UNSUPERVISED ENDMEMBER EXTRACTION: APPLICATION TO HYPERSPECTRAL IMAGES FROM MARS

Bin Luo, Jocelyn Chanussot[*]

GIPSA-Lab, 38402 Grenoble, France.
({Bin.luo, jocelyn.chanussot}@gipsa-lab.inpg.fr)

Sylvain Douté

LPG 38041 Grenoble, France.
(sylvain.doute@obs.ujf-grenoble.fr)

ABSTRACT

In this paper, we try to identify and quantify the chemical species present on the surface of planet Mars with the help of hyperspectral images provided by the instrument OMEGA. For this purpose, we suppose that the spectrum of each pixel is a linear mixture of the spectra of different endmembers. From this linear mixture hypothesis, our work is divided into two steps. Firstly, we propose a new unsupervised method for estimating the number of endmembers based on the eigenvalues of covariance and correlation matrix of the hyperspectral data. This method is then validated on synthetic data. With the help of the number estimated by the precedent step, we use the Vertex Component Analysis (VCA) to extract the spectra and the abundances of the endmembers. The results on hyperspectral image taken by the instrument OMEGA are shown.

1. INTRODUCTION

Visible and near infrared imaging spectroscopy is a key remote sensing technique to study the Earth and other planets. As solar light is partially transmitted, reflected and diffused back by interaction with the different constituents of the atmosphere and the surface, the analysis of reflectance spectra may allow the identification and quantification of the chemical species. In Figure 1, we show an example of a hyperspectral image cube captured by OMEGA on the south pole of Mars. From this image, we would like to identify the chemical species (such as CO_2, water, etc.) present in this image and estimate their abundances. For this purpose, we consider that the spectrum of each pixel in the hyperspectral image is a linear mixture of the spectra of different endmembers. Based on this linear mixture model, we divide our work into two steps.

The first step consists in estimating the number of chemical species present in the image. This number is required before performing unmixing. For this purpose, PCA-based thresholding is a standard method [1]. However, the cut-off threshold is not easy to determine since the eingenvalues caused by the signals and noise are sometimes very similar. In [2], an automatic eingenvalue thresholding is proposed based on the difference of the eigenvalues of the sample correlation matrix and the eigenvalues of the sample covariance matrix (referred to as HFC method). However, this approach requires to fix the false alarm probability α, which affects the estimated number of endmembers. Moreover, when the noise level of the data is high, the HFC approach cannot give the exact number. Therefore, in this paper, we propose a method based on the likelihood function of the distribution of the eigenvalue differences for determining the number of endmembers, which does not need any parameter and is more accurate with high noise level.

Afterwards, with the help of the linear mixture model, we can consider the problem of separating the mixture of spectra on each pixel as a blind source separation problem. In [1], the authors have proposed to use the Independent Component Analysis (ICA) to separate the spectra of OMEGA data by assuming that the spectra of the endmembers are independant. However, as mentioned in [3], the ICA could not lead to a good result, as the assumption of independence is not valid. The sum of abondances of the chemical species on a single pixel must be one (so called sum-to-one condition). In [4] , the Vertex Component Analysis (VCA) is proposed as an efficient method for extracting the endmembers of hyperspectral data by using both the linear mixture model and the sum-to-one condition. In this paper, we choose VCA to unmix the spectra of each pixel with the help of the number of endmembers estimated by the previous step.

The plan of this paper is as follows. In Section 2, we present the linear mixture model used in this paper. In Section 3, we introduce our approach for estimating the number of endmembers present in the image. This is the main contribution of the paper. In Section 4, we present very briefly the method VCA. In Section 5 and 6, we present the unmixing results on both synthetic and OMEGA data.

2. LINEAR MIXTURE MODEL

We note $\mathbf{X}$ the matrix representing the hyperspectral image cube with N_a pixels and N_s spectral bands, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_a}\}$ and $\mathbf{x}_k = \{x_{1,k}, x_{2,k}, \ldots, x_{N_s,k}\}^T$, $x_{l,k}$ is the value of the kth pixel at the lth band. We assume that the spectrum of each pixel is a linear mixture of the spectra of N_c endmembers, leading to the following model:

[*]This work is funded by French ANR project VAHINE.

