Hyperspectral Band Selection From Statistical Wavelet Models

Siwei Feng, Student Member, IEEE, Yuki Itoh, Student Member, IEEE, Mario Parente, Senior Member, IEEE, and Marco F. Duarte, Senior Member, IEEE

Abstract—High spectral resolution brings hyperspectral images with large amounts of information, which makes these images more useful in many applications than images obtained from traditional multispectral scanners with low spectral resolution. However, the high data dimensionality of hyperspectral images increases the burden on data computation, storage, and transmission; fortunately, the high redundancy in the spectral domain allows for significant dimensionality reduction. Band selection provides a simple dimensionality reduction scheme by discarding bands that are highly redundant, thereby preserving the structure of the data set. This paper proposes a new criterion for pointwise-ranking-based band selection that uses a nonhomogeneous hidden Markov chain (NHMC) model for redundant wavelet coefficients of each hyperspectral signature. The model provides a binary multiscale label that encodes semantic features that are useful to discriminate spectral types. A band ranking score considers the average correlation among the average NHMC labels for each band. We also test richer discrete-valued label vectors that provide a more finely grained quantization of spectral fluctuations. In addition, since band selection methods based on band ranking often ignore correlations in selected bands, we study the effect of redundancy elimination, applied on the selected features, on the performance of an example classification problem. Our experimental results also include an optional redundancy elimination step and test their effect on classification performance that is based on the selected bands. The experimental results also include a comparison with several relevant supervised band selection techniques.

Index Terms—Band selection, hidden Markov model, hyperspectral imaging, wavelet.

I. INTRODUCTION

HYPERSPECTRAL remote sensors collect reflected image data simultaneously in hundreds of narrow adjacent spectral bands, which makes it possible to derive a continuous spectrum curve for each image pixel. Compared with traditional multispectral techniques generating image cubes with low spectral resolution, hyperspectral remote sensors obtain a drastically increased number of spectral bands. Such an increase in data dimensionality provides the potential for better accuracy in discrimination among materials with similar spectral characteristics. However, the high dimensionality of hyperspectral data also has its drawbacks. With a fixed number of labeled training data, such a large dimensionality often causes the so-called Hughes phenomenon, i.e., a reduction in the generalization capability of classification system due to “overfitting” during the training procedure compared with lower dimensional data [1]. In fact, hyperspectral images tend to include significant redundancy in adjacent bands, leading to large amounts of redundant information being processed, stored, and transmitted [2], [3]. All these drawbacks of hyperspectral images present challenges to many conventional remote sensing data analysis problems. Thus, dimensionality reduction is often a necessary preprocessing step for hyperspectral data analysis.

One option for hyperspectral data dimensionality reduction is to transform the original data into a low-dimensional space while preserving a desired amount of information content. Such approaches are referred to as feature extraction, in which the generated features are generally combinations of the original reflectance values. The most popular feature extraction techniques include principal component analysis (PCA), which maximizes the total variance of the selected features, and Fisher’s linear discriminant analysis, which maximizes the class separability. A significant body of work exists on additional feature extraction techniques, such as orthogonal subspace projection [4], projection pursuit (PP) [5], and optimized information divergence PP [6]. However, the computation of features of new samples requires the entire spectrum to be available, which again poses a burden on computation, storage, and communication. Moreover, feature extraction changes the original data representation, which complicates the interpretation of the results of relevant hyperspectral data analysis. Particularly, feature extraction cannot be applied in cases where the physical meaning of individual bands needs to be maintained.

Another popular approach in hyperspectral dimensionality reduction is decision fusion, including approaches such as divide-and-conquer [1], [7]–[9]. In divide-and-conquer, one first groups bands adaptively into clusters and then a classifier is designed for each cluster independently. Finally, the classification decisions in each cluster are merged into a final global classification decision. In this way, the classification problem on the original high-dimensional data can be separated into several low-dimensional classification problems, which may reduce the negative effects brought by the large dimensionality. However, these methods can be hampered by the presence of redundancy between the selected adjacent spectral bands.
Furthermore, the common practice of feature extraction in each band group also affects the interpretability of each feature.

An alternative approach to dimensionality reduction is band selection [10]–[24]. As a complement to feature extraction, band selection aims to select a subset of the original bands, thus taking advantage of preserving the same feature space as that of the raw data, while avoiding the problem of high computational load as in feature extraction methods. For example, the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) spectrometer has a tiling mode in which only 72 selected channels are acquired [10]. The channels were carefully chosen by experts to capture a sufficiently large set of spectral species while maximizing the spatial field of view. CRISM has been applied in many scenarios. One example is the application in detecting the mineral species of the surface of Mars [25]. Apart from band selection based on expert knowledge, automatic band selection schemes greatly reduce the amount of manual work. In [26] and [27], adaptive methods are proposed to robustly discover symmetric absorption bands, which are selected to detect endmembers that consist of a given spectrum.

Though selecting bands with a specific physical meaning (e.g., absorption bands) may make the results more interpretable, these expert-knowledge-based methods may limit the exploiting of relationship between bands that cannot be precisely explained by the physical meaning. There exists a large volume of works in hyperspectral band selection area that employ more general statistical and signal processing methods to select bands without prespecifying types of bands needed. Band selection methods in this category can be roughly classified into two categories: groupwise selection methods [11]–[13], [15] and pointwise selection methods [16]–[22]. Groupwise selection methods aim at separating the entire set of spectral bands into several subsets, and final band sets are selected from these subsets. For example, [12] proposes a hierarchical clustering based method to separate the spectrum into clusters by maximizing the ratio between intercluster variance and intracluster variance. For each cluster, the band with the highest average correlation with other bands in the cluster is selected as the representative band; the final band subset consists of single representatives from each cluster. In another band clustering method [13], after all bands are grouped into \( k \) clusters through \( k \)-means clustering, one removes each cluster center in turn and evaluates the classification performance on the remaining \( k - 1 \) cluster centers. If the removal of a cluster center leads to optimal classification performance, then that cluster center will be removed and the remaining \( k - 1 \) clusters consist of the final output. When used for band selection, bands that are closest to each of the remaining \( k - 1 \) cluster centers can be used to form the final band set. Some groupwise band selection methods cluster bands according to their location (see [15]), while other methods cluster bands according to their statistical properties (see [11]–[13]). Additionally, band grouping is a common preprocessing step to methods involving decision fusion over several classifiers trained separately on each band group [1], [7]–[9].

