METHOD FOR ESTIMATING THE NUMBER OF MIXED SPECTRAL ENDMEMBERS BASED ON FEATURE-ENHANCED SPATIAL SPECTRAL MATCHING ALGORITHM **

Qingbo Li*, Qi Wang, Shaolin Shi

Beihang University, School of Instrumentation and Optoelectronic Engineering,
Precision Opto-Mechatronics Technology Key Laboratory of Education Ministry, Haidian District;
e-mail: qbleebuaa@buaa.edu.cn

Based on the characteristics of space target components, this paper proposes a spectral in-degree distribution curve-matching algorithm. Simulation experiments showed the average accuracy of the proposed algorithm is improved in comparison with the Harsanyi–Farrand–Chang and HySime algorithms. The simulation experiments were performed for a case in which the number of mixed spectral bands was reasonably small. The results showed the number of endmembers could be estimated accurately when the number of endmembers was three to eight. The algorithm proposed in this paper is suitable for estimating the number of endmembers of a space target.

Keywords: *intrinsic dimension, in-degree value, K-nearest neighbor, spectral matching, endmember number estimation.*

МЕТОД ОЦЕНКИ ЧИСЛА СМЕШАННЫХ СПЕКТРАЛЬНЫХ КОНЕЧНЫХ ЭЛЕМЕНТОВ НА ОСНОВЕ РАСШИРЕННОГО АЛГОРИТМА ПРОСТРАНСТВЕННОГО СПЕКТРАЛЬНОГО СОГЛАСОВАНИЯ

Q. Li*, Q. Wang, Sh. Shi

УДК 535.3

Университет Бейхан, округ Хайдянь,
Пекин, 100191, Китай; e-mail: qbleebuaa@buaa.edu.cn

(Поступила 7 марта 2019)

Предлагается алгоритм согласования кривой степенного спектрального распределения, основанный на характеристиках компонентов космической цели. Средняя точность предложенного алгоритма выше, чем алгоритмов Харсаньи–Фарранда–Чанга и Хайсима. Эксперименты проводились для случая, когда число смешиваемых спектральных полос достаточно мало. Результаты показывают, что число конечных спектральных элементов может быть оценено точно в пределах от трех до восьми. Алгоритм пригоден для оценки числа конечных спектральных элементов космической цели.

Ключевые слова: *внутренняя размерность, степенная величина, алгоритм k-ближайших соседей, спектральное согласование, оценка числа конечных спектральных элементов.*

Introduction. In the aerospace field, the term "space target" usually refers to manmade spacecraft and space debris in orbit beyond earth's atmosphere. With ongoing development of the aerospace industry in various countries, the necessity for regulated monitoring, i.e., identification and tracking of space targets, has become increasingly urgent. The main objective of this study was to explore the application of imaging spectroscopy in the field of space target recognition. By analyzing the mixed spectrum of a space target, the number of different types of material contained in the target can be obtained, which lays the foundation for further analysis of specific materials and material composition ratios.

The composition of space target material can be referred to as an endmember, and estimation of the number of endmembers is a prerequisite for endmember extraction. The exact number of endmembers will

[**] Full text is published in JAS V. 87, No. 2 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 87, No. 2 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

have a positive impact on the spectral unmixing algorithm. Owing to limitations imposed by both the measuring instruments and the measurement environment, a space target becomes a point target on an imaging spectrometer. The hyperspectral image of each component of a space target corresponds to a single spectrum, and the number of spectra is small. A space target contains a fixed number of materials, and it is likely that most targets will be composed of several types of materials listed in the library of space target materials. With consideration of the characteristics of the abovementioned space target material distribution, this paper proposes a mixed spectral in-degree curve-matching method based on the spectral library of known materials to provide an accurate estimation of the number of endmembers for the case of a small number of spectral bands.