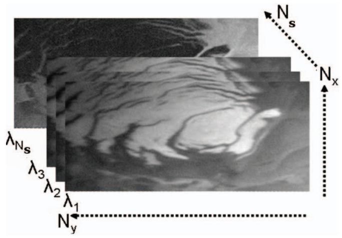


Fig. 1. Hyperspectral image taken by the instrument OMEGA on the south pole of Mars.

$$\mathbf{X} = \mathbf{MS} + \mathbf{n} \tag{1}$$

where $\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{N_c}\}$ is the mixing matrix where $\mathbf{m}_n$ denotes the spectral signature of the nth endmember. $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{N_c}\}^T$ is the abundance matrix where $\mathbf{s}_n = \{s_{n,1}, s_{n,2}, \ldots, s_{n,N_a}\}$ ($s_{n,k} \in [0,1]$ is the abundance of the nth endmember at the kth pixel). **n** stands for the additive noise of the image. For separating **M** and **S** from **X** without any *a priori* information, we first have to estimate the number N_c of endmembers . In a second step, we can perform a linear unmixing in order to obtain **M** and **S**.

3. ESTIMATION OF THE NUMBER OF ENDMEMBERS

In [2], it is proposed to threshold eigenvalues of the covariance and correlation matrix for estimating the number of endmembers. We note K the sample covariance matrix of **X** and R its correlation matrix. Suppose that λ_i and $\hat{\lambda}_i$ are respectively the ith eigenvalues of K and R with $i \geq 0$, $\lambda_i > \lambda_{i+1}$ and $\hat{\lambda}_i > \hat{\lambda}_{i+1}$. If the noise is Gaussian with zero mean and there are N_c endmembers present in **X**, the eigenvalues $\hat{\lambda}_i, (i > N_c)$ and $\lambda_i, (i > N_c)$ correspond to the noise variance, we have therefore:

$$\begin{cases} \hat{\lambda}_i - \lambda_i > 0, & i \leq N_c \\ \hat{\lambda}_i - \lambda_i = 0, & i > N_c \end{cases} \tag{2}$$

Noting $z_i = \hat{\lambda}_i - \lambda_i$, a Neyman-Person test can be used to threshold the z_i value in order to estimate the number of endmembers [2]. Even though in [2], experiments show that the proposed method gives much better results than the classical approaches such as MDL or AIC, since the noise of each band is not identical and independent. However, this method has two main drawbacks. Firstly, one has to fix the false alarm value in order to determine the threshold for z_i, which can affect the estimated number of endmembers. Secondly, if the SNR of the image is too low, this method can not give the exact number.

In this section, we propose to use the distribution of z_i values for estimating the number of endmembers without any parameter, which can give a precise number of endmembers even if the SNR is very low. The distribution of z_i can be modeled by [5]:

$$\begin{array}{ll} z_i \sim \aleph(\mu_i, \sigma_i^2), & i \leq N_c \\ z_i \sim \aleph(0, \sigma_i^2), & i > N_c \end{array} \tag{3}$$

where μ_i is unknown and σ_i can be given by $\sigma_i^2 \approx \frac{2}{N}(\hat{\lambda}_i^2 + \lambda_i^2)$, if the number of samples is sufficiently large (which is usually the case for hyperspectral images). [2][5]

According to Equation (3), we define a likelihood function $H(i)$ by $H(i) = \Pi_{l=i}^{N_s} \frac{1}{\sigma_l} \exp(-\frac{z_l^2}{2\sigma_l^2})$. We then take the logarithmic value of $H(i)$,

$$\tilde{H}(i) = \log H(i) = A(i) + B(i) \tag{4}$$

where $A(i) = -\sum_{l=i}^{N_s} \frac{z_l^2}{2\sigma_l^2}$ and $B(i) = -\sum_{l=i}^{N_s} \log \sigma_l$. It is evident that $H(N_c + 1) > H(i)$ when $i \leq N_c$, because the mean value of z_i for $i < N_c$ is not zero. When $i \geq N_c + 1$, $A(i)$ will change very slightly as the function of i. And if we normalize all the values of **X** into $[0, 1]$, $-\log \sigma_i$ will be positive and $B(i)$ decreases as a function of i when $i \geq N_c + 1$, as well as $\tilde{H}(i)$. We can have a global *maximum* of $\tilde{H}(i)$ at $i = N_c + 1$. Therefore the number of endmembers can be defined by:

