Endmember Extraction Using a Combination of Orthogonal Projection and Genetic Algorithm

Y. Rezaei, M. R. Mobasheri, M. J. Valaddan Zoej, and M. E. Schaepman, Senior Member, IEEE

Abstract—Common endmember extraction algorithms presume that the number of materials present is either known or may be predetermined by using spectral databases or other approaches. In this letter, we propose a new method called genetic orthogonal projection (GOP) for endmember extraction in imaging spectrometry. GOP is based on a fully unsupervised approach and uses convex geometric characteristics as well as a genetic algorithm. We compare GOP with existing endmember extraction algorithms and demonstrate that GOP partially outperforms them, without the need of a priori information.

I. INTRODUCTION

Imaging spectroscopy, with its capability of capturing hundreds of contiguous spectral bands, has introduced new perspectives for data (pre-)processing and applications alike [1]. Imaging spectroscopy provides data at high spectral resolution that can be used for object detection and for discriminating between these objects based on their spectral characteristics. However, as in all remotely sensed images, data are subject to significant mixing at the pixel level. The pixel-level measurement is a function of individual spectral signatures contributing to the total signal.

Therefore, decomposing the mixed pixel signals into their contributing materials (“endmembers”) has emerged as one of the most challenging problems in imaging spectrometer data processing [2], [3]. In addition, dimensionality algorithms are used to estimate the number of endmembers present. Projection techniques seek the best subspace to project data by minimizing an objective function. The principal component analysis [4] and singular value decomposition (SVD) [5] are the common projection-based algorithms. Harsanyi et al. [6] developed a Neyman–Pearson detection theory-based thresholding method (Harsanyi–Farrand–Chang (HFC)) or virtual dimensionality (VD) [3] to determine the number of spectral endmembers in hyperspectral data. The HFC method is based on a detector built on the eigenvalues of the sample correlation and covariance matrices. The dimensionality algorithms usually are threshold based and need a threshold as a criterion. Another method to estimate the intrinsic dimensionality is the hyperspectral signal identification by minimum error (Hysime) [7] method. Hysime is a fully unsupervised method. It first estimates the signal and noise correlation matrices and then selects the subset of eigenvalues that best represents the signal subspace in the least squared error sense [7]. Recently, several fully unsupervised methods have been developed for endmember extraction, such as the works described in [8] and [9], that estimate the number of endmembers and extract them.

As it is mentioned in [10], two issues are arising in endmember extraction algorithms: One is the knowledge of the number of endmembers that must be provided a priori, and the other is the initialization of endmember extraction algorithms, where most of these algorithms use the dimension reduction algorithms and initialize their endmember-searching processes by using randomly generated endmembers. Since endmember extraction algorithms are sensitive to initial endmembers, a properly selected set of initial endmembers can make significant improvements on the searching process [10].

The proposed method, called genetic orthogonal projection (GOP), is based on geometrical characteristics and a genetic algorithm (GA) and is used to estimate the number of endmembers and their extraction. The subsequent extraction of the endmembers from an image is based on three main stages as follows: 1) approximate estimation of the initial number of image-based endmembers using absorption features present in the spectrum of each pixel; 2) extraction of the initial endmembers by projecting the data onto an orthogonal subspace; and 3) determination of the exact number and location of each endmember using a GA. Here, the dimension of the subspace is varied until the best estimation of the subspace is achieved.

II. METHODS AND MATERIALS

A. Approximate Estimation of the Number of Endmembers

In imaging spectroscopy, when a high degree of spectral detail is present, i.e., distinct absorption phenomena dominate the spectral space, the pixel response can be related to the physical properties of the absorbing molecules within the pixel area [11]. In general, the spectral shape has a direct relationship with the composition of the material present in the target pixel, and consequently, the variations in the surface materials change the spectral signatures.

In this letter, the absorption features are determined by using gradient changes in the reflectance of three adjacent spectral...
Assume that \( \mathbf{R} = (L, N) \) represents a matrix of an imaging spectrometer data set consisting of \( N \) pixels, each represented by a pixel vector \( r_i \) in \( L \) spectral bands. If a spectral pixel vector \( r_i \) belongs to an absorption feature, then \( r_i < r_{i-1} \) and \( r_{i+1} > r_i \), or \( r_i - r_{i-1} < 0 \) and \( r_{i+1} - r_i > 0 \), can be expressed as follows:

\[
r_i - r_{i-1} < \Delta \quad \text{and} \quad r_{i+1} - r_i > \Delta
\]

(1)

where the quantity \( \Delta \) is a threshold value equal to the mean of the difference between every two adjacent bands. \( \Delta \) can be used to remove the undesired effects and can be calculated by

\[
\Delta = \frac{1}{L-1} \sum_{i=2}^{L} |r_i - r_{i-1}|.
\]

(2)

In order to decrease the computational cost, the average spectra of all pixels in an image are calculated. All absorption features in this averaged spectrum can be extracted by using (1).