The number of endmembers in the algebraic domain can be understood as the minimum number of independent variables required to describe the mixed spectral space, which is also named the intrinsic dimension (ID) of the spectral data space. For a linear mixed model of mixed spectra, the number of endmembers in the hyperspectral data set will be related linearly to the ID of the data set. There are many definitions of ID [1]. In the context of the application of hyperspectral data, the interpretation of the ID assumes that the hyperspectral data set is located on a smooth manifold in the spectral space, and that the manifold can be modeled by a given spectral mixture (i.e., a linear or nonlinear model). By definition, a manifold is similar to a given size of Euclidean space, and the ID of a manifold is the dimension of the Euclidean space. We believe that the number of endmember spectra is equivalent to the dimension of the mixed spectral space, i.e., the ID of the spectral space. Therefore, we apply geometric thinking to the composition of the mixed spectra, and we use ID estimation techniques to determine the number of endmembers [2].

Currently, the endmember number estimation algorithms used most commonly are the virtual dimension (VD) algorithm and the HySime algorithm. The original VD algorithm was proposed by Chang et al. [3]. The most well-known algorithm based on the VD algorithm is the Harsanyi–Farrand–Chang (HFC) algorithm, the essence of which is to find the minimum number of distinct signals in a spectral matrix from the perspective of target monitoring. The HySime algorithm proposed by Bioucas-Dias et al. [4] is based on the least squares principle. The HySime algorithm is able to estimate the dimensions of a hyperspectral data subspace, and it considers the dimensions of the subspace as the number of endmembers of mixed spectral data. The algorithm first estimates the signal and noise correlation matrix, and then it selects the subset of feature vectors that express the signal subspace optimally in the form of a minimum root mean square error. The algorithm calculation utilizes information of the noise matrix. In the case of a high signal-to-noise ratio, the mixed spectral data must be denoised and preprocessed. In such cases, the accuracy of the HySime algorithm decreases. With a large number of mixed spectra, the estimation accuracy of the algorithm drops significantly.

The above algorithm uses the VD of mixed spectral data to estimate the number of endmembers. Compared with the VD, the intrinsic dimension (ID) has greater correlation with the number of endmembers. The algorithm used most widely directly calculates the ID value of mixed spectral data; however, the accuracy of the result is low and the geometric in-degree distribution (IDD) value has strong dependence on the ID. Relative to the estimation of the ID, the estimation of the in-degree value of spectral pixels is more convenient and accurate. The algorithm proposed in this paper draws the in-degree curve by calculating the in-degree value of each mixed spectral pixel, and then compares it with the in-degree curve in the material spectral library to determine the real number of endmembers.

Calculation. *In-degree distribution curve*. The central phenomenon is the effect that occurs when the ID of the data set increases. We treat each mixed spectral data element as a data point. If we create a K-nearest neighbor (KNN) graph [5] in a random data set, we will observe that some data points appear more as the ID of the data set grows. The in-degree value is the number of times a target point appears as the nearest neighbor of other points. A schematic of the calculation of the in-degree value is shown in Fig. 1.

The IDD curve of the mixed spectral data first counts the in-degree value of each spectral point and then calculates the ratio of the number of mixed spectral points under each in-degree value to the total number of mixed spectral points. The formula is as follows:

$$P\left(x_i\right) = N_{x_i} / N_{\text{all}} , \tag{1}$$

where N_{x_i} represents the number of mixed spectral points at the in-degree value of x_i, and N_{all} represents the total number of mixed spectral points. It has been documented that the IDD curve for a random graph will become a Poisson distribution [6], which will become very unmanageable for real world data sets with high IDs. The central phenomenon exists in many high-dimensional data sets and is related closely to the ID of the data [7]. The IDD curve has strong similarity for mixed spectral data with different IDs.

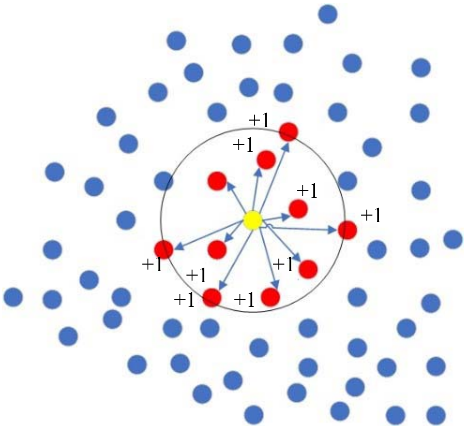

Fig. 1. Schematic of calculation of in-degree value.