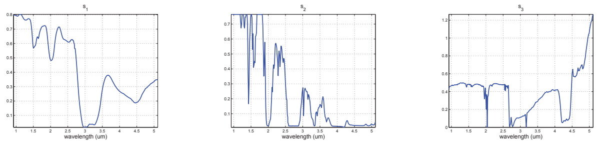
$$\hat{N}_c = \arg\max_i \{\tilde{H}(i)\} - 1 \tag{5}$$

4. VERTEX COMPONENT ANALYSIS

In [4], the Vertex Component Analysis (VCA) is proposed as an effecient method for extracting the endmembers which are linearly mixed. The main idea is to extract the vertex of the simplex formed by **M** which contains all the data vectors in **X**. According to the sum-to-one condition, the sum of the abundances of all the endmembers for each pixel is equal to one, *i.e.* $\forall k, \sum_{n=1}^{N_c} s_{n,k} = 1$. Therefore the data vectors $\mathbf{x}_l$ are always inside a simplex of which the vertex are the spectra of the endmembers if there is no noise. VCA iteratively projects the data onto the direction orthogonal to the subspace spanned by the endmembers already determined. And the extreme of this projection is the new endmember signature. The algorithm stops the iteration when all the p endmembers are extracted, where p is the number of endmembers which has to be fixed before performing VCA. In the practice, we fix $p = \hat{N}_c$ determined by the approach presented in Section 3.

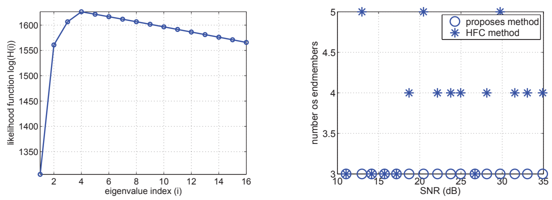
5. UNMIXING RESULTS ON SYNTHETIC DATA

The Grenoble Planetology Laboratory (LPG) provided the reference spectra, sampled at 256 wavelength in infra-red and visible light, of three chemical species (H_2O, CO_2 and the $Dust$) which are the most present on planet Mars. Therefore each spectrum is a discrete signal with $N_s = 256$ samples.



(a) $\mathbf{m}_1$: H_2O (b); $\mathbf{m}_2$: CO_2; (c) $\mathbf{m}_3$: $Dust$

Fig. 2. Reference spectral signatures of the 3 endmembers.



(a) (b)

Fig. 3. (a) Likelihood function $\tilde{H}(i)$ when $SNR = 11.1dB$; (b) Number of endmembers estimated by our approach (represented by circles) and the approach of [2] (represented by cross) as the function of SNR.

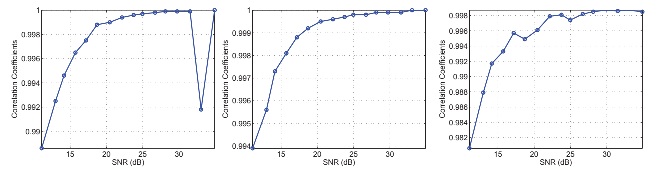
The three spectra, corresponding to the endmembers used to synthetize the data are shown in Figure 2.

The abundance of each endmember is a positive random signal with length $N_a = 32 \times 32$ generated by identical independent Gaussian distribution. We then normalize the abundances subject to the sum-to-one condition. Therefore, we have a matrix $\mathbf{M} \times \mathbf{S}$ with size $N_a \times N_s$. We add Gaussian noise at each row of the data matrix. The variance of the Gaussian noise for each row is generated randomly by a Gaussian distribution in the interval $[0, \sigma_{\max}]$. By varying $\sigma_{\max}$, we can generate images with different SNR.

In Figure 3(a), we have shown the likelihood function $\tilde{H}(i)$ (Equation (4)). It can be observed that, as explained in Section 3, $\tilde{H}(i)$ increases very fast at first since the first few values of z_i are not distributed as $\aleph(0, \sigma_i)$. After its global *maximum* reached at $i = 4$, it decreases relatively slowly.By using Equation (5), it can be drawn that the number of endmembers is 3, which is the number we used for simulating the data.

