighbourhood share the same sparsity pattern, as in JSRC, has the following limitations. First, the size of the window is a free parameter, and determining its optimal value requires some tuning that varies from one image to the other. Secondly, when the central pixel is located on or near the boundaries between different classes, its neighbouring pixels belong to multiple classes, violating the assumption of the JSRC model and causing classification errors in these border regions. Finally, in practice, both the shape and the size of nearly homogeneous regions may vary a lot across a real scene, which suggests adaptive neighbourhoods instead of the fixed ones. The price to pay for such adaptive instead of fixed neighbourhoods is that a certain type of segmentation is needed. However, it turns out that such an approach with adaptive neighbourhoods may be advantageous not only in terms of accuracy, but also in terms of the net computation time, since each small region can be classified simultaneously as we show next.

We develop here a robust JSRC model, where the spatial information is captured at a super-pixel level, instead of using fixed-size rectangular neighbourhoods. Super-pixel segmentation techniques [43] adaptively divide the image into non-overlapping super-pixels being nearly homogeneous regions according to some criterion. In our problem, each super-pixel is a relatively small arbitrarily shaped and nearly homogeneous region, composed of pixels that belong to the same class. Let $\mathbf{X}$ now denote a matrix composed of pixels within the same super-pixel. With the same reasoning as in the previous section, we assume the model in Equation (17). Note that now the size of $\mathbf{X}$ is not fixed, but, otherwise, the formal description remains equivalent to the previous one, with the optimisation problem defined in Equation (18).

An important difference, both formally and practically, is that now we can assign $\mathbf{X}$ simultaneously to a given class instead of its central pixel alone in Equation (19). Now, we have

$$class(\mathbf{X}) = \underset{i=1,2,\ldots,C}{\arg\min} \|\mathbf{X} - \mathbf{D}_i\mathbf{A}_i - \mathbf{S}\|_F. \tag{20}$$

Here, the class label of a super-pixel is simultaneously calculated, which means also that the sparse coding problem, calculation of class residuals and the minimization over these is calculated only once per non-overlapping super-pixel. On the contrary, in Section 3.2, all these operations are performed in each sliding window, centred around each image pixel. A typical hyperspectral image in remote sensing often has the size of thousands by thousands or more amounting to over million pixels, while we segment it into a couple of hundreds or thousands of super-pixels. This indicates a tremendous saving in computation. The concrete example are given in Section 4.2.

3.4. Optimization Algorithm

Here, we present an optimization algorithm to solve the proposed robust model by an alternating minimization strategy. A general derivation for the optimization in a matrix form is shown in Algorithm 1, where the input matrix $\mathbf{X}$ can represent a patch in R-JSRC or a super-pixel in R-SJSRC or reduce to a single vector in R-SRC. We employ alternating minimization similarly as in [28,31,36,44] to split a difficult problem into two easily solvable ones by fixing one variable in the other sub-problem, and alternating the process iteratively. In the $(k+1)$th iteration, we update $\mathbf{A}$ and $\mathbf{S}$ as follows:

$$\mathbf{A}^{(k+1)} = \underset{\|\mathbf{A}\|_{row,0} \leq K_0}{\arg\min} f(\mathbf{A}, \mathbf{S}^{(k)}), \tag{21}$$

$$\mathbf{S}^{(k+1)} = \underset{\mathbf{S}}{\arg\min}\, f(\mathbf{A}^{(k+1)}, \mathbf{S}). \tag{22}$$

Problem in Equation (21) can be solved by the SOMP algorithm [42], and for problem in Equation (22), the optimization with respect to $\mathbf{S}^{(k+1)}$ is formulated by

$$\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{D}\mathbf{A}^{(k+1)} - \mathbf{S}\|_F^2 + \lambda||\mathbf{S}||_1, \tag{23}$$

which is the well-known shrinkage problem. By introducing the following soft-thresholding operator:

$$\Re_\Delta(x) = \begin{cases} sgn(x)(|x| - \Delta), & if\ |x| \geq \Delta, \\ 0, & if\ |x| < \Delta, \end{cases} \tag{24}$$

the solution of Equation (23) could be given by

$$\mathbf{S}^{(k+1)} = \Re_{\lambda/2}(\mathbf{X} - \mathbf{D}\mathbf{A}^{(k+1)}). \tag{25}$$

Note that, for the vector form of R-SRC in Algorithm 1, the sparse coefficients $\boldsymbol{\alpha}$ in step 4 are obtained by OMP algorithm [41] and $\mathbf{s}'$ in step 5 is derived by $\Re_{\lambda/2}(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})$. The class in step 8 is labeled by Equation (16).

Algorithm 1 Generic pseudo-code of the proposed approach

1: **Input**: input matrix $\mathbf{X}$, dictionary $\mathbf{D}$, K_0 and λ
2: **Initialize**: $\mathbf{A}^{(0)} = \mathbf{0}$, $\mathbf{S}^{(0)} = \mathbf{0}$ and $k = 0$
3: **While** stop criterion is not satisfied **do**
4: Obtain $\mathbf{A}^{(k+1)}$ by solving the sub-problem in Equation (21)
5: Obtain $\mathbf{S}^{(k+1)}$ by Equation (25)
6: **end**
7: **Return**: $\mathbf{A} = \mathbf{A}^{(k+1)}$, $\mathbf{S} = \mathbf{S}^{(k+1)}$
8: **Output**: Class label is obtained by residuals (Equation (19) for R-JSRC and Equation (20) for R-SJSRC).

4. Experiments

We evaluate the performance of our methods on both simulated and real hyperspectral images, in comparison with SVM with radial basis function (RBF) kernel [45], SRC [19], JSRC [20] and NLW-JSRC [21]. As quantitative performance measures, we adopt the common indicators: overall accuracy (OA), average accuracy (AA) and Kappa coefficient (κ). All the reported results represent the average of ten runs. In each run, the training samples are randomly selected and the remaining labeled samples are used for testing.

4.1. Results for the Simulated HSI Experiment

The Washington DC image shown in Figure 2a was collected by the HYDICE. Due to its high quality, this image is commonly used to simulate data degraded with different kinds of noise. The image is of size 280 × 307 × 210 with the spectrum ranging from 0.4 to 2.4 μm and has six classes in total. In this experiment, we reduce the number of bands to 191 by removing the opaque bands. Five percent of labeled samples were randomly selected as training samples and the remainder as test samples as shown in Table 1.