### B. Extraction of Initial Endmembers

To extract the initial endmembers, the improved Automatic Target Generation Process algorithm [3] was used. This algorithm projects data repeatedly on an orthogonal subspace and identifies the vertices of the simplex stepwise.

The inputs of the algorithm are the imaging spectrometer data sets and the initial number of endmembers \( (m) \) estimated from the data.

**Step 1** involves calculating the norm for each pixel vector and finding the pixel with the maximum norm. This pixel is labeled as the first endmember \( e_1 \).

In **Step 2**, the pixel with the largest distance to \( e_1 \) is searched and labeled as the second endmember \( e_2 \). The endmember matrix is then formed as \( U = [e_1, e_2] \).

Finally, in **Step 3**, the new endmembers are extracted as in the Vertex Component Analysis (VCA) method [13]. In this step, all other pixel vectors are projected onto the subspace orthogonal to the space spanned by \( U \), by using \( r_p = o^T R \), and the endmember matrix \( U \) is constructed using the endmembers previously defined. \( o \) is the orthonormal projection operator, and it is calculated by

\[
o = (I - UU^+w)/||(I - UU^+)w||,
\]

where \( I \) is the identity matrix, \( w \) is a zero-mean random Gaussian vector of covariance \( I \) [13] that controls the randomness of the process, and \( U^+ \) is the pseudoinverse of \( U \), \( U^+ = (U^T U)^{-1}U^T \). The projected data in the new space also have a convex shape [13]. This means that the endmembers are still occupying the vertices of the simplex. The pixel vector with the maximum norm in the projected data set \( (r_p) \) still represents an endmember; the index of these pixels is determined, and the \( m \)th endmember is extracted. During each iteration, the endmember matrix \( U \) is updated, and **Step 3** is repeated until all endmembers \( (m) \) are determined.

### C. Determining the Exact Number of Endmembers Using a GA

GAs are based on the genetic process of biological organisms and represent adaptive methods that may be used to solve search and optimization problems. They work with a population of individuals, each representing a possible solution to a given problem, and an individual is produced by assigning values to each gene [14]. Each of these individuals is assigned a fitness score according to the qualification of that solution for the given problem. Individuals are randomly selected from the population and recombined using the mechanism of crossover and mutation to produce offsprings (new solutions) comprising the next generation. Mutation randomly alters each gene with a small probability. It is responsible for reintroducing lost gene values inadvertently, providing a small amount of random search in the vicinity of the population [14].

In order to prepare a GA, the coding scheme, chromosome design, and a suitable fitness function must be determined.

**Coding scheme:** When solving an optimization problem using a GA, the first step is to model a typical solution as a set of ordered genes called chromosomes. In this letter, a chromosome is designed by having \( m \) genes where \( m \) represents the number of primary endmembers produced in the previous step. To obtain a possible solution to the endmember extraction problem, each gene is considered to be an endmember, and each individual in the population is considered to be a combination of endmembers. To generate a coding system, a string of binary digits in the chromosome is designed where a gene equals either “1,” when the endmember contributes to the calculation of a fitness function, or “0,” when the endmember is not involved in the calculation of the fitness function. The design of a chromosome is shown in Fig. 2.

**Fitness function:** Under the linear mixing model, the spectral data set from a scene resides within a simplex in a multidimensional space whose endmembers represent the vertices of the simplex [13], [15], and the volume defined by the endmembers is maximal [15]. Therefore, the objective is to find the minimum number of initial endmembers that provide the maximum volume among the set of all simplices.
with the same dimension. The volume of a simplex in an $n$-dimensional space with $n+1$ vertices $(v_0, \ldots, v_n)$ can be defined as

$$Vol. = \frac{1}{n!} |\det(v_1 - v_0, v_2 - v_0, \ldots, v_n - v_0)|. \quad (3)$$

Each column in the $n$ by $n$ determinant represents the subtraction of two vectors each representing two vertices. In (3), $Vol.$ is the volume of the simplex defined by $n$ endmembers, $\det(.)$ is the determinant of the matrix, $|.|$ denotes the absolute value, and $n$ is the number of endmembers contributing to the volume calculation. In other words, $n$ equals the number of genes in each chromosome that are assigned a value of “1” in the binary system as defined in the coding scheme.

