Unsupervised Band Selection in Hyperspectral Images using Autoencoder

Mateus Habermann, Vincent Frémont, Elcio H. Shiguemori

Unsupervised Band Selection in Hyperspectral Images using Autoencoder

M. Habermann[1,2], **V. Fremont**[1], **E.H. Shiguemori**[2]

[1]Sorbonne Universités, Université de Technologie de Compiègne,
CNRS, Heudiasyc UMR 7253, CS 60319, 60203 Compiègne cedex, France.
[2]Institute for Advanced Studies, Brazilian Air Force, Brazil.

Keywords: Band selection, Autoencoder, Masking noise, Hyperspectral Images, Ranking approach.

Abstract

Hyperspectral images provide fine details of the observed scene from the exploitation of contiguous spectral bands. However, the high dimensionality of hyperspectral images causes a heavy burden on processing. Therefore, a common practice that has been largely adopted is the selection of bands before processing. Thus, in this work, a new unsupervised approach for band selection based on autoencoders is proposed. During the training phase of the autoencoder, the input data samples have some of their features turned to zero, through a masking noise transform. The subsequent reconstruction error is assigned to the indices with masking noise. The bigger the error, the greater the importance of the masked features. The errors are then summed up during the whole training phase. At the end, the bands corresponding to the biggest indices are selected. A comparison with four other band selection approaches reveals that the proposed method yields better results in some specific cases and similar results in other situations.

1 Introduction

Hyperspectral images (HSI) consist of many continuous spectrum bands with high resolution [2]. This means that a broad range of spectrum could be covered, providing, consequently, lots of information about the observed scene. Therefore, those images are beneficial to Remote Sensing (RS) tasks, such as image classification and target detection for Earth analysis.

However, a large amount of bands can cause difficulties for data transmission, storage and processing. In addition to that, this high number of bands can lead to the so-called Hughes phenomenon —or curse of dimensionality [7]. These facts represent a challenge in numbers of RS applications and in HSI data classification, for example. So, it normally requires dimensionality reduction.

Feature extraction (FE) is a common method used for dimensionality reduction. Under the FE approach, new features are generated by linear, or non-linear combinations of the original ones. The new features have a lower dimension, and they still retain much of the original information. Principal Component Analysis (PCA) is a very popular FE technique. A negative aspect of feature extraction is that it mixes up the spectral bands, and this may be a problem when the original band information is necessary [8].

Another possibility for dimensionality reduction is *band selection* (BS). BS seeks to reduce the dimensionality of the original data without losing much useful information. When it comes to classification tasks, the focus of a BS method is done on the bands that provide a good class separability. Because band selection algorithms preserve the original information of the HSI, the results are more interpretable as presented in [9].

Further, BS methods can be divided into two groups: *supervised* and *unsupervised*. The supervised band selection methods normally have better performances compared to unsupervised ones as highlighted in [14]. However, supervised approaches require labeled training samples, which are very expensive to be gathered. Thus, unsupervised methods are a feasible and wise choice for band selection.

The selection of bands can normally be done by ranking, clustering, or searching methods [8] . Since unsupervised BS approaches do not rely on class information, the aim of such methods is to select features that preserve the structure of the original dataset [1].

Autoencoders have the ability of learning the structure of a dataset. An autoencoder (AE) can be considered as an artificial neural network, and it is used to learn a representation of the data samples in an unsupervised fashion, since such architectures seek to reconstruct the input vector [6].

Inspired by [1], this paper proposes an unsupervised BS framework based on simple autoencoders, that is, with only one hidden layer. The input, hidden and output layers have the same number of neurons. The input data samples undergo a masking operation, *i.e.*, some bands of such vectors are forced to 0, what is also called *masking noise* [12]. For each sample, its reconstructed vector is estimated by the autoencoder, and the error between this reconstruction and the original vector without masking is attributed to the vector's positions turned into 0. After thousands of iterations, it is possible to rank the bands, and, the bigger the reconstruction error, the better the ranking. At the end, the best-ranked bands are selected. A multiplicative aggregation function (MAF), which takes into account the correlation amongst bands, is placed in the hidden layer of the AE architecture. Consequently, redundant information can be reduced.