Experiment 1 (Synthetic simulation): In this simulated experiment, four kinds of noise were added as follows:

1. Zero-mean Gaussian noise in all bands with SNR value for each band varying from 10 to 20 dB.
2. Impulse noise with 20% of corrupted pixels in bands 30–40.
3. Dead lines in bands 70–73 with width ranging from one line to three lines.

4. Strips in bands 101–104 with width ranging from one line to three lines.

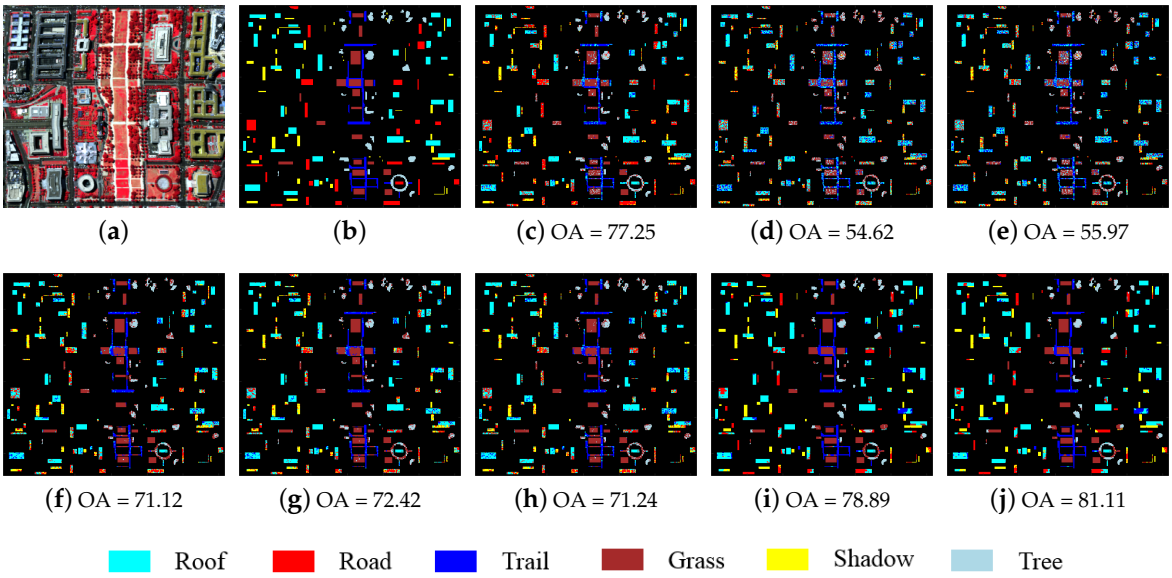


Figure 2. Washington DC image. (**a**) false color image; (**b**) ground truth; and classification results (OA in percentage) obtained by (**c**) SVM; (**d**) SRC; (**e**) R-SRC; (**f**) JSRC; (**g**) R-JSRC; (**h**) NLW-JSRC; (**i**) SJSRC; (**j**) R-SJSRC.

Table 1. Results for simulated data with different classifiers *.

Class	Class Name	Train	Test	SVM	SRC	R-SRC	JSRC	R-JSRC	NLW-JSRC	SJSRC	R-SJSRC
1	Roof	146	2770	0.7466	0.5842	0.5873	0.7727	0.7918	0.7790	0.7897	**0.7962**
2	Road	91	1728	**0.6742**	0.4100	0.4197	0.5219	0.5374	0.5204	0.5122	0.5425
3	Trail	64	1200	0.7585	0.6900	0.7070	0.7417	0.7540	0.7543	**0.9110**	0.9099
4	Grass	90	1700	0.8726	0.7536	0.7712	0.9463	0.9496	0.9468	0.9801	**0.9834**
5	Shadow	56	1064	0.7087	0.4234	0.4487	0.5778	0.5738	0.5617	0.8237	**0.8273**
6	Tree	65	1216	**0.7970**	0.4792	0.5042	0.5954	0.6038	0.5881	0.6846	0.7160
	OA			0.7595 ±0.0095	0.5650 ±0.0087	0.5787 ±0.0101	0.7109 ±0.0142	0.7218 ±0.0148	0.7114 ±0.0156	0.7792 ±0.0208	**0.7912 ±0.0192**
	AA			0.7596 ±0.0129	0.5567 ±0.0142	0.5730 ±0.0154	0.6941 ±0.0144	0.7017 ±0.0152	0.6917 ±0.0160	0.7836 ±0.0230	**0.7959 ±0.0190**
	κ			0.7034 ±0.0119	0.4623 ±0.0123	0.4797 ±0.0140	0.6421 ±0.0174	0.6553 ±0.0182	0.6426 ±0.0192	0.7284 ±0.0258	**0.7432 ±0.0232**

* Abbreviations: support vector machines (SVM), sparse representation classification (SRC), robust SRC (R-SRC), joint SRC (JSRC), robust JSRC (R-JSRC), nonlocal weighted JSRC (NLW-JSRC), super-pixel level JSRC (SJSRC) and robust SJSRC (R-SJSRC). The best result in each row is denoted in bold and suboptimal result is underlined.

The optimal parameters of the proposed robust methods were determined empirically as: $\lambda = 0.01$, $K = 5$ for the R-SRC model, $\lambda = 0.02, K_0 = 30, T = 25$ for the R-JSRC model and $\lambda = 0.02$, $K_0 = 30, Ns = 7000$ for the R-SJSRC model, where Ns denotes the number of super-pixels. For other classification methods in Table 1, all the parameters were tuned to give the best results, which are denoted in bold and suboptimal results are underlined. In order to be able to evaluate the contribution of each of the components of the proposed approach separately (both the robust nature and handling of spatial context), we also implemented the super-pixel level joint sparse representation classification (SJSRC) method with the same segmentation map as R-SJSRC. The results in Table 1 and Figure 2 show that the R-SJSRC model yielded a superior performance in terms of OA, AA and Kappa coefficient. The improvement due to the better spatial modelling can be clearly seen by comparing the performance of the super-pixel based SJSRC with the original JSRC. In terms of OA, this improvement was above 9.6%. Further improvement in the performance results from the improved noise model in R-SJSRC

(the OA increases by other 1.5% compared to SJSRC). Similarly, the robust versions R-SRC and R-JSRC improve consistently over the corresponding SRC and JSRC methods, respectively.