In contrast, pointwise selection methods perform a gradual band selection procedure without relying on partitioning. Pointwise selection methods can also be separated into two groups. Subset search methods [16]–[18], [23] aim at optimizing some criterion via search strategies, sequentially increasing or decreasing the number of selected bands until the desired size is achieved. In contrast, band ranking methods [19]–[22] assign rankings to individual bands to measure their priority in a given task based on some criteria; then, bands with higher rankings are selected. Compared with subset search, the criterion used to judge whether a band should be selected depends only on the band itself. In subset search, the criterion often involves other factors, which makes the computation more complex.

In terms of whether class label information is used during band selection procedure, band selection approaches can be classified as supervised, unsupervised, or semisupervised. Supervised band selection methods [14]–[16], [18], [19], [21], [23] assume a priori knowledge of the class label information for different spectra during the selection process. In contrast, unsupervised band selection methods [11], [12], [17], [20], [24] do not assume any prior class information. Finally, several semisupervised band selection algorithms have recently been proposed [13], [22], leveraging both labeled and unlabeled training samples for band selection.

In this paper, we propose a supervised pointwise hyperspectral band selection scheme featuring a nonhomogeneous hidden Markov chain (NHMC) model that is trained and applied on the wavelet transforms of the training and testing spectra, respectively. The NHMC model provides discrete labels through the Viterbi algorithm [28]–[30] for each wavelet coefficient based on the coefficient statistics among the training samples. The labels are then collected into a discrete-valued feature matrix for each spectrum. The obtained features encode the scientifically meaningful structural information of each pixel in a hyperspectral image, which are referred to as semantic features.\(^1\) Instead of using the raw data to generate a band ranking, we use labels obtained by the NHMC model for each band and a variety of scales for each pixel. We use these labels, averaged over each class, to calculate pairwise class correlations for each band as a criterion for ranking-based band selection. To the best of our knowledge, neither wavelet analysis nor hidden Markov models have been fully exploited in the field of hyperspectral band selection in the past.

The designed NHMC-based feature used in this paper has also been employed on hyperspectral signature classification [31], data universality [32], and hyperspectral unmixing [33]. The first work [31] focuses on classifying pixels from a hyperspectral image. The second work [32] measures the degree to which the performance of individual supervised learning problems depends on the data set from which the feature design was based on (with the options being local

\(^1\)We define structural features in hyperspectral curves that capture the chemical makeup of corresponding material as semantic features, since they can be used for material discrimination.
versus global sets or, in other words, a data set specific to the supervised learning problem versus a universal data set. The third work [33] focuses on decomposing a spectrum into “endmembers” that contribute to the shape of the spectrum. The designed feature is used as the output feature in all the three works. The experimental results of the three works demonstrate the advantages of the designed features both in interpretability of the obtained features and in experimental performance, which further indicates the excellent discriminant performance of the designed features.

We also present a comparison with some other supervised and unsupervised hyperspectral band selection methods (covering both groupwise and pointwise band selection methods), decision fusion methods, and wavelet-based dimensionality reduction methods. The comparison involves several pixel-level classification problems with hyperspectral images, and we use the classification accuracy as a performance metric for band selection.

We summarize the main contributions of this paper as follows. First, we propose a supervised pointwise hyperspectral band selection scheme based on an undecimated wavelet transform (UWT) and NHMC. Most existing wavelet-based hyperspectral dimensionality reduction approaches rely on a discrete wavelet transform for preprocessing and simple dimensionality reduction via thresholding. In contrast, our band selection scheme use the UWT for preprocessing, thus preserving the band information in the wavelet coefficient domain, which provides convenience in the subsequent band selection operation. To the best of our knowledge, the use of UWT in band selection is not present in the literature. Second, we propose a criterion to measure the relevance of bands to a particular task for a band ranking method. Bands with high relevance are prioritized. Third, we conduct a numerical comparison with a wide range of hyperspectral dimensionality reduction methods (going beyond band selection) that provides an all-around contrast of the dimensionality reduction performance in terms of the classification accuracy on hyperspectral images. Fourth, we perform redundancy elimination on the selected bands to reduce the mutual dependence between selected bands. We compare the band selection performance measured by classification accuracy on hyperspectral image pixels before and after redundancy elimination to see the influence of redundancy elimination on band selection methods.

This paper is organized as follows. Section II reviews some related work and introduces the mathematical background behind our proposed band selection scheme. Section III provides an overview of the proposed hyperspectral signature classification system. Section IV describes our experimental validation setup as well as the corresponding results. Finally, Section V provides some conclusions.

II. BACKGROUND AND RELATED WORK

In this section, we provide an overview of the NHMC models that will be used by our proposed method. Furthermore, we review existing approaches to feature selection that will be used in Section IV.

A. Wavelet Analysis

The wavelet transform transform is a popular tool in many signal processing applications. In the field of hyperspectral data processing, the wavelet transform is widely used as a preprocessing step [34]–[36]. The wavelet transform of a signal provides a multiscale analysis of a signal’s content that effectively encodes the locations and scales at which the signal structure is present in a compact fashion. In this paper, we use the UWT to obtain multiscale analysis. We choose the UWT because it provides maximum flexibility on the choice of scales and offsets used in the multiscale analysis and allows for a simple characterization of the spectrum structure at each individual spectral band. Our analysis uses the Haar wavelet, which is more sensitive to a larger range of fluctuations than other wavelets. Thus, the Haar wavelet enables the detection of both slow-varying fluctuations and sudden changes in a signal [37], while it is not particularly sensitive to small discontinuities (i.e., noise) on a signal, in effect averaging them out over the wavelet support.