Compared with an endmember number estimation algorithm using VDs, an algorithm based on the ID is more accurate for estimating the number of endmembers. The VD algorithm also regards natural features, anomalies, and disturbances during calculations as endmembers; thus, it overestimates the true number of endmembers [8]. The proposed IDD curve-matching algorithm is based on the ID, and experiment proved that its performance is superior to other endmember number estimation methods. Before drawing an IDD curve, it is necessary to establish a KNN graph for the mixed spectral data, i.e., the KNN points of each spectral point must be determined. There are many measures of distance such as Euclidean distance and spectral information divergence (SID). Here, we use SID values of different spectral projections in the feature enhancement space as a measure of the distance between spectral points [9].

Feature enhancement SID algorithm. Generally, the distance between different spectra is calculated according to the Euclidean distance, i.e., the KNN map of the mixed spectrum is constructed using the measure of Euclidean distance. In this paper, the feature-enhanced SID algorithm is used to calculate the difference between different mixed spectral data, because it is better able to reflect the degree of similarity between spectra. Full details of the SID algorithm can be found elsewhere [10], so only a brief introduction is presented here.

SID is derived from the field of information theory, and it is a commonly used measure of the degree of difference between two spectral curves (or multidimensional vectors). This method treats the two spectra to be compared as two random vectors, based on the probability of the difference between the elements in the two vectors, to reflect the degree of difference between the two vectors.

Suppose $a=[a_1,a_2,\cdots,a_L]$, $b=[b_1,b_2,\cdots,b_L]\in R^L$, and the information $I_i(a)$ and $I_i(b)$ is as follows:

$$I_i(a)=-\log p_i\,,\ I_i(b)=-\log q_i\,, \tag{2}$$

where $p_i=a_i/\sum_{j=1}^{L}a_j$ and $q_i=b_i/\sum_{j=1}^{L}b_j$. Then, the SID of a and b is

$$\mathrm{SID}(a,b)=D(a\,\|\,b)+D(b\,\|\,a)=\sum_{i=1}^{L}p_i\log(p_i/q_i)+\sum_{i=1}^{L}q_i\log(q_i/p_i). \tag{3}$$

The SID value in Eq. (3) represents the degree of similarity between the spectra, where a small SID value corresponds to a high degree of similarity.

The feature-enhanced SID algorithm adds a feature extraction process before the spectral matching algorithm, maps the original spectral data to the feature enhancement space, and enhances important band information when calculating the similarity of the spectra in the feature enhancement space. Important band information refers to more accurate matching results [11]. A brief introduction of the feature-enhanced SID algorithm is presented in the following.

Suppose Y is a mixed spectral matrix, λ_1, λ_2, …, $\lambda_n(\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_n)$ is the characteristic value of YY^T, $e_1,e_2,\cdots,e_n$ is the feature vector, and n represents the band number. Then, select m eigenvalues according to the eigenvalue size, where the selection criterion of m is that the maximum value should make $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_m>>\lambda_{m+1}\geq\lambda_{m+2}\geq\ldots\geq\lambda_n$. Thus, a new diagonal array Λ can be constructed as follows:

$$\Lambda=\begin{bmatrix} \sqrt{\frac{\lambda_1}{\text{sum}}} & 0 & \cdots & 0 \\ 0 & \sqrt{\frac{\lambda_2}{\text{sum}}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{\frac{\lambda_m}{\text{sum}}} \end{bmatrix}, \tag{4}$$

where $\text{sum} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$; thus, the algorithm considers that the larger the value of λ, the higher the value of the matching feature in the mixed spectrum. Define $P = [e_1, e_2, \cdots, e_m]$. Then, the original mixed spectrum is first projected into the m-dimensional subspace, and the features in the subspace are enhanced to obtain a spectral feature enhancement space. Define $F = \Lambda P$, where F is regarded as the spectral feature enhancement space. The data set after data set Y is projected into the feature enhancement space is Z. Therefore, the original mixed spectral data can be expressed as follows:

$$Z = \Lambda X = \Lambda PY\ . \tag{5}$$

The original data set Y is projected into the feature enhancement space and becomes data set Z. The degree of similarity between data set Z and the original spectrum is determined by calculating the SID value of Z. The feature-enhanced matching algorithm has the following advantages: it reduces the redundancy of spectral data, improves matching precision, and reduces the influence of noise on the data. In this paper, the feature-enhanced SID value is used as the standard for constructing the mixed spectrum KNN map, and the IDD curve-matching algorithm is used on this basis.