Experiment 2 (Effect of sparse noise): In this experiment, we analyse the robustness of our models to degradations dominated by sparse noise. We attempt to simulate a realistic situation where at least a small amount of Gaussian noise is always present and where sparse noise only affects certain bands or pixels of HSIs. Therefore, we first add a small amount of zero mean white Gaussian noise, such that the resulting SNR is 30 dB, and subsequently we introduce sparse noise. Let S_b denote the fraction of bands affected by sparse noise and S_p the fraction of affected pixels in each band. We perform experiments with $S_b = S_p = S \in \{0, 0.05, 0.1, 0.2\}$. The results are reported in Figure 3. R-SJSRC is the most stable method among all the tested ones, while SRC degrades sharply with the increasing level of sparse noise. Clearly, the performance of R-SJSRC is less sensitive to sparse noise than that of other methods. Moreover, the robust methods R-SRC, R-JSRC and R-SJSRC yield consistent improvements over the original models as expected.

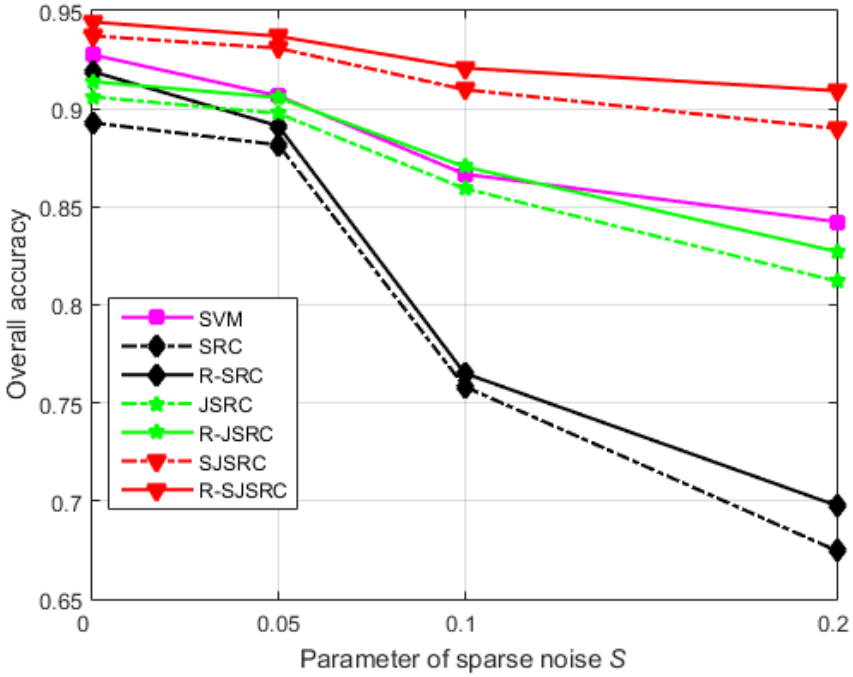


Figure 3. The influence of sparse noise on the classification performance of different classifiers. The overall accuracy is plotted against the parameter S reflecting the level of sparse noise.

Experiment 3 (Effect of sparsity constraint λ): In this experiment, we study the effects of the parameter λ on the classification performance for our methods. The test image was firstly degraded by Gaussian noise such that the SNR is 30dB, and then corrupted by sparse noise with $S_b = S_p = 0.2$. The classification performance for R-SRC, R-JSRC and R-SJSRC is reported in Figure 4. Note that when the parameter λ is set as zero, R-SRC, R-JSRC and R-SJSRC reduce to SRC, JSRC and SJSRC, respectively.

We can observe in Figure 4 that the overall accuracies of the three models show similar trends in a function of the parameter λ. When the value of λ is relatively low, which means we enforce a smaller weight on the sparse noise, the performance of the proposed methods is not significantly improved over the results with $\lambda = 0$. As the value of λ increases, the classification performance also improves, reaching its highest values at $\lambda = 10^{-3}$ for R-JSRC and R-SJSRC, and at $\lambda = 10^{-2}$ for R-SRC. The improvements for R-SRC, R-JSRC and R-SJSRC show the benefit of incorporating the effect of sparse noise in our models.

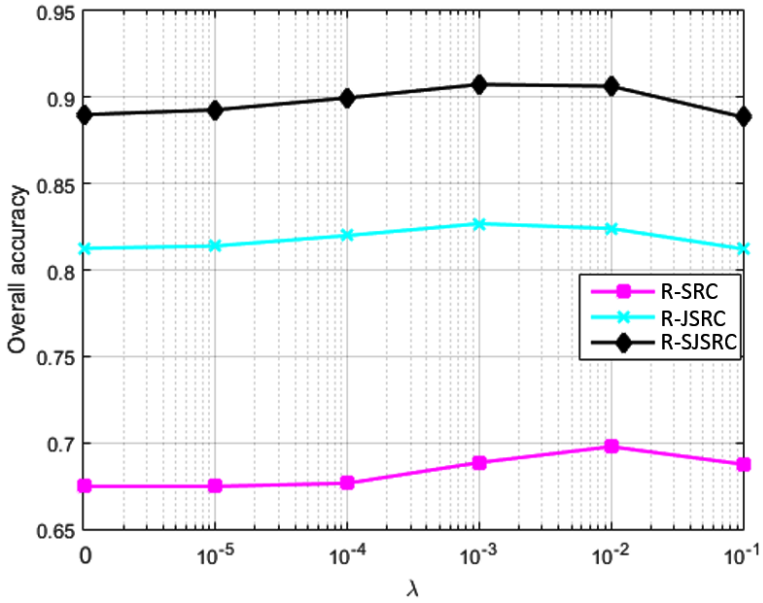


Figure 4. Classification performance of our methods with respect to parameter λ.

4.2. Results for Real HSI Experiment

In this section, two real HSI datasets are used: Indian Pines data set and an urban area HYDICE data set.