In the first step, a randomly generated initial population, the mutation probability, the location of the primary endmembers that have already been extracted, and the projected data set on orthogonal subspace by using SVD [5] are introduced to the algorithm.

In the second step, the fitness function for each individual is computed. This procedure is performed by determining the number of genes that are equal to “1.” For calculating the volume of the simplex, we need to know the dimensions of the space and the positions of simplex vertices. The algorithm creates the simplex by changing the positions of the simplex vertices and the dimensions of the subspace. After determining the simplex, its volume is calculated by using (3), and it shows the fitness function for each individual.

The selection of pairs for the mating pool is performed during the third step. We used the roulette wheel selection method as defined in [16]. If $f_i$ is the fitness of an individual $i$ in the population, its probability of being selected is $r_{bi} = f_i / \sum_{j=1}^{N} f_j$, where $N$ is the number of individuals in the population. The next two steps are crossover and mutation. They select the position and number of endmembers from among the pixels that are the vertices of the simplex and dimensions of the subspace.

After crossover and mutation, the level of convergence of the algorithm is calculated for each iteration. When it equals 95%, i.e., 95% of all genes in each individual and 95% of all individuals in each population are the same, the algorithm stops. The locations of the final endmembers and their numbers correspond to the genes equaling “1” in the fittest individual.

III. EVALUATION OF GOP WITH SIMULATED DATA

Simulated data were used in order to evaluate the GOP method. For simulating spectral data sets, we used a linear mixing model based on the methods described in [7] and [13]. The endmembers were selected from the United States Geological Survey (USGS) digital library [17], and the abundance fractions were generated according to Dirichlet distribution principles. The Dirichlet distribution enforces positivity and full additive constraints resulting in a wide range of shapes depending upon the parameters of the distribution [7]. By adding simulated noise to these data, a Gaussian shape centered at the band center $L/2$ was obtained and used, as outlined in [7].

To evaluate the performance of endmember extraction, angular errors between estimated and real endmembers were calculated using the spectral angle mapper (SAM) [18]. Based on $SAM(s_i, s_j)$ criteria, for $p$ endmembers, the spectral angle error (SAE) rmse was estimated using $SAE_{rmse} = (\text{Sum}(SAM(s_i, s_j)^2/p))^{0.5}$.

A. Estimation of Number of Endmembers

The processing of simulated data was conducted using 3, 5, and 15 endmembers, respectively, with $L = 224$ spectral bands and $N = 10^4$ pixels. First, the GOP was evaluated with respect to the SNR for both white and nonwhite noise. In the second experiment, the Hysime [7], VD [3], and the proposed method were compared for their abilities in estimating the number of endmembers. Table I shows the results of the GOP, Hysime, and VD algorithms for different SNR and noise types. These results are based on ten Monte Carlo runs. The GOP first estimates the initial number of endmembers and then extracts the final endmembers. The initial endmembers depend on the type of reference material. Different numbers were estimated using the ten rounds of iteration (see Table I).

As can be seen in Table I, the GOP algorithm shows improved estimation for larger SNR values and white noise, i.e., GOP works well with white noise and for SNR = 35 and 25. The results of GOP are better than VD with a false alarm probability $P_f$ equal to $10^{-3}$. For smaller SNR values, the GOP usually produces overestimations.

In the next experiment, the GOP algorithm was compared with two other endmember extraction algorithms, namely, N-FINDR [15] and VCA [13]. The evaluation was performed using 3 to 15 endmembers and all 224 spectral bands of the Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS) convolved library spectra selected from the USGS spectral library. The abundance fractions were generated according to a Dirichlet distribution, and additive noise was added to the data. The results of this comparison are shown in Fig. 3. The GOP
Fig. 3. RMSE (SAE) of GOP, VCA, and N-FINDR algorithms using 3–15 endmembers, \( N = 10^4 \) pixels, and an SNR = 35 [dB], with a Gaussian-shaped noise.

Fig. 4. False color composite of the subimage (\( R = 855.87 \), \( G = 645.96 \), and \( B = 547.20 \) nm).

IV. EVALUATION OF THE GOP ALGORITHM USING AVIRIS DATA

We evaluated the GOP using actual AVIRIS data from the Cuprite test site. This study is based on a subimage as shown in Fig. 4. After removing noisy and water vapor affected bands [7], a 250 × 191 pixel image was selected to be used for the GOP evaluation. A total of 188 remaining bands were used (i.e., 389.53–1314.699 nm (bands 3–103), 1423.61–1752.15 nm (bands 114–147), and 1948.56–2467.1599 nm (bands 168–220)).