The contributions of this paper can be summarized as follows: *i)* To the best of our knowledge, it is the first time that simple autoencoders are used for hyperspectral bands selection; and *ii)* Following [1], we propose modifications on the multiplicative aggregation function. Besides, we use another activation function in the autoencoder.

The rest of the paper is organized as follows: In Section 2, a literature review is presented. Section 3 provides a detailed description of the proposed framework. The dataset, classifiers used, the competitors and the results are found in Section 4. Finally, Section 5 concludes this paper.

2 Literature Review

Since the proposed method is unsupervised, we will cite in this section only recent state-of-the-art unsupervised works. In the literature, one can find a plethora of approaches addressing the BS subject under several perspectives and mathematical tools.

Evolutionary computation with optimization have been largely used by BS methods. For example, in [13], the authors propose an incorporated rank-based multi-objective band selection framework, to avoid conflicting objective functions, such as Jeffreys-Matusita (JF) and Bhattacharyya distances. During the processing, the spectral bands are transformed into binary vectors, whose elements are subjected to flipping with a certain probability.

Due to the high HSI dimensionality, the different classes existing in the image may lie in manifolds embedded in subspaces of the original feature space. Furthermore, it is also possible to explore the sparsity of the dataset in order to find a more meaningful data representation. For example, in [17], the authors propose a BS framework that can capture the inter-band redundancy through low-rank modeling. Then, by using an affinity matrix and concepts of data quality, the most representative bands are selected.

Another criterion that can be used in BS strategies is the HSI data information analysis. For example, in [10], the authors propose a method based on information-assisted density peak index. It takes into account the intra-band information entropy into the local density and inter-cluster distance to ensure cluster centers with a high quality. Besides, the channel proximity and band distance are integrated to control the local density compactness. The bands with top-ranked scores may get clear global distinction, good local density and also high informative quality.

Using graph theory, in [16] the authors propose a Multi-graph Determinant Point Process (MDPP). The aim is to capture the structure amongst bands and find the optimal band subset. For this, multiple graphs are designed to capture the intrinsic relationship amongst bands. Besides, the proposed MDPP is used to model the multiple dependencies in graphs, providing an efficient search strategy for the BS process.

Clustering techniques can also be used in band selection methods. For instance, in [15], the authors propose a framework based on dual clustering that takes into account the contextual information. For this, a novel descriptor that reveals the image context is devised, in order to select the representatives of each cluster, and taking into consideration the mutual effects of each cluster.

In sum, one can notice that there are many different techniques used in BS frameworks. Since the band selection we want to perform is also unsupervised, resorting to a intrinsically unsupervised method, such as autoencoder, is a plausible option.

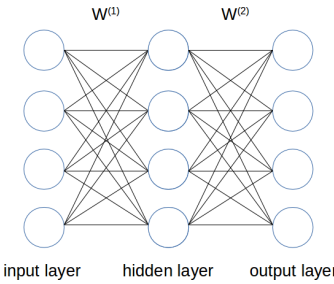


Figure 1: Example in reduced size of the autoencoder used in the proposed framework. All the layers have d neurons. $W^{(1)}$ and $W^{(2)}$ the sets of weights.

3 Proposed method

In this section, the proposed unsupervised band selection approach is described.

3.1 Definitions

Let X be the hyperspectral dataset whose elements $x^{(i)} \in \mathbb{R}^{1\times d}$ represent the pixels of this image, with $i = 1, 2, ..., n$, where n is the cardinality of X and d is quantity of spectral bands.

Let $N : \mathbb{R}^{1\times d} \rightarrow \mathbb{R}^{1\times d}$ be an autoencoder whose hidden layer has d neurons. Its cost function is quadratic. And let $W^{(1)}$ and $W^{(2)}$ be the matrices of weights between the input and hidden layers and between the hidden and output layers, respectively, as shown in Fig. 1. And let $B^{(1)}$ and $B^{(2)}$ be the vectors of biases of the hidden and output layers, respectively. Let $a_s \in A$ be the output of the s^{th} neuron of the hidden layer, with $s = 1, 2, \ldots, d$, and $A \in \mathbb{R}^{1\times d}$. Let S be the set containing the original d spectral bands. And let R be a set, whose element r_h is the ranking of the band $s_h \in S$, with $h = 1, 2, \ldots, d$.