4.2.1. Classification Results on the Real Datasets

The first experiment was conducted on the Indian Pines image, which was acquired by the Airborne/Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in northwestern Indiana in 1992 as shown in Figure 5a. This image has 16 classes and 220 spectral reflectance bands ranging from 0.4 to 2.5μm. In this experiment, 20 water absorption spectral bands in 104–108, 150–163 and 200 are removed; therefore, the real hyperspectral image size is $145 \times 145 \times 200$. Nine percent of the labeled samples are randomly selected as training samples and the remainder as test samples as shown in Table 2.

Table 2. Reference classes for the Indian Pines.

No.	Class Name	Train	Test
1	Alfalfa	6	40
2	Corn-notill	129	1299
3	Corn-mintill	83	747
4	Corn	24	213
5	Grass-pasture	48	435
6	Grass-trees	73	657
7	Grass-pasture-mowed	5	23
8	Hay-windrowed	48	430
9	Oats	4	16
10	Soybean-notill	97	875
11	Soybean-mintill	196	2259
12	Soybean-clean	59	534
13	Wheat	21	184
14	Woods	114	1151
15	Bldgs-grass-trees-drives	39	347
16	Stone-steel-towers	12	81
	Total	958	9291

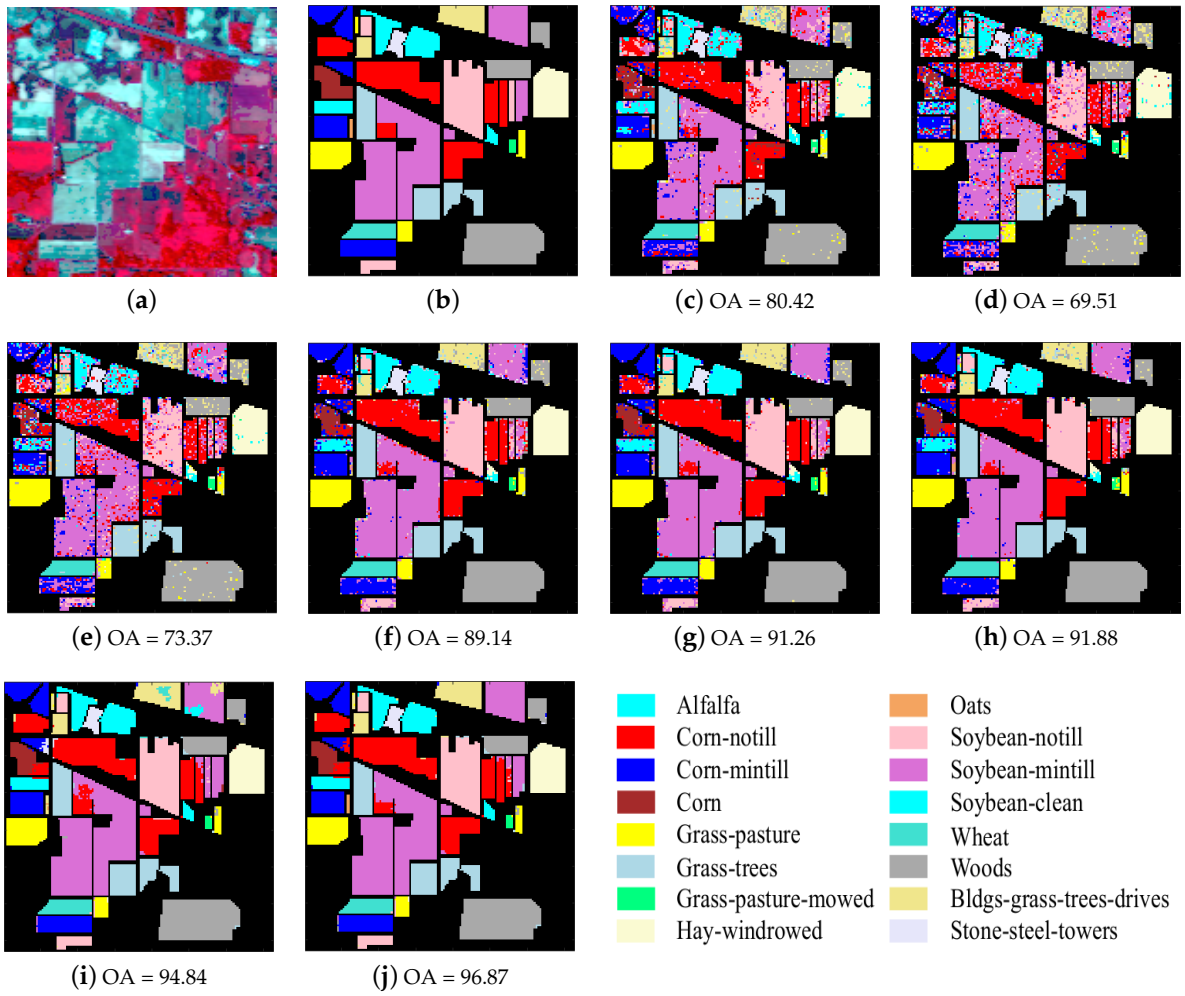


Figure 5. Indian Pines image. (**a**) false color image; (**b**) ground truth; and classification results (OA in percentage) obtained by (**c**) SVM; (**d**) SRC; (**e**) R-SRC; (**f**) JSRC; (**g**) R-JSRC; (**h**) NLW-JSRC; (**i**) SJSRC; (**j**) R-SJSRC.

The optimal parameters of our methods were: $\lambda = 4 \times 10^{-4}$, $K = 11$ for R-SRC, $\lambda = 1.5 \times 10^{-3}$, $K_0 = 30, T = 49$ for R-JSRC and $\lambda = 0.003, K_0 = 50, Ns = 300$ for R-SJSRC. For JSRC, the optimal window size was 7×7 and sparsity level was 30. In NLW-JSRC, the parameters were chosen from the recommendation of [21]. For SVM and SRC classifiers, we tuned the parameters such to produce the best classification results. The results are listed in Table 3 and Figure 5. In most cases, our method R-SJSRC yields better results than other classifiers. Based on super-pixel segmentation, the SJSRC model had at least 2.7% improvement over the reference methods JSRC and NLW-JSRC. Considering the sparse prior for multiple noise in the HSIs, our proposed R-SJSRC further improves OA by 1.5% over SJSRC. Moreover, the proposed robust models show a superior performance over SRC, JSRC and SJSRC, respectively. In Table 3, it should be noted that, even though the number of training samples for classes 1, 7 and 9 is very limited, both SJSRC and R-SJSRC still achieve a very high classification accuracy over others, which is largely due to the exploitation of super-pixel segmentation. Both R-SJSRC and SJSRC on a super-pixel level classification are able to alleviate the effect of unbalanced training samples on the performance to a certain degree.