TABLE II

<table>
<thead>
<tr>
<th></th>
<th>GOP</th>
<th>VCA</th>
<th>N-FINDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buddingtonite</td>
<td>3.89</td>
<td>3.89</td>
<td>3.89</td>
</tr>
<tr>
<td>Montmorillonite</td>
<td>3.14</td>
<td>3.23</td>
<td>3.14</td>
</tr>
<tr>
<td>Nontronite</td>
<td>4.94</td>
<td>4.94</td>
<td>5.25</td>
</tr>
<tr>
<td>Quartz</td>
<td>5.87</td>
<td>4.92</td>
<td>6.33</td>
</tr>
<tr>
<td>Alunite+Dickite</td>
<td>3.89</td>
<td>5.41</td>
<td>3.85</td>
</tr>
<tr>
<td>Kaolinite #1</td>
<td>3.19</td>
<td>3.25</td>
<td>3.34</td>
</tr>
<tr>
<td>Kaolinite #2</td>
<td>3.04</td>
<td>3.02</td>
<td>3.05</td>
</tr>
<tr>
<td>Montmorillonite-Na</td>
<td>2.95</td>
<td>1.05</td>
<td>0.52</td>
</tr>
<tr>
<td>Opalized Tuff</td>
<td>2.11</td>
<td>2.77</td>
<td>Goethite (7.38)</td>
</tr>
<tr>
<td>Stonewl_Ply</td>
<td>4.21</td>
<td>6.49</td>
<td>3.65</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>3.86</td>
<td>4.17</td>
<td>4.43</td>
</tr>
</tbody>
</table>

The VCA-, N-FINDR-, and GOP-algorithm-extracted endmembers were compared to the USGS spectral library data convolved to AVIRIS [19] and also to the ground truth as presented in [20]. In order to estimate the number of endmembers to be introduced into the VCA and N-FINDR algorithms, the Hysime method is used. By applying Hysime on the AVIRIS subset (see Fig. 4), the number of endmembers was estimated to be \( p = 18 \). This value of \( p \) was introduced to the VCA and N-FINDR algorithms. We applied the GOP algorithm to this data set, and its first estimation was \( p = 18 \) while the final estimation was \( p = 10 \). The comparison of the results of GOP with those of VCA and N-FINDR is shown in Table II. The computational performance of all three algorithms is very similar, although GOP uses no a priori information.

The VCA and N-FINDR algorithms extracted 18 endmembers of which some endmembers were extracted multiple times (e.g., Montmorillonite #1, #2, #3; Montmorillonite-Na #1, #2; Alunite #1, #2; and Kaolinite #1, #2, #3). The rmse estimation is based on the same ten endmembers that all three algorithms detected. Duplicated endmembers extracted using VCA and N-FINDR were removed for the rmse analysis.

The VCA and N-FINDR algorithms cannot detect the number of endmembers automatically due to the a priori requirement. Even the introduction of varying number of endmembers to these algorithms will not change the result; they will still extract the same number of endmembers. While ambiguity with VCA and N-FINDR remains high, the ambiguous endmembers remain low with GOP.

V. CONCLUSION

In this letter, we have proposed a new endmember extraction algorithm named GOP. GOP is an unsupervised approach and is based on the geometry of convex sets and GAs. It exploits the fact that endmembers occupy the vertices of a simplex. The GOP algorithm assumes the presence of pure pixels in a given data set, similar to other approaches such as VCA, Pixel Purity Index, and N-FINDR. The results show that the proposed GOP algorithm can well estimate the number of endmembers and works well in comparison with other endmember extraction
REZAEI et al.: ENDMEMBER EXTRACTION USING A COMBINATION OF ORTHOGONAL PROJECTION AND GA

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

algorithms. The GOP algorithm shows the good results in high SNR in simulated data, and it has a better rmse in real data in comparison of the VCA and N-FINDR algorithms. However, in the future work, it is intended to focus on noise estimation and on improving the GOP performance for the low values of SNR. The main advantage of the GOP algorithm compared to other common endmember extraction algorithms is that there is no need to enter the number of the endmembers estimated to be present. The GOP estimates and extracts the endmembers in a completely unsupervised fashion.

ACKNOWLEDGMENT

The authors would like to thank their employing institutions for the support. The authors would also like to thank the anonymous reviewers for the helpful comments.

REFERENCES