The specific steps of the algorithm are as follows:

1. Establish a material spectrum library with n band numbers.
2. Randomly select 3 to n spectral material spectra from the database to generate mixed spectral data, and plot IDD curves to form an IDD curve library corresponding to the number of endmembers.
3. The KNN graph of the mixed spectrum is established using the SID of the feature enhancement space, counting the in-degree value of each mixed spectral point, and drawing the IDD curve.
4. Match the IDD curve with a curve in the established IDD curve library to find the closest curve and determine the number of endmembers.

Results and discussion. *FE-IDD algorithm validity analysis.* The proposed IDD algorithm is intended to overcome the problem of estimating the number of endmembers for the case of a small number of bands. Experiments were conducted to verify the validity of the IDD algorithm using 16-band spectral data obtained from the United States Geological Survey (USGS) database. Referring to the USGS database, 12 spectra were randomly selected as pure material endmembers with a wavelength range of 370–2500 nm and an overall total of 224 bands. The 12 end elements of the material library are shown in Fig. 2. The selected 12 endmember spectra were used to simulate the end-band spectra of the 16 bands. As the number of USGS bands used was 224, the average of groups of 14 bands was simulated to produce an endmember spectrum of each of the 16 wavelengths.

First, three endmember spectra were randomly selected, and an abundance value was randomly generated using the Dirichlet distribution function. The simulation experiment was performed 100 times, and the number of mixed spectra was 10,000 each time. The feature enhancement SID value was used as a standard to construct the KNN graph of the mixed spectrum. The value of K was 10, which means for each mixed spectrum the 10 spectra closest to the mixed spectrum were determined. Then, the in-degree value was specified as the number of times that a spectrum appeared as the nearest neighbor of the other spectrum. Finally, the in-degree value of each mixed spectrum was counted. To produce the graph, the in-degree value is plotted on the x-axis and the IDD curve is drawn on the ordinate with the ratio of the total number of mixed spectra under the same in-degree value.

The results of the 100 simulation experiments are shown in Fig. 3. The IDD curves of the mixed spectra are similar, with only two IDD curves plotted outside the specific curve trend. In the case of the same endmember with different abundance values, the IDD curve obtained by the mixed spectrum has a certain regularity, and the degree of similarity is very high, demonstrating that different abundance values of the same endmember can produce a stable IDD curve.

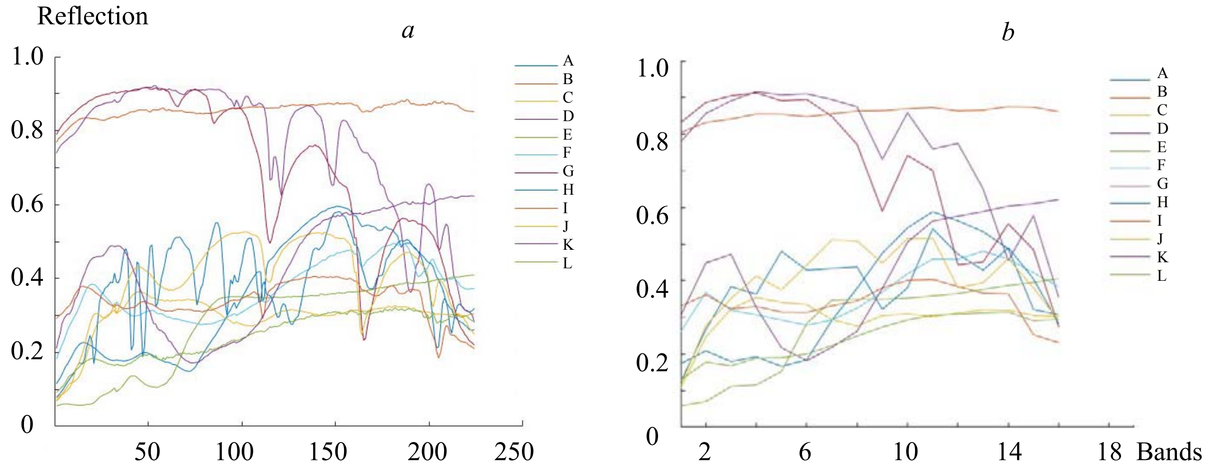

Fig. 2. (a) The 12 endmembers selected from the USGS database with 224 bands and (b) the 12 endmembers with 16 bands generated by the endmembers in (a).