A 1-D real-valued UWT of an \( N \)-sample signal \( x \in \mathbb{R}^N \) is composed of wavelet coefficients \( w_{s,n} \), each labeled by a scale \( s \in 1, \ldots, L \) and an offset \( n \in 1, \ldots, N \), where \( L \leq N \). The coefficients are defined using inner products as \( w_{s,n} = \langle x, \phi_{s,n} \rangle \), where \( \phi_{s,n} \in \mathbb{R}^N \) denotes a sampled version of the mother wavelet function \( \phi \) dilated to scale \( l \) and translated to offset \( n \)

\[
\phi_{s,n}(\lambda) = \frac{1}{\sqrt{s}} \phi \left( \frac{\lambda - n}{s} \right).
\]

All the coefficients can be organized into a 2-D matrix \( W \) of size \( L \times N \), where rows represent scales and columns represent offsets. In this case, each coefficient \( w_{s,n} \), where \( s < L \), has a child coefficient \( w_{s+1,n} \) at scale \( s + 1 \). Similarly, each coefficient \( w_{s,n} \), at scale \( s > 1 \) has one parent \( w_{s-1,n} \) at scale \( s - 1 \). Such a structure in the wavelet coefficients enables the representation of fluctuations in a spectral signature by chains of large coefficients appearing within the columns of the wavelet coefficient matrix \( W \). The second figure of Fig. 1 shows an example of the wavelet coefficient matrix of UWT with nine scales of a sample hyperspectral signature (shown in the top figure of Fig. 1).

B. Statistical Modeling of Wavelet Coefficients

The statistical model is motivated by the compression property of wavelet coefficients, which states that the wavelet transform of a piecewise smooth signal generally features a small number of large coefficients and a large number of small coefficients, where “small” and “large” describe the relative magnitudes of wavelet coefficients. This property motivates the use of a zero-mean Gaussian mixture model (GMM) to encode wavelet coefficients into discrete features. Among the numerous variations of GMMs, the simplest one consists of only two Gaussian components and best captures the compression property of wavelet coefficients. In such a binary-state GMM, the Gaussian component with smaller variance characterizes wavelet coefficients with “small” magnitudes, while the Gaussian component with larger variance characterizes
wavelet coefficients with “large” magnitudes. By combining a binary GMM with suitable hidden state discovery schemes (e.g., the Viterbi algorithm, described in Section II-C), a binary-state GMM can assign to each wavelet coefficient one of two possible states: “large” or “small,” which can then be used as a discrete 2-D feature for the spectrum; we call this feature the state label in the sequel. The third figure of Fig. 1 shows the matrix of binary state labels for the hyperspectral signature shown in the top figure, where 0/1 corresponds to the “small/large” states, respectively. Fig. 1 shows that a wavelet coefficient is more likely to be characterized as “small” than “large,” which is a good endorsement of the compression property of wavelet coefficients.

However, apart from their simplicity (i.e., computational complexity) and embodiment of the compression property, binary-state GMMs have a potential drawback due to their overly coarse characterization of wavelet coefficient magnitudes. As mentioned above, binary-state GMMs only assign wavelet coefficients with either a “large” state or a “small” state. For wavelet coefficients close to the boundary between “large” state and “small” state, the encoding scheme can separate two similar wavelet coefficients into two different groups. Such a coarse characterization may have negative influence on subsequent processing steps. Therefore, in addition to binary-state GMMs, we also employ GMMs with more than one Gaussian components to capture the compression property. We associate each wavelet coefficient \( w_{s,n} \) with an unobserved hidden state \( S_{s,n} \in \{0,1,...,k-1\} \), where the states have prior probabilities \( p(S_{s,n} = i) \) for \( i = 0,1,...,k-1 \). Here the state \( i = 0 \) represents smooth regions of the spectral signature, which can be regarded as “the smallest” state, while \( i = 1,...,k-1 \) represent states with increasingly stronger spectral signature fluctuations, which can be regarded as a path from “the second smallest” state to “the largest” state. Each state is characterized by a zero-mean Gaussian distribution for the wavelet coefficient with variance \( \sigma^2_{i,s,n} \). The value of \( S_{s,n} \) determines which of the \( k \) components of the mixture model is used to generate the probability distribution for the wavelet coefficient \( w_{s,n} \): 

\[ p(w_{s,n}|S_{s,n} = i) = \mathcal{N}(0, \sigma^2_{i,s,n}) \]

We can then infer that 

\[ p(w_{s,n}) = \sum_{i=0}^{k-1} p_i S_{s,n} p(w_{s,n}|S_{s,n} = i) \]

The bottom figure of Fig. 1 shows the state label matrix with six states \( (S_{s,n} \in \{0,1,...,5\}) \) of an example hyperspectral signature shown in the top figure.

In cases where a UWT is used, the persistence property of wavelet coefficients [38], [39], which implies a high probability of a chain of wavelet coefficients to be consistently small or large across adjacent scales, can be accurately modeled by an NHMC that links the states of wavelet coefficients in the same offset. Because of the overlap between wavelet functions at a fixed scale and neighboring offsets, adjacent coefficients may have correlations in relative magnitudes [40]. However, for computational reasons, in this paper we consider only the parent–child relationship of the wavelet coefficients in the same offset. More specifically, we train an NHMC separately on each of the \( N \) wavelengths sampled by the hyperspectral acquisition device. This means that the state \( S_{s,n} \) of a coefficient \( w_{s,n} \) is affected only by the state \( S_{s-1,n} \) of its parent (if it exists) and by the value of its coefficient \( w_{s,n} \). The Markov chain is completely determined by the likelihoods for the first state and the state transition matrices for the different parent–child label pairs \( (S_{s-1,n}, S_{s,n}) \) for \( s > 1 \). In analogy with the binary GMM case, we can also define a \( k \times k \) transition probability matrix