Let $y^{(i)} \in \mathbb{R}^{1\times d}$ be the output of the autoencoder N.

Finally, let σ be the sigmoid function,

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

which will be used as activation function of the autoencoder. Since every $x_k^{(i)} \in x^{(i)}$ belongs to the $[0, 1]$ interval, we adopted the sigmoid as the activation function.

3.2 Description

Autoencoders seek to reconstruct in their output layer, the encoded information of the input vector.

Therefore, autoencoders are composed of two parts: *encoder* and *decoder*. The encoding of the input vector takes place in the hidden layer. And, then, the output layer performs the decoding process.

Mathematically, the output $y^{(i)} = N(x^{(i)})$ is defined as

$$y^{(i)} = \sigma(W^{(2)}A + B^{(2)}), \tag{2}$$

where the vector A is given by

$$A = \sigma(M + B^{(1)}), \tag{3}$$

where M is a vector containing d multiplicative aggregation functions $m_k(\cdot)$.

3.2.1 Multiplicative aggregation function

The multiplicative aggregation function (MAF) is an important component of the proposed framework, and this function is used to soften the redundancy present in the dataset. MAFs are placed in the hidden layer, and in order to exploit the correlation amongst all bands, the input and hidden layers have the same size. Thus, less redundant information will be fed towards the output layer.

In this paper, we propose a MAF simpler than the one proposed in [1], in order to speed up the processing —20% faster—, and it also yields simpler equations for the back-propagation algorithm.

For each neuron of the hidden layer, there is an associated multiplicative aggregation function $m_k : \mathbb{R}^{1\times d} \to \mathbb{R}$, with $k \in \{1, 2, \ldots, d\}$, given by

$$m_k(x^{(i)}) = x_k^{(i)}(w_{kk}^{(1)})^2(1 + \sum_{l\neq k} -2\rho_{lk}^2 w_{kl}^{(1)} x_l^{(i)}), \tag{4}$$

where ρ_{lk} is the correlation between the bands l and k, and the weights $w^{(1)} \in W^{(1)}$.

Finally, the output $a_s \in A$ of each hidden neuron is

$$a_s = \sigma(m_s(x^{(i)}) + b_s^{(1)}), \tag{5}$$

where $b_s^{(1)} \in B^{(1)}$.

According to Equation (4), the negative summation makes m_k smaller. More precisely, the bigger the correlation amongst band k and the other bands, the smaller the value of m_k, and consequently, the smaller the magnitude of a_k.

3.2.2 Spectral bands ranking

The outcome of the proposed band selection framework is the ranking of all spectral bands. At the end of the whole processing, for each band $s_h \in S$ there will be a correspondent $r_h \in R$ indicating its ranking.

During the training of the autoencoder N, every input data sample $x^{(i)}$ is subjected to the masking noise transform t, which has the following properties: *i)* each $x_k^{(i)} \in x^{(i)}$ has equal probability p to get masked; and *ii)* no band $x_k^{(i)}$ is masked in two consecutive iterations.

Thus, let $\tilde{y}^{(i)}$ be the output of the autoencoder when $\tilde{x}^{(i)}$ is the input sample. That is, $\tilde{y}^{(i)} = N(\tilde{x}^{(i)})$, where $\tilde{x}^{(i)} = t(x^{(i)})$. Likewise, $y^{(i)} = N(x^{(i)})$, without masking the input sample.

Initially, $R^{(0)} = 0$, and at iteration q, the calculation of the rankings $r_h^{(q)} \in R^{(q)}$ is

$$r_h^{(q)} = \frac{1 + v_1}{1 + v_2} + r_h^{(q-1)}, \tag{6}$$

if $\tilde{x}_h^{(i)}$ is masked. Where

$$v_1 = \sum_{k=1}^{d} (\tilde{y}_k^{(i)} - x_k^{(i)})^2,$$

and

$$v_2 = \sum_{k=1}^{d} (y_k^{(i)} - x_k^{(i)})^2.$$

Let us suppose that $\tilde{x}_h^{(i)} = 0$, that is, the h-th position of the input vector is masked. The extent to which $v_1 > v_2$ indicates the importance of the feature —or band —h.