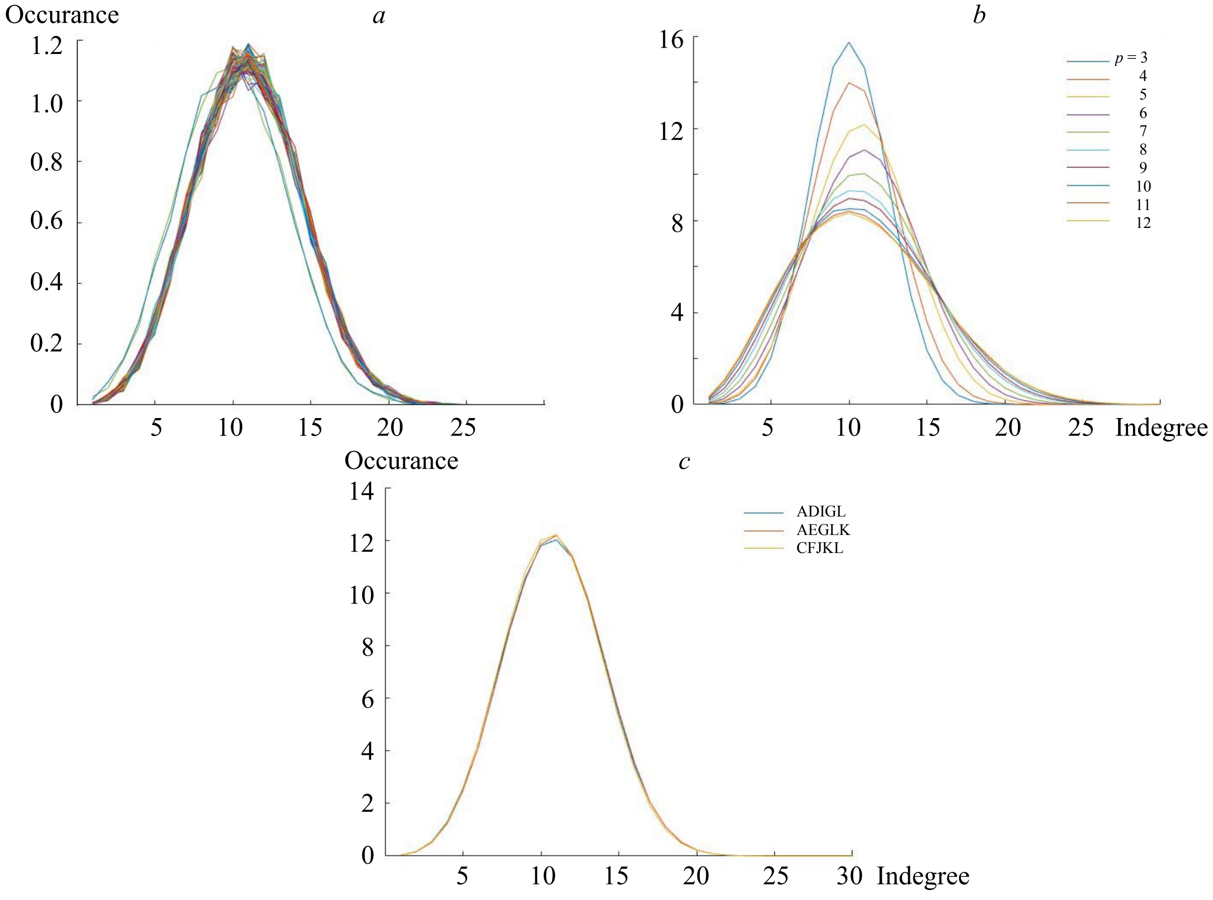

Fig. 3. (a) IDD curves corresponding to the mixed spectrum of different endmembers, (b) IDD curves corresponding to the mixed spectrum of different endmember numbers, and (c) IDD curves corresponding to the mixed spectrum of different endmembers with the same number of endmembers.