\[
A_{s,n} = \begin{pmatrix}
P_{0\rightarrow0,s,n} & P_{0 \rightarrow 1,s,n} & \cdots & P_{0 \rightarrow k-1,s,n} \\
P_{1\rightarrow0,s,n} & P_{1 \rightarrow 1,s,n} & \cdots & P_{1 \rightarrow k-1,s,n} \\
& & \cdots & \cdots \\
& & & \cdots
\end{pmatrix}
\]

with

\[ p_{i\rightarrow j,s,n} = p(S_{s,n} = j|S_{s-1,n} = i), \]

where \( i, j = 0,1,...,k-1 \). Note that the probabilities in the diagonal of \( A_{s,n} \) are expected to be larger than those in the off-diagonal elements due to the persistence property of wavelet transforms. Note also that all state probabilities \( p_{i,s,n} \) for \( s > 1 \) can be derived from the matrices \( \{A_{s,n}\}_{s=2}^{L} \) and the probabilities for the first scale \( \{p_{i,1,n}\}_{i=0}^{k-1} \).

The training process of the NHMC model is based on the expectation maximization algorithm. The set of NHMC parameters \( \theta_n \) of a certain spectral band \( n \) includes the probabilities for the first scale \( \{p_{i,1,n}\}_{i=0}^{k-1} \), the state transition matrices \( \{A_{s,n}\}_{s=2}^{L} \), and the Gaussian variances \( \{\sigma^2_{0,s,n}, \sigma^2_{1,s,n}, \cdots, \sigma^2_{k-1,s,n}\}_{s=1}^{L} \). Since we use a zero-mean GMM in this paper, the Gaussian means are not regarded as a parameter.

C. Label Computation

Given the model parameters \( \theta = \{\theta_n\}_{n=1}^{N} \), the state label values \( l_{s,n} \) (where \( s = 1, \ldots, L \) and \( n = 1, \ldots, N \)) for a given observation are obtained using a Viterbi algorithm [30] as follows. We use \( l_{s,n} = i \), where \( i = 0,1,...,k-1 \), to denote the state of wavelet coefficient \( w_{s,n} \). For a particular wavelet coefficient \( w_{s,n} \), a conditional probability vector is defined as

\[ p(w_{s,n}|S_{s,n} = i) = \frac{1}{\sqrt{2\pi\sigma^2_{i,s,n}}} \exp\left(-\frac{w_{s,n}^2}{2\sigma^2_{i,s,n}}\right) \]
with \( i = 0, 1, \ldots, k - 1 \). We define a variable \( \delta_{i,s,n} \) as the “best score” that ends in a particular state \( i \) at scale \( s \) from its previous state and another variable \( \psi_{i,s,n} \) as the most likely state at a particular scale \( s - 1 \) to have children \( s \) with state \( i \). For the first scale \((s = 1)\), the two variables are initialized as
\[
\psi_{1,1,n} = 0, \quad \delta_{1,1,n} = p_{1,1,n} \cdot p(w_{1,n}|S_{1,n} = i).
\]
For subsequent scales \((s = 2, \ldots, L)\), we have the following mathematical expressions for the two variables:
\[
\psi_{i,s,n} = \arg \max_j \delta_{j,s-1,n} \cdot p_j \cdot p(w_{i,n}|S_{i,n} = j), \\
\delta_{i,s,n} = \psi_{i,s,n} \cdot \delta_{i,s,n-1} \cdot p(w_{i,n}|S_{i,n} = i).
\]
The output of the algorithm is twofold. The first output is a set of estimated states for each wavelet coefficient \( w_{s,n} \). When \( s = L \), the estimated state is \( l_{s,n} = \arg \max_j \delta_{j,s,L,n} \), i.e., the state with maximum likelihood. When \( s = 1, \ldots, L - 1 \), the estimated state is \( l_{s,n} = \psi_{l_{s-1,n},s,n} \); this implies that the state estimation proceeds backward through the wavelet scales. The second output is the likelihood \( p(W|\Theta) \) of a wavelet coefficient matrix \( W \) under the model \( \Theta \) as a byproduct. We propose the use of the state label matrix \( S \) as a descriptive feature for the original hyperspectral signal \( x \). The feature captures the presence of fluctuations in the spectrum (often described as semantic information that allows for discrimination between different types of spectra) by describing the relative magnitudes of the wavelet coefficients (as “large” or “small”) as well as their connection with adjacent wavelet coefficients in terms of their statistics.

D. Survey of Feature/Band Selection Algorithms

We now review several methods for band selection present in the literature, which will be used for comparison purposes in Section IV.

In feature weighting (FW) [21], a PCA matrix is generated for the spectra in each of the classes and a weight for each band–class pair is obtained from the row of the class PCA matrix corresponding to the band. The weights for a given band and all classes are fused to obtain a score used in band ranking. These scores aim to capture the weight that a given band has in the PCA decomposition for the classes considered.

Mutual information (MI) measures the degree of dependence between two random variables [41] and has been an important step in many unsupervised band selection approaches [12], [24]. MI can also be used for supervised band selection: one seeks the bands that feature maximal MI with the corresponding class labels over the hyperspectral data set.

Minimum estimated abundance covariance (MEAC) [18] is a subset search-based band selection algorithm that iteratively selects the bands with maximal dissimilarity to those already chosen, using sequential forward (i.e., greedy) selection. More specifically, MEAC is initialized by selecting the band pair that has the largest dissimilarity in terms of maximum linear prediction error, as proposed in [17]. Following sequential forward (i.e., greedy) selection, eligible bands (i.e., bands minimizing a target function after being combined with previously selected bands) are selected sequentially until the desired band size is reached. The supervised band selection scheme proposed in this paper requires a training procedure. In order to conduct a fair comparison, our experiments in Section IV apply MEAC on a training set and then select bands from testing samples based on the band selection result in the training procedure.