Table 3. Overall classification accuracy for Indian Pines with different classifiers.

Class	SVM	SRC	R-SRC	JSRC [20]	R-JSRC	NLW-JSRC [21]	SJSRC	R-SJSRC
1	0.6275	0.4125	0.5075	0.5625	0.6350	0.5950	**0.9800**	**0.9800**
2	0.7807	0.6122	0.6546	0.8570	0.8780	0,8917	**0.9799**	0.9427
3	0.7106	0.5396	0.5750	0.8371	0.8541	0,8617	**0.9601**	0.9426
4	0.5362	0.3286	0.3770	0.6892	0.7469	0,7113	**0.9920**	0.8441
5	0.8968	0.8478	0.8678	0.9159	0.9292	**0,9366**	0.9172	0.9163
6	0.9534	0.9307	0.9470	0.9962	0.9970	0,9976	**1.0000**	0.9976
7	0.8130	0.7565	0.8261	0.6304	0.6652	0,6783	**0.9696**	**0.9696**
8	0.9584	0.9170	0.9598	0.9988	0.9993	**0,9995**	0.9977	0.9977
9	0.5813	0.5125	0.5813	0.4125	0.4750	0,6625	**1.0000**	0.8000
10	0.7506	0.6103	0.6466	0.8312	0.8519	0,8665	0.8574	**0.9271**
11	0.8053	0.7000	0.7291	0.8726	0.8977	0,9137	0.9099	**0.9508**
12	0.7315	0.5075	0.5772	0.8384	0.8936	0,9026	0.9296	**0.9700**
13	0.9544	0.9538	0.9668	**0.9967**	**0.9967**	**0,9967**	0.9951	0.9951
14	0.9308	0.9056	0.9145	0.9791	0.9815	**0,9856**	0.9569	0.9818
15	0.5545	0.4596	0.4937	0.7960	0.8499	0,8369	0.8939	**0.9677**
16	0.9346	0.8531	0.8605	0.9840	0.9852	**0,9938**	0.9790	0.9679
OA	0.8096 ±0.0066	0.7015 ±0.0039	0.7333 ±0.0024	0.8851 ±0.0047	0.9055 ±0.0040	0.9137 ±0.0064	0.9407 ±0.0008	**0.9547 ±0.0095**
AA	0.7825 ±0.0211	0.6780 ±0.0137	0.7178 ±0.0155	0.8248 ±0.0226	0.8523 ±0.0228	0.8644 ±0.0283	**0.9574 ±0.0016**	0.9469 ±0.0271
κ	0.7827 ±0.0074	0.6588 ±0.0043	0.6952 ±0.0026	0.8690 ±0.0053	0.8921 ±0.0046	0.9014 ±0.0074	0.9325 ±0.0009	**0.9483 ±0.0109**

We also test the computation time saving of R-SJSRC compared to R-JSRC. The experiment was implemented in Matlab R2015a on the computer with Intel Core i7-3930K CPU and 64 GB RAM, and recorded time consumption of one iteration including super-pixel segmentation and classification map generation for R-SJSRC and classification map generation for R-JSRC. The results show that R-JSRC spends 321 s, while R-SJSRC only takes 5 s for one iteration, which indicates the benefit of R-SJSRC in terms of time saving. The reason for the high complexity of R-JSRC mainly comes from the computation of sparse coefficient when using the sliding window, which has to be calculated multiple times.

The second image that we use for evaluation is HYDICE Urban captured by the HYDICE sensor [46]. The original image size is of $307 \times 307 \times 210$ and there are five classes in total. In this experiment, we tested our method on a part of this image with size 200×200 as shown in Figure 6a. The number of bands was reduced to 188 by removing the bands 104–108, 139–151 and 207–210, which were seriously polluted by the atmosphere and water absorption. We used this image as it contains different types of noise including strips, dead lines, impulse noise and Gaussian noise [46]. The number of samples used for training and test are shown in Table 4.

The quantitative results and classification maps from different methods are shown in Table 4 and Figure 6. The optimal parameters of R-SRC, R-JSRC and R-SJSRC methods are obtained, respectively, by $\lambda = 2.5 \times 10^{-3}, K = 4$, $\lambda = 0.01, K_0 = 10, T = 25$ and $\lambda = 0.01, K_0 = 12, Ns = 1450$. For other classification methods, we tuned the parameters in order to yield the best results. The results in Table 4 and Figure 6 show clearly that the proposed R-SJSRC model performs better than other classification methods on the HYDICE Urban image in terms of quantitative measurements and visual evaluation. A superior performance can be also viewed for other robust models, i.e., R-SRC and R-JSRC, over SRC and JSRC.