Second, it can be seen from Fig. 3b that the shapes of the IDD curves corresponding to different endmember numbers are similar but with obvious differences. Thus, they could be used as a basis for judging the number of endmembers. The IDD curves corresponding to different endmembers are similar to the pin curve of a normal distribution, but the widths and heights of the curves are different, making them easily distinguishable.

Finally, it is judged whether the mixed spectral data of different endmembers under the same number of endmembers have the same IDD curve. The same number of endmember spectra are arbitrarily selected in the spectral library, and the generated IDD curve is as shown in Fig. 3c. It can be seen from Fig. 3c that in the same spectral library, even if the endmember spectra are different, the IDD curves corresponding to the mixed spectra produced by the same number of endmembers are different. This indicates that the algorithm could be used to predict the real number of endmembers from a mixed spectrum produced using a spectrum from the spectral library.

In summary, the IDD curves obtained using the mixed spectrum for the cases of different endmembers and different numbers of end elements have a certain degree of distinguishability. Moreover, the IDD curves corresponding to the mixed spectrum of different endmembers with the same number of endmembers also have strong similarities. Therefore, the curves could be used as a method to judge the number of mixed spectral endmembers based on the premise of having a spectral library reference.

Experiments. The experiments were intended to verify the accuracy of the algorithm using a small number of bands. The simulation experiments were conducted to estimate the number of endmembers for mixed spectral data of 16 bands. The experiments used the endmember spectral data from the USGS database. Twelve spectra were randomly selected from the spectral library as pure material endmembers with a wavelength range of 370–2500 nm. As the number of USGS bands used was 224, the average of groups of 14 bands was simulated to produce an endmember spectrum of each of the 16 wavelengths. The 12 end elements of the bands generated by the simulation are shown in Fig. 2b. Endmember numbers of 3, 5, and 7 and the abundance were randomly assigned to different endmembers according to the Dirichlet distribution function to generate a spectral mixing matrix containing 1000 spectra. Considering the noise factor, the spectral mixing matrix is superimposed with Gaussian white noise with a signal-to-noise ratio of 50–20 dB.

In the experiments, the HySime, HFC, and IDD curve-matching algorithms were compared. The HFC algorithm needs a false alarm rate to be specified manually. Different false alarm rates were chosen for the comparison of the estimation results of the HFC algorithm ($P_f = 10^{-3}$, 10^{-4}, 10^{-5}). The HySime algorithm requires preprocessing to be performed to obtain a noise correlation matrix. Each algorithm was run more than 100 times, and the average result for each was determined (Table 1).

As shown Table 1, all the algorithms could accurately estimate the number of endmembers for the case of a high signal-to-noise ratio. However, as the signal-to-noise ratio decreased, the accuracy of all algorithms also decreased. The accuracy of the HFC algorithm decreases because as the noise increases, the noise variance becomes larger. The noise variance of more bands is greater than the warning value, which leads to an increase in the estimation of the number of endmembers. The accuracy of the HySime algorithm is degraded because the correlation between the bands is masked by the increase in noise.

In comparison with the HFC and HySime algorithms, the average value of the IDD algorithm is closer to the true value, and the accuracy is higher. Nevertheless, as the signal-to-noise ratio increases, the accuracy of the algorithm decreases, indicating that increased noise has certain influence on the IDD curve-matching algorithm. The algorithm cannot improve the accuracy of estimation for cases of high signal-to-noise ratio. Therefore, the IDD algorithm should be used in conjunction with a denoising preprocessing algorithm to

TABLE 1. Estimation Results of Three Algorithm

		Gaussian white noise (1000 pixel)			
SNR, dB	Method	$P = 3$	$P = 5$	$P = 7$	$P = 10$
50	IDD	**3.02**	**5.08**	**7.12**	**9.68**
	Hysime	3.14	5.54	7.95	8.05
	HFC (P_f= 10^{-3})	2.24	3.95	6.21	7.49
40	IDD	**3.25**	**5.21**	**7.25**	**9.87**
	Hysime	3	4.95	6.63	6.89
	HFC (P_f= 10^{-3})	3.58	4.24	6.45	8.00
30	IDD	**3.87**	**5.24**	**6.54**	**8.54**
	Hysime	3.00	4.21	4.52	4.03
	HFC (P_f= 10^{-3})	9.71	10.67	12.53	14.04
20	IDD	**3.94**	**6.54**	**7.98**	**10.98**
	Hysime	2.48	2.56	2.50	2.11
	HFC (P_f= 10^{-3})	15.44	15.28	15.84	16.00