The parameters update of the autoencoder is done by the back-propagation algorithm, based on the quadratic error between the output with masked input and the input without masking noise. That is,

$$e = \frac{1}{2}(\tilde{y}^{(i)} - x^{(i)})^2. \tag{7}$$

Algorithm 1 shows the steps of the proposed BS method. The indices of the biggest values of R are those of the best bands to be selected.

Algorithm 1 Proposed method.

1: input : X
2: initialize: $R^{(0)} = 0$
3: **for** $q = 1$: MaxIterations **do**
4: $\quad y^{(i)} = N(x^{(i)})$
5: $\quad \tilde{y}^{(i)} = N(\tilde{x}^{(i)})$
6: $\quad$ Update $R^{(q)}$ using (6)
7: $\quad$ Update the weights and biases of N using the back-propagation algorithm, according to the error calculated in (7)
8: return: R

4 Results

In this section, the results of the proposed method are shown. Furthermore, they will be compared with other BS methods by analyzing the accuracy of two supervised classifiers —K-Nearest Neighbors (KNN) and Classification and Regression Trees (CART) —, which have as input the selected bands.

The image used in this work is the Indian Pines, which consists of 145×145 pixels and 224 spectral reflectance bands in the $0.4 - 2.5$ μm wavelength range. Regarding the ground truth, there are 16 classes, which are used only for classification comparison purposes.

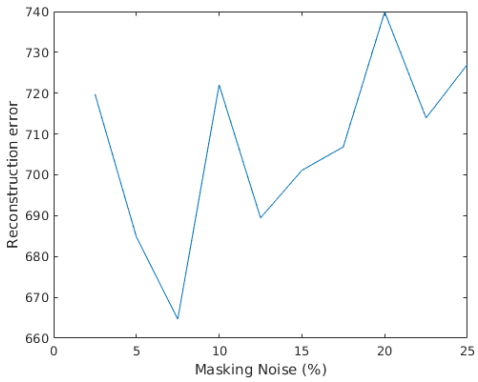


Figure 2: Reconstruction error with different masking noise probabilities. The lower the reconstruction error, the better the autoencoder.

4.1 Competitors

The classification performance of the proposed method is compared with four other methods from the literature.

One method is clustering-based [11], which will be referred to as `WaLuDi`. The other approach uses both clustering and ranking techniques for band selection [4], which will be called `CR`. Another competitor uses band elimination with partitioned image correlation [5], and this method will be called `EM`. And the last competitor is based on information divergence, and it will be referred to as `ID` [3].

Finally, the proposed method will be called `AE`.

4.2 Masking noise percentage

In [1], each feature $x_k^{(i)} \in x^{(i)}$ has a probability of $p = 0.25$ to be masked. However, in this work we run our algorithm with ten different probability values, with $p_v = v * 2.5\%$, where $v = 1, 2, \ldots, 10$. For each p_v, we summed up the reconstruction error for every input data sample, according to (7). In Fig. 2, one can see the reconstruction error for each masking noise probability. The masking noise probability that yielded the smallest reconstruction error was 7.5%, so this setting is kept throughout this work.

4.3 Selected bands

Firstly, the bands selected by the proposed method can be found in Table 1. We let them available to other researchers who may be interested in comparing results. Because we have access to only the first 18 best-ranked bands of our competitors, we restrict the analysis of results to this number of bands.

Table 1: Selected bands in order of importance, according to the rankings R.

Selected bands	106, 215, 43, 99, 123, 82, 118, 209, 144, 73, 13, 98, 11, 137, 133, 77, 174, 190.

4.4 Results comparison

All the classification results shown in this work are the mean values over ten runs. The standard-deviation values are also calculated.

In Table 2, one can see the results of the KNN classifier. The proposed method `AE` got the best results in almost all cases. It is illustrated in Fig. 3 (a).

Table 3 exhibits the overall results achieved by the CART classifier. The proposed method got the best result in only one case, with 12 spectral bands. In Fig. 3 (b), it is possible to have a visual idea of the results.