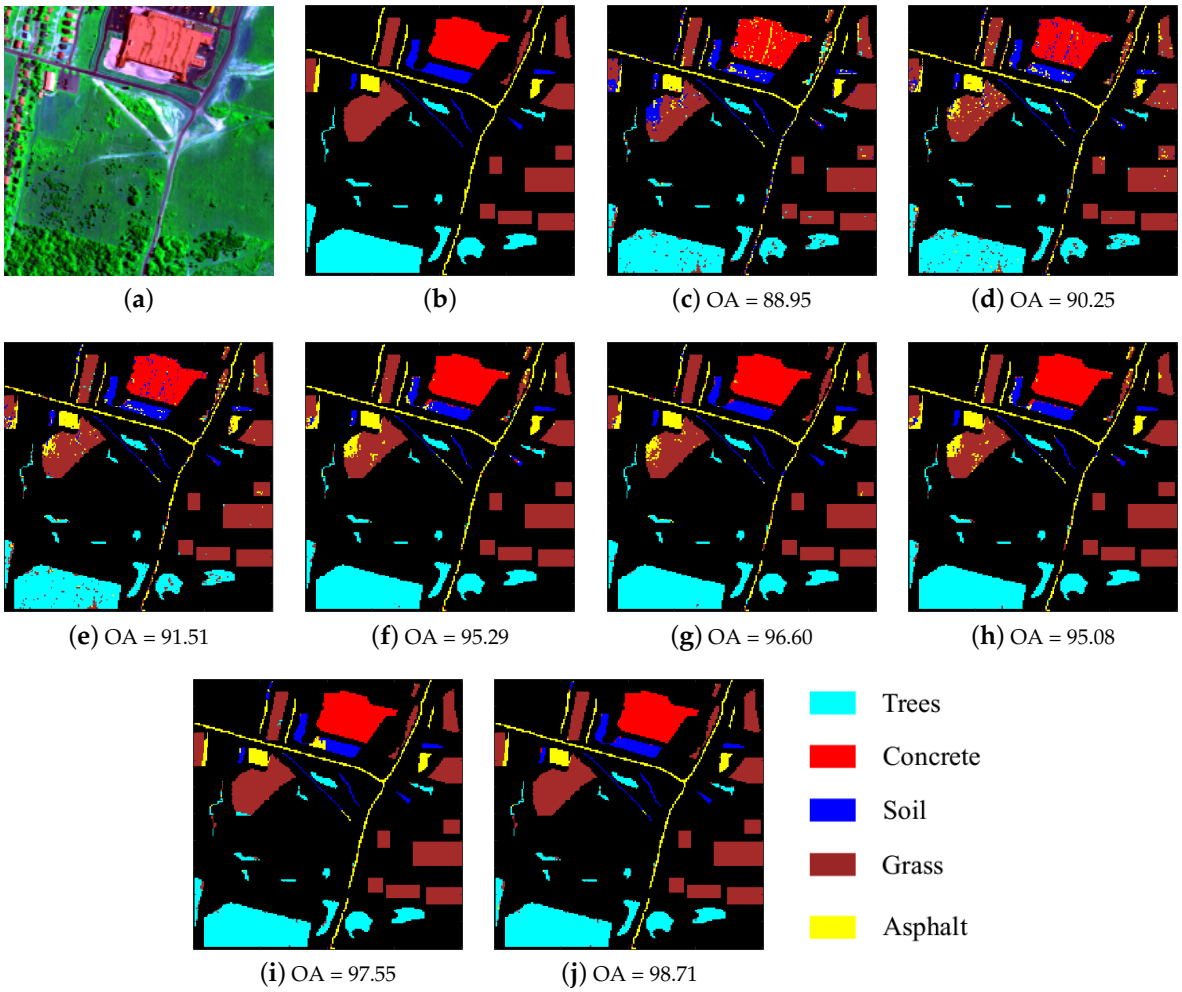


Figure 6. HYDICE Urban data. (**a**) false color image; (**b**) ground truth; and classification results (OA in percentage) obtained by (**c**) SVM; (**d**) SRC; (**e**) R-SRC; (**f**) JSRC; (**g**) R-JSRC; (**h**) NLW-JSRC; (**i**) SJSRC; (**j**) R-SJSRC.

Table 4. Overall classification accuracy for urban with different classifiers.

Class	Class Name	Train	Test	SVM	SRC	R-SRC	JSRC	R-JSRC	NLW-JSRC	SJSRC	R-SJSRC
1	Trees	30	3093	0.9251	0.9230	0.9269	0.9817	**0.9856**	0.9812	0.9691	0.9737
2	Concrete	30	1380	0.9696	0.9787	0.9874	0.9978	0.9990	0.9977	**1**	0.9999
3	Soil	30	607	0.8638	0.8359	0.8624	0.7685	0.8891	0.7802	0.8611	**0.9432**
4	Grass	30	4014	0.9055	0.8984	0.9208	0.9421	0.9590	0.9426	**0.9915**	0.9863
5	Asphalt	30	882	0.7832	0.7953	0.8194	0.9117	0.9179	0.9027	0.9621	**0.9711**
	OA			0.9071 ±0.0123	0.9041 ±0.0139	0.9191 ±0.0114	0.9488 ±0.0086	0.9649 ±0.0049	0.9488 ±0.0083	0.9752 ±0.0003	**0.9803 ±0.0058**
	AA			0.8894 ±0.0107	0.8863 ±0.0094	0.9034 ±0.0073	0.9204 ±0.0094	0.9501 ±0.0084	0.9209 ±0.0093	0.9568 ±0.0007	**0.9748 ±0.0094**
	κ			0.8706 ±0.0166	0.8662 ±0.0189	0.8869 ±0.0156	0.9284 ±0.0119	0.9508 ±0.0069	0.9284 ±0.0115	0.9651 ±0.0004	**0.9723 ±0.0081**

4.2.2. The Effect of the Training Sample Size

Here, we examine the effect of the training set size on the classification performance, using HYDICE Urban image as a case study. The number of training samples per class was set as 5, 10, 20, 40, 80 and 160, respectively, and the parameters for different methods were fixed as earlier specified. The results shown in Figure 7 reveal that the OA of all the methods gets improved

significantly with the increase of training sample size, and R-SJSRC consistently achieves the best performance over all other tested methods. It can be observed that the highest improvement of R-SJSRC over SJSRC, as well as R-JSRC over JSRC and R-SRC over SRC is obtained when the number of training samples is the smallest (five per class). This improvement, resulting from accounting for the sparse noise in our model, turns out to be less significant when the size of the training set increases. This demonstrates that our robust model is especially effective when the training samples are limited.