Heuristics such as genetic algorithms (GAs) [15] and particle swarm optimization (PSO) [23] are nowadays popular in the field of hyperspectral band selection because of their global searching capability, and so are expected to provide better performance than greedy algorithms such as sequential forward selection. When using PSO to select bands such as in [23], possible band combinations (which are called particles) and a solution update (which is called velocity) are first initialized and then recursively updated until some criterion is satisfied. The procedure based on GA for band selection is similar to that based on PSO. In [15], bands are first put into different groups based on conditional MI. Then, the GA is employed to choose one representative in each group. The rationale behind this setup is to make the search space of the GA computationally feasible as well as to reduce the redundancy between selected the bands. In fact, both methods in [15] and [23] contain algorithms to adaptively determine the optimal number of bands. However, in order to make the comparison with our method feasible, we set the number of bands in both methods to be fixed.

Apart from the pointwise band selection methods mentioned above, there also exist several relevant groupwise band selection methods. For example, Ward’s linkage strategy using MI and Ward’s linkage strategy using divergence [12] put bands into different clusters based on MI and Kullback–Leibler divergence, respectively. Meanwhile, an agglomerative strategy combined with Ward’s linkage is employed to merge bands into a predefined number of clusters. One representative from each cluster is selected to form the final band set for classification.

III. PROPOSED FRAMEWORK

We provide an overview of the NHMC-based band selection procedure in Fig. 2. The system consists of two modules: an NHMC-based feature training module and a band selection module. The second module assigns rankings to each band and results in the final band subset selection. The NHMC parameter training stage uses a training library of spectra containing pixels randomly sampled from the raw hyperspectral image cube and runs them through the UWT. The wavelet representations are then used to train a single NHMC model, which is then used to compute state labels for each of the training spectra using a Viterbi algorithm. The feature for each class is then constructed via averaging of all state arrays for the samples in that class. After that, pairwise class average correlation is computed for each band and the average correlation value for each band is then used as the criterion for ranking-based band selection.

A. Criterion for Band Selection

After obtaining state label matrices for each training sample, we construct a classwise feature by calculating the
element-wise average value of the state label matrices among training spectra in a certain class. Assuming that $I_{s,n}^{c,j}$ denotes the state label of sample $j$ from class $c$ at the band $n$ and scale $s$, the feature of class $c$ at band $n$ and scale $s$ is denoted by

$$I_{s,n}^c = \frac{\sum_{j=1}^{N_c} I_{s,n}^{c,j}}{N_c}$$  \hspace{1cm} (5)$$

where $N_c$ denotes the number of training samples in class $c$. Then for each band $n$, the correlation coefficients of class $p$ and class $q$ can be calculated as

$$\rho_{p,q}^n = \frac{\sum_{s=1}^{S} \sum_{j=1}^{N_p} I_{s,n}^{p,j} \cdot I_{s,n}^{q,j}}{\sqrt{\sum_{s=1}^{S} (\sum_{j=1}^{N_p} I_{s,n}^{p,j})^2 \cdot \sum_{s=1}^{S} (\sum_{j=1}^{N_q} I_{s,n}^{q,j})^2}}$$  \hspace{1cm} (6)$$

The criterion for the ranking of a certain band $n$ is the average of the all pairwise correlation coefficient values for band $n$

$$J_n = \frac{2}{C(C-1)} \sum_{p=1}^{C-1} \sum_{q=p+1}^{C} \rho_{p,q}^n$$  \hspace{1cm} (7)$$

where $C$ is the number of classes. We then rank the bands in increasing order of correlation (i.e., the band with the lowest correlation is ranked first). Although in this paper we use normalized correlation to measure the similarity between two classes, other metrics (e.g., Bhattacharyya distance, Kullback-Leibler divergence, and Jeffries-Matusita distance) can also be employed; the use of different similarity metrics may lead to different band selection results.

**B. Evaluation of Redundancy in Selected Bands**

Huang and He [21] claim that band selection schemes based on band ranking consider only the priority of bands for a given task, while ignoring the possible redundancy between the selected bands. Therefore, they argue that a redundancy elimination operation should be performed on ranked bands. However, there is a tradeoff here: after redundancy elimination, the overall relevance of selected bands to the specific problem is inevitably weakened. More specifically, bands with low priority may be selected to replace some bands with higher priority that are highly correlated with other selected bands. This behavior could negatively affect the final classification performance. In Section IV, we will verify the influence of band redundancy elimination on the final classification performance.

Following [21], after bands are ranked according to their priorities, the correlation coefficients of each band (starting from the second band) with all bands before it are calculated. A band will be discarded once its correlation coefficient with any band before it is greater than a prespecified threshold $T$. Final band selection will be conducted among the remaining bands.

We describe in Algorithm 1 a minor variant of the redundancy elimination scheme proposed in [21] that is used in this paper. One important difference between Algorithm 1 and the original approach proposed in [21] is that instead of empirically setting a threshold $T$ to select bands with low redundancy with other bands, we predefine the number of bands $N_T$ that are retained after redundancy elimination. We do this in order to conduct a convenient comparison with the experimental results of band selection without redundancy elimination, since all band selection approaches involved in our experiment fix the number of selected bands. The inputs to this algorithm include the ranked band set $B = \{b_1, b_2, \ldots, b_N\}$ with higher ranked bands being listed first and the maximum correlation vector $V$ that is computed as follows. We first compute the pairwise normalized correlation $N \times N$ matrix $D$ of those ranked bands by calculating band correlation coefficients across the training data set. More specifically, $d_{i,j}$ corresponds to the normalized correlation value of bands $b_i$ and $b_j$. We denote the training data set by a matrix $X \in \mathbb{R}^{N \times N_T}$, where $N_T$ denotes the cardinality of the training set, and each
Algorithm 1 Band Redundancy Elimination (Modified From [21])

**Input:** Target number of bands $N_b$, Ranked band set $B \in \mathbb{R}^{N \times 1}$, Training dataset $X \in \mathbb{R}^{N \times N_T}$.