4.5 Remarks about the results

In general, KNN results are superior than CART accuracies, 73.36% and 62.65%, respectively. This is because CART divides the feature space into several regions corresponding each one to a class. Thus, once a $x^{(i)}$ lies in a region of the class α, for example, it will be given the label α, even if it belongs to class β. On the other hand, in such a situation, KNN would inquire the K nearest neighbors of $x^{(i)}$ to assign it a label. Therefore, KNN is better than CART in highly non-linear separating boundaries. Furthermore, from Tables 2 and 3 it is possible to notice that, in general, the accuracies increase as more bands are used. Fig. 4 depicts both facts.

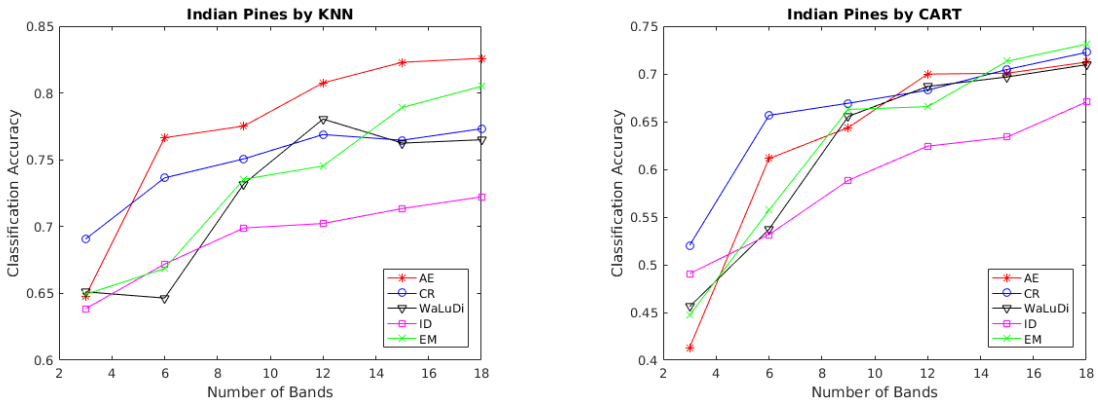
When it comes to the BS approaches, with the KNN classifier, the proposed method gets the best results in almost all situations. In fact, the standard-deviation values of Table 2 indicate that the `AE` method have statistically better results. Considering the CART classifier, the proposed method achieves the best accuracy with 12 bands. With 9, 15 and 18 bands, `AE` has results similar in relation to its competitors.

5 Conclusion

Hyperspectral images provide fine spectral details about the scene under analysis. However, the large amount of bands can also bring drawbacks in terms of storage and processing. Thus, in order to alleviate those problems, this work proposes a band selection method to decrease the HSI dimensionality.

The proposed BS method is based on autoencoders. During the training phase of the autoencoder, each input data sample is subjected to a masking noise transform, which flips some features of the input vector into zero, following a certain probability. Then, the output error is assigned to those indices with masking noise. The errors are summed up to their respective positions during the whole training phase. At the end, there is a ranking of the bands, and the most important are the ones with the biggest rankings.

According to the results, one can conclude that the KNN classifier is better than CART for the Indian Pines image. Also, the bigger the number of bands, the better the classifier accuracy. It is worth noting that the we selected from 3 up to 18 spectral bands. Regarding the proposed method, it achieved the best results in almost all situations using the KNN classifier. With CART, the proposed method got the best results in



(a) (b)

Figure 3: Indian Pines image classification results. In (a), KNN classifier. In (b), results achieved by the CART classifier.

Table 2: KNN results.

	3 bands		6 bands		9 bands		12 bands		15 bands		18 bands	
Method	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
AE	64.79%	0.24%	**76.65%**	0.37%	**77.53%**	0.34%	**80.76%**	0.53%	**82.30%**	0.50%	**82.59%**	0.37%
WaLuDi	65.12%	1.02%	64.65%	0.25%	73.19%	0.72%	78.05%	0.56%	76.25%	0.19%	76.50%	0.67%
CR	**69.06%**	0.52%	73.65%	1.03%	75.07%	1.43%	76.89%	1.07%	76.47%	1.14%	77.32%	0.22%
EM	64.92%	1.15%	66.86%	1.03%	73.54%	0.28%	74.54%	1.07%	78.92%	0.41%	80.50%	0.52%
ID	63.85%	079%	67.20%	0.22%	69.90%	0.18%	70.23%	1.16%	71.35%	0.47%	72.23%	1.34%

Table 3: CART results.