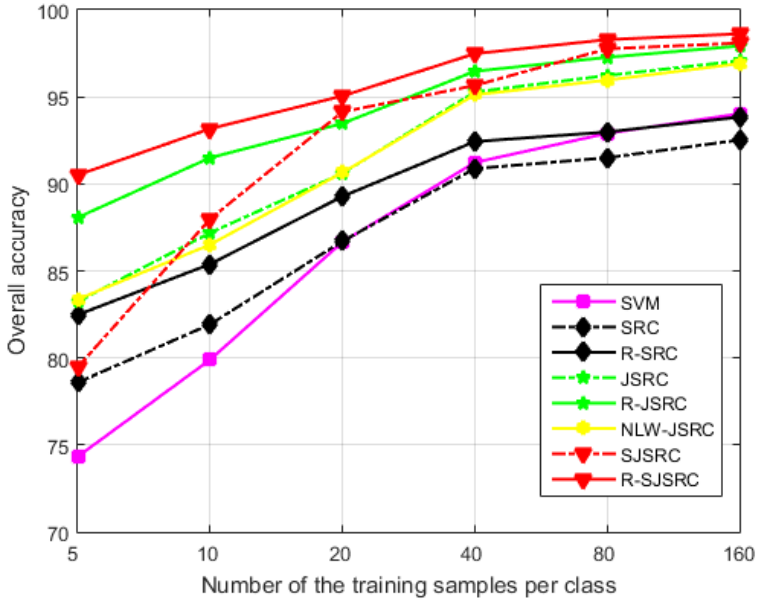


Figure 7. The influence of the training sample size on the performance of different methods. The test image is HYDICE Urban.

4.2.3. The Influence of the Segmentation Granularity

To investigate the influence of the segmentation granularity on the performance of R-SJSRC, we conduct the experiments with varying number of super-pixels and record the resulting overall classification accuracy. Figure 8 shows the results for HYDICE Urban image and Indian Pines, where the number of super-pixel Ns is ranging from 100 to 3200, and the parameters of R-SJSRC are as specified earlier. The results demonstrate that the OA of HYDICE Urban image is less sensitive to the number of super-pixels than Indian Pines image for R-SJSRC. The OA of HYDICE Urban image stably increases to 97.93% when the value of Ns is less than 1600. The OA of Indian Pines image rapidly reaches to the top of 95.47% at $Ns = 300$, and then drops down to 85% at $Ns = 3200$. The reason for the stronger sensitivity of Indian Pines to Ns may be caused by the large diversity of the ground truth in the same class. When the number of super-pixels is large, more homogeneous regions will be separated into many small pieces, which results in the constraint relaxation of joint sparsity for the pixels in the same super-pixel and deteriorates the performance of R-SJSRC.

4.3. Practical Specification of the Parameters

In our experiments, we make sure that the comparison between different methods is fair by presenting for all of them the best achievable performance, assuming that the parameters were set optimally. In practice, ground truth data are rarely available. We advise the user in this case to optimize the parameters (using e.g., a widely adopted grid search) for images that are similar (in resolution and variability) to the ones being tested and for which ground truth data are available. The parameter values that we give may also be used without extensive decrease of the performance on a wide range of images of two types: AVIRIS and high-resolution urban images. The diagrams where we report

the influence of the different parameters should also serve as a useful guideline in this respect. Figure 4 shows that λ can be chosen in a relatively wide range around the optimal value, without strongly affecting the performance. The same holds for the segmentation parameter, especially for Urban types of images.

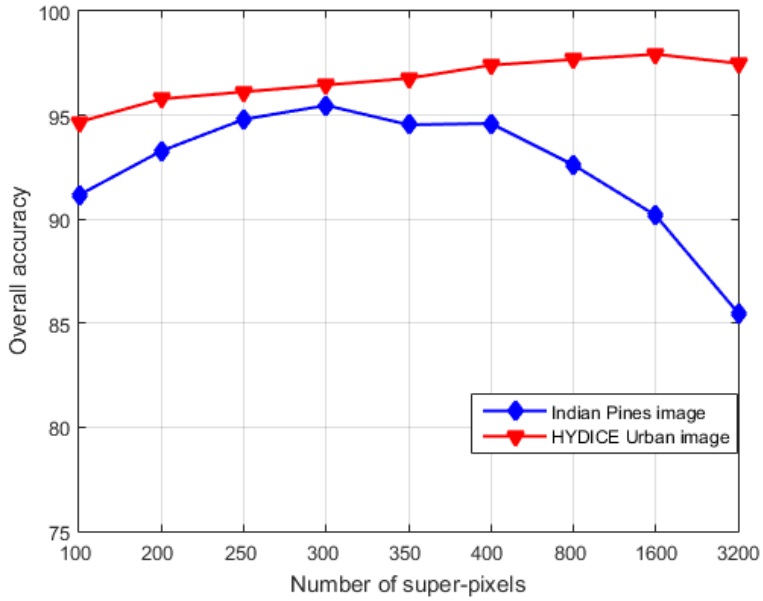


Figure 8. The effect of the number of super-pixels on the performance of R-SJSRC with two real HSIs: Indian Pines and HYDICE Urban.

5. Conclusions

In this work, we have proposed a robust classification model for HSIs, which combines an appropriate statistical model for the sparse noise and the representation coefficients of test samples into a unified framework, explicitly accounting for both Gaussian noise and sparse noise. An alternating minimization strategy is utilized to solve the resulting optimization problems. The robust model can easily generalize the off-the-shelf classification model to a robust version. The superior performance of the proposed methods over the existing methods is confirmed by the experiments on both real and simulated data, which demonstrates the effectiveness of the proposed robust model.

Acknowledgments: This work was supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) project: G.OA26.17N Dictionary Learning and Distributed Inference for the Processing of Large-Scale Heterogeneous Image Data (DOLPHIN), by grants from the China Scholarship Council (CSC) and UGent Bijzonder Onderzoeksfonds (BOF) cofunding-CSC.

Author Contributions: Shaoguang Huang and Hongyan Zhang conceived and designed the robust sparsity-based classification methods for hyperspectral images and Aleksandra Pižurica formulated the derivation of the model; Hongyan Zhang designed the experiments for validation of our model; Shaoguang Huang performed the experiments and Aleksandra Pižurica helped to formulate the overall approach and to analyse the results; the final manuscript was finished by Shaoguang Huang and revised by Aleksandra Pižurica and Hongyan Zhang.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Datt, B.; McVicar, T.R.; Van Niel, T.G.; Jupp, D.L.; Pearlman, J.S. Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1246–1259.