**Output:** Selected band set $S$.

1. Define $D \in \mathbb{R}^{N \times N}$ whose elements will be the normalized correlation for each band pair.
2. for $i = 2, i \leq N, i + +$ do
3. for $j = 1, j < i, j + +$ do
4. Calculate the normalized correlation $d_{i,j}$ between $b_i$ and $b_j$ using (8).
5. end for
6. end for
7. Define $V \in \mathbb{R}^{N \times 1}$ whose elements will be the maximum correlation coefficient values between each band and all bands ranked with higher priorities.
8. for $i = 2, i \leq N, i + +$ do
9. Calculate the maximum correlation coefficient value $v_i$ between each band $b_i$ and all bands with higher priority ranking using (9).
10. end for
11. Rank the elements of $V$ in ascending order, and select the first $N_b$ bands in that order to form the final selected band set $S$.
12. return $S$

The maximum correlation vector $V$ contains the maximum correlation coefficient values between each band $b_i$ and all bands ranked with higher priority than $b_i$; more specifically, we can write the $i^{th}$ entry of the vector $V$ as

$$v_i = \max_{1 \leq j < i} d_{i,j}, \quad i = 2, 3, \ldots, N.$$  \hfill (9)

We note that $v_i$ is undefined and is not used in the algorithm. That is because the band with the highest ranking is always preserved. After we get the maximum correlation vector $V$, we sort the values of all elements in $V$ and determine the final band set by selecting the $N_b$ bands with the lowest values in $V$.

### IV. Experiment and Result Analysis

This section presents the experimental results for the comparison between our proposed method and relevant techniques including both pointwise and groupwise band selection.

#### A. Data Set Descriptions

In order to test the classification performance of the proposed method and its competitors, three different hyperspectral images were used in the experiment.

1. **Botswana:** This image has a spatial resolution of 30 m with 242 bands covering the 400–2500 nm spectral range with spectral resolution of 10 nm. Uncalibrated and noisy bands were removed and 145 bands remained for data analysis. The image has $1476 \times 256$ pixels. For classification purposes, 14 classes were used in this experiment.

2. **Kennedy Space Center (KSC):** This image has a spatial resolution of 18 m with 224 bands covering the 400–2500 nm spectral range. Bands with low signal-to-noise ratio or corrupted by water absorption were removed and 176 bands remained for data analysis. The image has $512 \times 614$ pixels. For classification purposes, 13 classes were used in this experiment.

3. **Whole Indian Pines (WIP):** The 92AV3C is a well-known hyperspectral image acquired by AVIRIS with $145 \times 145$ pixels, 224 spectral bands, and 16 classes, which is a small portion of a larger image known as Indian Pines. In this experiment, we consider the whole Indian Pines image, which has $2166 \times 614$ pixels and 58 classes. However, performing classification on such a large database with a time-consuming classifier [support vector machine (SVM)] takes a significant amount of time. We reduced the number of pixels for our simulation by preserving only those classes containing at least 1000 pixels, and we randomly select 1000 pixels for each of these classes. Finally, we have removed the bands covering the region of water absorption with 200 bands remaining. For classification purposes, 39 classes were used in this experiment.

The average value of the spectra for each class from each processed image is shown in Fig. 3.

#### B. Experimental Setup

In order to increase the statistical significance of our experimental results, the final classification accuracy of each method corresponds to the average from five-fold cross-validation testing experiments. For each fold, data from each class were separated into a training set and a testing set into a split of 20% and 80%; we refer to this average as the overall classification rate. We choose SVM [42] with a radial basis function (RBF) kernel as the exclusive classifier for all testing experiments. Due to the large amount of computation in our experiments, particularly those performed on WIP, we decide to choose only one classifier for testing. Our choice of SVM with RBF kernel is due to its simplicity and good performance. We use LibSVM [43] to implement our SVM-based classification.

#### C. Classification Results

Fig. 4(a)–(c) illustrates our experimental results, corresponding to each of the images listed in Section IV-A, respectively. The classification accuracy using all bands (i.e., with no band selection) for each image is displayed as a pink dashed line in Fig. 4 (denoted by “lossless” and explained later in this section). These results show that both Botswana and KSC have similar complexities, while WIP provides the
most difficult classification problem. Fig. 3 shows the average spectra from each class of each hyperspectral image: spectra from different classes are quite similar in all the three images, but the number of classes in WIP is much larger than in the other two (39 versus 13 and 14). With such a small distinction between classes, tripling the number of classes potentially results in a clear and significant increase in the complexity of classification.

GMM with two states is the simplest GMM model for NHMC, therefore providing the lowest computational cost. However, such a simplicity also brings the potential drawback of an overly coarse characterization of spectral semantics, since there are only two labels “small” and “large” describing the magnitudes of corresponding wavelet coefficients. This may cause low discriminability of the designed features. Compared with binary-state GMM, GMMS with multiple Gaussian components feature high computational cost as well as potentially better discriminability because of their finer structural information characterization. In order to compare the performance of binary-state and multistate GMMs for this band selection task, we use in Fig. 4 pairs of connected circle markers for the NHMC results corresponding to the classification rates for $k = 2$ GMM states and to the highest classification rate among $2 \leq k \leq 8$ GMM states. The numbers above the plot denote the value of $k$ that achieves maximum performance. From Fig. 4, we can find that for most cases in Botswana and WIP, multistate GMM achieves better performance than binary-state GMM. However, the advantage of multistate GMM is usually less than 1%. That is, although a large number of GMM states capture more structural information in hyperspectral data, the resulting improvement in performance is quite limited. Though not shown explicitly in this paper, our observations show that in some cases, the classification rates using multistate GMM is even worse than those of binary-state GMM. There are two possible reasons. First, the magnitude of a wavelet coefficient is not the only factor of determining its GMM state. Besides its magnitude, the GMM state of a particular wavelet coefficient $w_{s,n}$ is also determined by the state label of its parent $S_{s-1,n}$ as well as the transition probability matrix element $A_{s,n}$, which is introduced in Section II-B. In this paper, we train the NHMC model and generate features (i.e., state labels) for wavelet coefficients in each band independently. We observe that features assigned to adjacent wavelet coefficients at neighboring bands often disagree with their relative magnitudes. That is, a wavelet coefficient with larger magnitude may have a smaller GMM state and vice versa. This is due to the fact that the label of a wavelet coefficient is determined by multiple factors, as mentioned above. This in turn may cause different maps between wavelet coefficient magnitudes and GMM states across scales and offsets ($s,n$). The variance often makes it difficult to assess the semantic information in the label matrix of a spectral signature. In practice, this variance may sometimes affect the interpretability of features obtained from GMM labels.
Furthermore, the likelihood of such variability in the value-to-state mappings could increase when more states are used.