In Figure 3(b), we have compared the number of endmembers estimated by our appoach presented in Section 3 and the results of the eingenvalue thresholding method presented in [2]. For the method of [2], we fix the false alarm probability always to $\alpha = 0.001$. It can be observed that the number estimated by the method in [2] varies when the SNR changes. In contrast, our approach does not need any parameter and the estimated number of endmembers is always 3, which is correct and robust.

Afterwards we show the unmixing results obtained by VCA on synthetic data. As previously discussed, the proposed approach provided a robust unsupervised estimation of the number of endmembers ($\hat{N}_c = 3$). Therefore by using VCA, for each data matrix $\mathbf{X}$ at a given SNR, we can obtain 3 endmember signatures. We note $\hat{\mathbf{M}} = \{\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \hat{\mathbf{m}}_3\}$ the spectral signatures of the endmembers extracted by VCA. In order to compare the obtained signature ($\hat{\mathbf{M}}$) with the signature used for simulating the data ($\mathbf{M}$) , we calculate the correlation coefficients matrix $\mathbf{R}_m$ between $\hat{\mathbf{M}}$ and $\mathbf{M}$. In Figure 4, we show the correlation coefficients between the spectral signatures of the endmembers extracted by VCA with different noise levels and the most similar reference signatures. It is observed that the correlation coefficients are always close to 1. In Figure 5, we show the reference spectral signatures and the signatures extracted by VCA when $SNR = 11dB$. It can be observed that even though the data is noisy, as well as the extracted signatures, they fit very well the references.



(a) (b) (c)

Fig. 4. Correlation coeffients for (a) $\mathbf{m}_1$; (b) $\mathbf{m}_2$; (c) $\mathbf{m}_3$ calculated under different SNR.

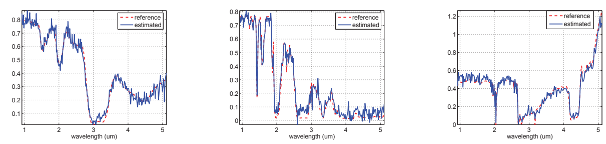


Fig. 5. Extracted spectra of H_2O, CO_2 and $Dust$ by VCA when $SNR = 11dB$. Solid lines represent the spectral signatures obtained by VCA; while the dash lines represent the most similar reference spectra.

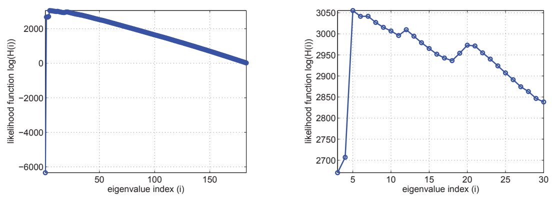
6. UNMIXING RESULTS ON OMEGA DATA

In this Section, we apply the proposed unmixing procedure on the hyperspectral image from the south pole of Mars shown in Figure 1. The spatial resolution of the image is $300m$. Each image contains 256 bands which cover the visible and infrared range. For the unmixing task, we have removed 73 bands which are mainly water absorbtion bands and corrupted by the atmospheric correction.

In Figure 6, the likelihood $\tilde{H}(i)$ calculated on the image is shown. It can be observed that at $i = 5$, there is a global *maximum*. Hence the number of endmembers is $\hat{N}_c = 4$.

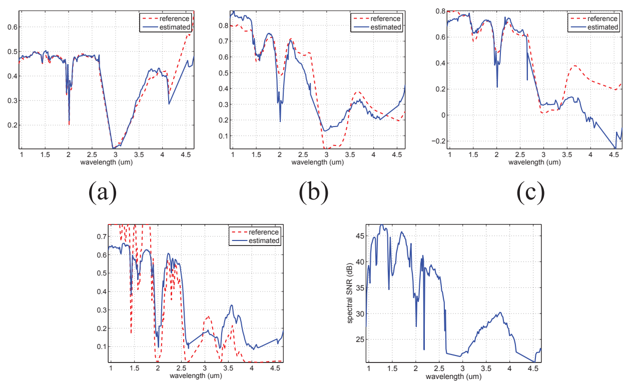
In Figure 7, we show the 4 spectral signatures ($\hat{\mathbf{M}} = \{\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \hat{\mathbf{m}}_3, \hat{\mathbf{m}}_4\}$) of the endmembers extracted by VCA. By using the linear assumption, we can approximate the abundances of the endmembers $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \hat{\mathbf{s}}_3, \hat{\mathbf{s}}_4\}$ by

$$\hat{\mathbf{S}} \approx \hat{\mathbf{M}}^{-1}\mathbf{X} \tag{6}$$



(a) (b)