2. Lee, M.A.; Huang, Y.; Yao, H.; Thomson, S.J.; Bruce, L.M. Determining the effects of storage on cotton and soybean leaf samples for hyperspectral analysis. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2562–2570.
3. Eismann, M.T.; Stocker, A.D.; Nasrabadi, N.M. Automated hyperspectral cueing for civilian search and rescue. *IEEE Proc.* **2009**, *97*, 1031–1055.
4. Camps-Valls, G.; Tuia, D.; Bruzzone, L.; Benediktsson, J.A. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Process. Mag.* **2014**, *31*, 45–54.
5. Zhang, B.; Wu, D.; Zhang, L.; Jiao, Q.; Li, Q. Application of hyperspectral remote sensing for environment monitoring in mining areas. *Environ. Earth Sci.* **2012**, *65*, 649–658.
6. Ratle, F.; Camps-Valls, G.; Weston, J. Semisupervised neural networks for efficient hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2271–2282.
7. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098.
8. Roscher, R.; Waske, B.; Forstner, W. Incremental import vector machines for classifying hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3463–3473.
9. Chen, C.; Li, W.; Tramel, E.W.; Cui, M.; Prasad, S.; Fowler, J.E. Spectral–spatial preprocessing using multihypothesis prediction for noise-robust hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 1047–1059.
10. Prasad, S.; Li, W.; Fowler, J.E.; Bruce, L.M. Information fusion in the redundant-wavelet-transform domain for noise-robust hyperspectral classification. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3474–3486.
11. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790.
12. Tuia, D.; Ratle, F.; Pozdnoukhov, A.; Camps-Valls, G. Multisource composite kernels for urban-image classification. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 88–92.
13. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491.
14. Dalla Mura, M.; Benediktsson, J.A.; Waske, B.; Bruzzone, L. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3747–3762.
15. Liao, W.; Bellens, R.; Pizurica, A.; Philips, W.; Pi, Y. Classification of hyperspectral data over urban areas using directional morphological profiles and semi-supervised feature extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1177–1190.
16. Liao, W.; Pizurica, A.; Bellens, R.; Gautama, S.; Philips, W. Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 552–556.
17. Liao, W.; Dalla Mura, M.; Chanussot, J.; Bellens, R.; Philips, W. Morphological Attribute Profiles With Partial Reconstruction. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1738–1756.
18. Prasad, S.; Cui, M.; Li, W.; Fowler, J.E. Segmented mixture-of-Gaussian classification for hyperspectral image analysis. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 138–142.
19. Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; Ma, Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 210–227.
20. Chen, Y.; Nasrabadi, N.M.; Tran, T.D. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3973–3985.
21. Zhang, H.; Li, J.; Huang, Y.; Zhang, L. A nonlocal weighted joint sparse representation classification method for hyperspectral imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2056–2065.
22. Chen, Y.; Nasrabadi, N.M.; Tran, T.D. Hyperspectral image classification via kernel sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 217–231.
23. Fang, L.; Li, S.; Kang, X.; Benediktsson, J.A. Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7738–7749.
24. Wang, J.; Jiao, L.; Liu, H.; Yang, S.; Liu, F. Hyperspectral Image Classification by Spatial–spectral Derivative-Aided Kernel Joint Sparse Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2485–2500.
25. Li, J.; Zhang, H.; Zhang, L. Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 5338–5351.

26. Fu, W.; Li, S.; Fang, L.; Kang, X.; Benediktsson, J.A. Hyperspectral Image Classification Via Shape-Adaptive Joint Sparse Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 556–567.
27. Bian, X.; Chen, C.; Xu, Y.; Du, Q. Robust Hyperspectral Image Classification by Multi-Layer Spatial–Spectral Sparse Representations. *Remote Sens.* **2016**, *8*, 985.
28. Chen, C.; Chen, N.; Peng, J. Nearest Regularized Joint Sparse Representation for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 424–428.
29. Wang, Z.; Nasrabadi, N.M.; Huang, T.S. Spatial–spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4808–4822.
30. Wang, Z.; Nasrabadi, N.M.; Huang, T.S. Semisupervised hyperspectral classification using task-driven dictionary learning with Laplacian regularization. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1161–1173.
31. Fang, L.; Li, S.; Kang, X.; Benediktsson, J.A. Spectral–spatial classification of hyperspectral images with a superpixel-based discriminative sparse model. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4186–4201.
32. Soltani-Farani, A.; Rabiee, H.R.; Hosseini, S.A. Spatial-aware dictionary learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 527–541.
33. Sun, X.; Nasrabadi, N.M.; Tran, T.D. Task-driven dictionary learning for hyperspectral image classification with structured sparsity constraints. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4457–4471.
34. Zhang, H.; Zhai, H.; Zhang, L.; Li, P. Spectral-Spatial Sparse Subspace Clustering for Hyperspectral Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3672–3684.
35. He, W.; Zhang, H.; Zhang, L. Sparsity-regularized robust non-negative matrix factorization for hyperspectral unmixing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4267–4279.
36. He, W.; Zhang, H.; Zhang, L.; Shen, H. Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 178–188.
37. Zhang, H.; He, W.; Zhang, L.; Shen, H.; Yuan, Q. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4729–4743.
38. Aggarwal, H.; Majumdar, A. Hyperspectral unmixing in the presence of mixed noise using joint-sparsity and total variation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4257–4266.
39. Giannakis, G.B.; Mateos, G.; Farahmand, S.; Kekatos, V.; Zhu, H. USPACOR: Universal sparsity-controlling outlier rejection. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 1952–1955.
40. Huang, S.; Zhang, H.; Liao, W.; Pizurica, A. Robust joint sparsity model for hyperspectral image classification. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017.
41. Tropp, J.A.; Gilbert, A.C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666.
42. Tropp, J.A.; Gilbert, A.C.; Strauss, M.J. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Process.* **2006**, *86*, 572–588.
43. Liu, M.; Tuzel, O.; Ramalingam, S.; Chellappa, R. Entropy rate superpixel segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 2097–2104.
44. Pham, D.; Venkatesh, S. Joint learning and dictionary construction for pattern recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
45. Bruzzone, L.; Chi, M.; Marconcini, M. A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 3363–3373.
46. Yuan, Q.; Zhang, L.; Shen, H. Hyperspectral image denoising employing a spectral–spatial adaptive total variation model. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 3660–3677.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).