Overall, we can see that for all the eight methods, the overall classification accuracy has a sharp increasing slope for a small number of selected bands. We also note that the results for the WIP image have gentler slope than those for its two counterparts, likely due to the increased number of classes featured in that problem. The slope flattens after a certain number of selected bands that vary between methods and images. This behavior is described in [12] in terms of transitory and flat zones in the spectrum. More specifically, the length of the transitory zone of a band selection scheme measures its “speed” of convergence to its achievable stable performance status. Additionally, the overall performance in the flat zone shows the upper limit of the performance of that scheme. However, [12] gives only a description and a qualitative comparison regarding these two aspects. In this paper, we consider a much crisper quantitative measure of performance for the band selection methods. First, we determine the smallest number of bands that decreases the classification performance metric by up to 1% of its original value (i.e., the smallest number of bands needed to achieve a classification accuracy whose value is 99% of that using all bands) because, according to our observations, the classification accuracy curves of different band selection schemes switch from the transitory zone to the flat zone at this point. We denote this smallest number of bands by the approximate performance point. For the flat zone, we measure the minimum number of bands for which the classification performance metric meets or exceeds the value obtained when all bands are used, which we term the lossless performance point. With these two criteria, we can quantitatively compare the multiple band selection schemes considered in this section. Intuitively, band selection schemes with small values for both the approximate and lossless performance points have fast convergence speed and high stable performance and can be regarded as good band selection schemes. This is in keeping with the goal of band selection in classification: to reduce computational and storage load caused by high data dimensionality while minimizing the effect of subsampling on the classification performance.

Table I shows the estimated values of both the approximate and lossless points for our proposed method (NHMC) and the competing methods mentioned in Section II-D. The second column of Table I lists the approximate performance points and the third column lists the lossless performance points, respectively. Because of the high computational burden of experiments on this image, our search uses a step size of five bands for both Botswana and KSC and ten bands for WIP. The number in parentheses on the right of each point is the ratio between the actual classification accuracy at the performance point and the classification accuracy without band selection. For example, in the experiment on Botswana using NHMC, the smallest number of bands that achieves 99% of the classification accuracy without band selection is 30, where the classification rate is 99.02% of that achieved without band selection. Table I shows that except for KSC, our proposed method uses the smallest number of bands to achieve both approximate and lossless performances.

For KSC, our proposed method and FW use the fewest bands to reach approximate and lossless performance, respectively. The discrete nature of the designed features arguably enhances the discriminability between spectra with different structural features.

Note that for some approaches, the smallest number of bands corresponding to the lossless performance is actually equal to the total number of bands (145 for Botswana, 176 for KSC, and 200 for WIP). Particularly for WIP, only our method and FW can achieve performance better than using all spectral bands by performing band selection. Furthermore, in our experiment, FW actually uses 190 out of 200 bands to achieve that goal, which is much larger than the number of bands (140) for our method. However, we do not focus on these improvements in the classification of band selection methods for two reasons. First, in most cases, the number of bands needed for improved classification performance is too large to enable the computational load reduction that motivates band selection. Second, in all three tested images, none of the tested methods achieved a classification rate greater than 1% above that using all bands. This means that such advantages are negligible.

We also compare the classification performance of our proposed band selection method with those of related techniques that go beyond band selection. The first, based on decision fusion and proposed in [7], starts by clustering the original bands. After that, a classification decision is made on each band cluster and the corresponding classification accuracy is calculated. In our implementation, we use SVM combined...
with RBF as the classifier. Finally, a weighted majority voting-based decision fusion is conducted to get the final classification decision, where the weight for each band group is dependent on the corresponding classification accuracy. Note that in the original method, band groups with classification rates below 50% are not considered in the decision fusion process, i.e., they receive a zero-valued weight. However, since in some cases (especially for WIP) the classification rates of almost all band groups are less than 50%, we remove this requirement in our implementation for feasibility. The final classification rates for all three images are shown in the second column of Table II. We can find that the corresponding classification performance is similar to that of our proposed method when five bands are used for both Botswana and KSC and ten bands are used for WIP, while noting that all bands are used in this decision fusion-based approach.

The second approach proposed in [34] relies on a discrete wavelet transform of the spectra. Though the choice of mother wavelet function can be flexible, we choose the Haar wavelet in our implementation to match the choice made in our NHMC implementations. The receiver operating characteristic (ROC) curves are computed to perform feature selection. First, we calculate the ROC curve when each wavelet coefficients from each band is used for detection. Then, the wavelet coefficient with the largest ROC curve area (denoted by $A_{Z1}$) is selected as the initial feature. The selection proceeds greedily by adding the remaining wavelet coefficient that when selected produces the largest increase in the ROC curve area for the selected features; the search ends when no increases can be obtained by further adding a feature. While the description of [34] does not specify how to calculate the ROC curve area for a multiclass problem, in our implementation we calculate the average of the ROC curves of the individual pairwise classification problems as the area for the aggregated feature. The classification performance on the three hyperspectral images are listed in the third column of Table II. The classification performance is slightly worse than that of the divide-and-conquer method [7], which is shown in the second column of Table II.