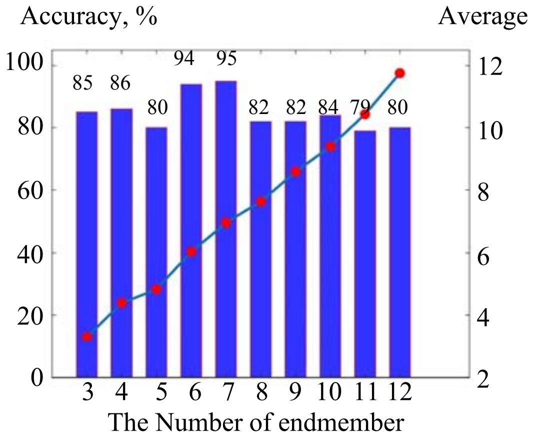

Fig. 4. IDD algorithm accuracy.

make the estimation result more accurate. In the case of a high signal-to-noise ratio, the difference between the average value of the estimated number of endmembers obtained from the IDD algorithm and the actual number of end elements was less than 1 in 100 simulation experiments. The accuracy of the algorithm in 100 simulation experiments is shown in Fig. 4. It can be seen from Fig. 4 that the average estimation accuracy of the IDD algorithm is >80% for cases with different numbers of endmembers. Although the estimated result is poor when the number of endmembers is 11 and 12, the error between the estimated average value of the number of endmembers and the actual value is <1.

Conclusions. An endmember number estimation algorithm is the premise of an endmember extraction algorithm. Current research in the field of endmember quantity estimation is imperfect, and most algorithms rely on experience to determine the number of endmembers. Consequently, the accuracy of spectral unmixing algorithms can be affected by noise in the spectral data and the personal experience of the interpreter. Compared with the traditional HFC and HySime endmember number estimation algorithms, the IDD algorithm proposed in this paper demonstrated improved average accuracy based on 100 simulation experiments. Given the small number of spectral bands of most space targets, simulation experiments were performed for the case of 16 mixed spectral bands. The results showed that the number of endmembers could be estimated accurately when the number of endmembers paper, which is based on a known spectral library of the types of constituent material of a space target, is suitable for estimation of the number of end elements of a space target.

Acknowledgments. This work was supported by the National Natural Science Foundation of the P.R. China under Grant No. 61575015.

REFERENCES

1. R. Heylen, P. Scheunders, *IEEE J. Select. Top. Appl. Earth Observ. Rem. Sens.*, **6**, N 2, 570–579 (2013).
2. E. Levina, P. J. Bickel, *Int. Conf. Neural Inform. Proc. Systems*, MIT Press, 777–784 (2004).
3. I. Chein Chang, *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.*, **11**, N 4, 1285–1305 (2018).
4. S. Devika, S. M. K. Chaitanya, *IEEE Int. Conf. Res. Adv. Integrated Navigation Systems* (2016).
5. J. M. Keller, M. R. Gray, J. A. Givens, *IEEE Trans. Syst., Man. Cybern.*, SMC-15(4), 580–585 (2012).
6. M. E. Newman, S. H. Strogatz, D. J. Watts, *Phys. Rev. Statist. Nonlinear Soft Matter Phys.*, **64**, N 2, 026118 (2001).
7. O. Gaci, S. Balev, *IEEE Int. Conf. Bioinform. Biomed. Workshop*, 107–112 (2009).
8. A. Robin, K. Cawse-Nicholson, A. Mahmood, M. Sears, *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.*, **8**, N 6, 2854–2861 (2015).
9. Nelson Antunes, Shankar Bhamidi, Tianjian Guo, Vladas Pipiras, Bang Wang, *Sampling-Based Estimation of In-degree Distribution with Applications to Directed Complex Networks*, **40**, N 5, 501–505 (2018).
10. Wanjun Liu, Xiuhong Yang, Haicheng Qu, Yu Meng, *J. Comput. Appl.*, **35**, N 3, 844–848 (2015).
11. Q. Li, C. Niu, *J. Appl. Remote Sens.*, **9**, N 1, 096008 (2015).