	3 bands		6 bands		9 bands		12 bands		15 bands		18 bands	
Method	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
AE	41.32%	0.55%	61.13%	0.84%	64.38%	1.17%	**69.98%**	1.48%	70.07%	1.06%	71.23%	0.80%
WaLuDi	45.62%	1.00%	53.71%	1.23%	65.55%	0.95%	68.68%	0.28%	69.68%	0.75%	70.96%	1.15%
CR	**52.03%**	1.14%	**65.66%**	0.39%	**66.93%**	0.37%	68.29%	1.48%	70.46%	0.99%	72.25%	1.91%
EM	44.72%	0.93%	55.72%	1.04%	66.28%	0.52%	66.57%	1.24%	**71.33%**	0.76%	**73.12%**	0.51%
ID	49.07%	0.82%	53.16%	1.35%	58.85%	1.42%	62.43%	1.67%	63.37%	1.01%	67.06%	0.87%

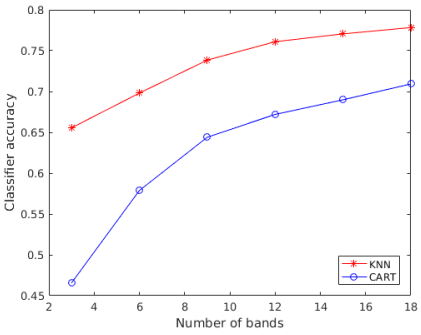


Figure 4: Mean results of all methods together.

one situation and similar to other competitors' results in other situations.

Concerning the future works, we will investigate some heuristics to choose the features to be masked, instead of using a uniform distribution.

Acknowledgements

This work was carried out in the framework of the Labex MS2T and DIVINA challenge team, which were funded by the French Government, through the program Investments for the Future managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

We are also thankful for the support provided by Brazilian Air Force and Institute for Advanced Studies (IEAv).

References

[1] B. Chandra and R. K. Sharma. Exploring autoencoders for unsupervised feature selection. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, July 2015.

[2] C. I. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, volume 1. Springer, 2003.

[3] Chein-I Chang and Su Wang. Constrained band selection for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 44(6):1575–1585, June 2006.

[4] A. Datta, S. Ghosh, and A. Ghosh. Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2814–2823, June 2015.

[5] Aloke Datta, Susmita Ghosh, and Ashish Ghosh. Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination. *International Journal of Remote Sensing*, 35(2):554–577, 2014.

[6] Simon S. Haykin. *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition, 2009.

[7] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968.

[8] S. Khalid, T. Khalil, and S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378, Aug 2014.

[9] J. Li and H. Liu. Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 32(2):9–15, Mar 2017.

[10] X. Luo, R. Xue, and J. Yin. Information-assisted density peak index for hyperspectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1870–1874, Oct 2017.

[11] A. Martnez-UsMartinez-Uso, F. Pla, J. M. Sotoca, and P. Garca-Sevilla. Clustering-based hyperspectral band selection using information measures. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):4158–4171, Dec 2007.

[12] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December 2010.

[13] X. Xu, Z. Shi, and B. Pan. A new unsupervised hyperspectral band selection method based on multiobjective optimization. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2112–2116, Nov 2017.

[14] H. Yang, Q. Du, H. Su, and Y. Sheng. An efficient method for supervised hyperspectral band selection. *IEEE Geoscience and Remote Sensing Letters*, 8(1):138–142, Jan 2011.

[15] Y. Yuan, J. Lin, and Q. Wang. Dual-clustering-based hyperspectral band selection by contextual analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1431–1445, March 2016.

[16] Y. Yuan, X. Zheng, and X. Lu. Discovering diverse subset for unsupervised hyperspectral band selection. *IEEE Transactions on Image Processing*, 26(1):51–64, Jan 2017.

[17] G. Zhu, Y. Huang, S. Li, J. Tang, and D. Liang. Hyperspectral band selection via rank minimization. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2320–2324, Dec 2017.