### D. Impact of Redundancy Elimination

Next, we test the effect of redundancy elimination (see Section III-B) on band selection performance. Our experiment compares the classification performance for the bands selected with and without redundancy elimination for each band selection scheme. We let the retained number of bands $N_b$ vary between 10 and 100 with a step size of 10; the classification performance results are displayed in Fig. 5. Fig. 5 shows that redundancy elimination improves classification performance only in a limited number of cases, and almost all instances are for MI. To illustrate the diversity of features selected with and without redundancy elimination, Fig. 6(a) and (b) shows the distributions of the selected bands before and after redundancy elimination for Botswana with $N_b = 30$. For MI, it is apparent that redundancy elimination greatly improves the diversity in band distribution. Because of the high correlation between nearby bands in hyperspectral images, selecting bands from a wider, more diverse distribution makes the selected bands more representative of the overall structure of a hyperspectral image (and therefore more informative). From Fig. 5(b), we see that redundancy elimination improves classification rates in almost every case for MI. However, we also note that redundancy elimination does not bring significant improvement in classification performance for NHMC and FW. In the first and third columns of Fig. 6, we can find that the selected bands before redundancy elimination are already quite diverse and redundancy elimination does not seem to improve this situation further. In Fig. 5(a), we can find that for almost all cases of NHMC and for some case of FW, redundancy elimination degrades the classification performance. This is because redundancy elimination includes more bands with low priority in the selected band set, which may have a negative effect on the classification accuracy. For band selection approaches generating selected bands with high correlation between nearby bands such as the case of MI for Botswana, redundancy elimination is worth trying since the benefit of reducing band redundancy may exceed the negative
effects brought by sacrificing band priority. However, for methods generating selected bands that are already spread enough like NHMC and FW for Botswana, redundancy elimination may not be necessary step since the advantages from reducing band redundancy may not be larger than the damage caused by lowering the band priority.

V. CONCLUSION

We propose a supervised band selection framework that reduces redundancy in hyperspectral image bands while preserving useful semantic information. The proposed scheme uses an NHMC model in conjunction with a UWT to design features capturing the semantic information in the structure of each single-pixel hyperspectral datum. A band selection criterion is constructed based on classwise features, aiming to minimize the correlation between those features. The obtained experimental results demonstrate the advantages of our method over other relevant techniques. In addition, we also tested the effect of increasing the number of GMM states and redundancy elimination. The results tell us the sufficiency of a simple GMM and the necessity of a high band diversity. In the future, we will focus on the fusion of band selection and spatial information in hyperspectral classification problems, and the extension to unsupervised band selection will also be considered. Moreover, though the choice of Haar wavelet makes the obtained wavelet coefficients less sensitive to small discontinuities including noise than other wavelets, due to observation the discretization procedure from wavelet coefficients to state labels may amplify the existence of noise. And it is likely that the performance of our NHMC-based band selection method is partially damaged due to noise. Therefore, in the future, a preliminary denoising step will be conducted on wavelet coefficients in order to obtain robustness to noise.

ACKNOWLEDGMENT

The authors would like to thank P. Fung for providing an efficient implementation of our NHMC training code that greatly decreased the time of running experiments.

REFERENCES


**Siwei Feng** (S’14) received the B.S. degree in electronic and information engineering (with Hons.) from Soochow University, Suzhou, China, in 2012, and the M.S. degree in electrical and computer engineering from the University of Massachusetts, Amherst, MA, USA, in 2015, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

His current research interests include high-dimensional signal processing, machine learning, and computer vision.

**Yuki Itoh** (S’14) received the B.S. (with Hons.) and M.S. degrees in aeronautics and astronautics from the University of Tokyo, Tokyo, Japan, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Massachusetts, Amherst, MA, USA.

His current research interests include hyperspectral remote sensing, machine learning, and compressed sensing.

Dr. Itoh was a recipient of the scholarship, the Student Exchange Support Program (Long Term Study Abroad) during 2013–2016 from the Japanese Student Service Organization.
Mario Parente (M’05–SM’13) received the B.S. and M.S. (summa cum laude) degrees in telecommunication engineering from the University of Naples Federico II, Naples, Italy, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA.

He was a Post-Doctoral Associate with the Department of Geosciences, Brown University, Providence, RI, USA. He is a Supporting Researcher of several NASA-funded mission teams, including the NASA Compact Reconnaissance Imaging Spectrometer for Mars, the Moon Mineralogy Mapper, and the Biologic Analog Science Associated with Lava Terrains. He is also a Consultant for the Johns Hopkins Applied Physics Laboratory, Laurel, MD, USA, and the MIT Lincoln Laboratory, Lexington, MA, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA. His research involves combining physical models and statistical techniques to address issues in remote sensing of Earth and planetary surfaces and also involves the study of physical models of light scattering for remote sensing applications. His current research interests include developing machine learning and statistical signal processing techniques for spectroscopic analysis and imaging spectrometer data modeling, reduction, and calibration for terrestrial and planetary remote sensing, and also include the integration of sensing and decision-making for data prioritization and identification of scientifically attractive targets in manned and unmanned, rover, and orbiter-based missions.

Dr. Parente is an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Marco F. Duarte (S’99–M’09–SM’14) received the B.Sc. degree in computer engineering (with distinction) and the M.Sc. degree in electrical engineering from the University of Wisconsin-Madison, Madison, WI, USA, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from Rice University, Houston, TX, USA, in 2009.

He was an NSF/IPAM Mathematical Sciences Post-Doctoral Research Fellow in the Program of Applied and Computational Mathematics with Princeton University, Princeton, NJ, USA, from 2009 to 2010, and with the Department of Computer Science, Duke University, Durham, NC, USA, from 2010 to 2011. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA. His research interests include machine learning, compressed sensing, sensor networks, and computational imaging.

Dr. Duarte is a member of Tau Beta Pi. He received the Presidential Fellowship and the Texas Instruments Distinguished Fellowship in 2004 and the Hershel M. Rich Invention Award in 2007, all from Rice University. He also received the IEEE Signal Processing Society Overview Paper Award (with Y. Eldar) in 2017. He is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.