Fig. 6. (a) Likelihood function $\tilde{H}(i)$ calculated on the image of Mars; (b) $\tilde{H}(i)$ values at the range of $i \in [3, 30]$. It can be observed that $\tilde{H}(i)$ reaches its global *maximum* at $i = 5$.



(a) (b) (c)

(d) (e)

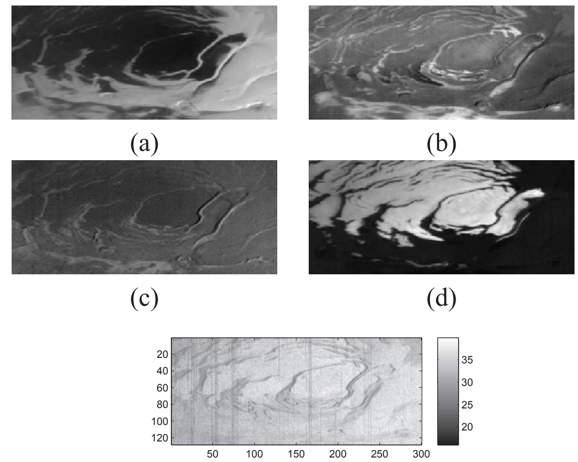
Fig. 7. In (a)-(d), the solid lines represent the 4 spectral signatures obtained by VCA; while the dash lines represent the most similar reference spectra. (e) Spectral SNR of the approximation made by Equation (6).

In Figure 8, we show the spatial abundances of the endmembers. Figure 7(e) and 8(e) show respectively the spectral and spatial SNR which represent the approximation error caused by Equation (6).

In order to identify the extracted endmembers, we have compared the extracted spectral signatures $\hat{\mathbf{M}}$ with the reference spectral signatures of CO_2, H_2O and $Dust$ shown in Section 5. For each signature extracted by VCA, we also show the most similar reference signature in Figures 7(a)-(d). It can be observed that the first, second and the fourth extracted signatures correspond to the signatures of $Dust$, H_2O and CO_2 respectively. Therefore the Figure 8(a), (b) and (d) are the approximate abundances of $Dust$, H_2O and CO_2. From Figure 7(c) we can see that this signature is similar to the reference signature of H_2O. However, Figure 8(c) shows that this endmember corresponds to the line shift of the OMEGA instrument between the visible and the infra-red range.

7. CONCLUSION

In this paper, we try to indentify and quantify the chemical species present on the hyperspectral image of Mars based on the hypothesis that the spectra of each pixel is a linear mixture



(a) (b)

(c) (d)

(e)

Fig. 8. (a)-(d) Spatial abundances estimated by Equation (6) by using the 4 endmembers extracted by VCA. (e) Spatial SNR of the approximation made by Equation (6).

of spectra of different chemical species. Our work include two steps. Firstly, we have proposed a new method for estimating the number of endmembers based on the eigenvalues of the covariance and correlation matrix of the hyperspectral data. Our method is parameter free. The validation on synthetic data shows that our method can accurately estimate the number of endmembers. This estimation is then used as an input for the VCA algorithm, which separates the spectra and the abundances of the different chemical species of each pixel.

8. REFERENCES

[1] S. Moussaoui, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J. Benediktsson, "On the decomposition of mars hyperspectral data by ICA and bayesian positive source separation," *Neurocomputing, Neurocomputing for Vision Research; Advances in Blind Signal Processing*, vol. 71, pp. 2194–2208, 2008.

[2] C. I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 608–619, March 2004.

[3] J. M. P. Nascimento and J. M. B. Dias, "Does independent component analysis play a role in unmixing hyperspectral data?," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 1, pp. 175 – 187, January 2005.

[4] J. M. P. Nascimento and J. M. B. Dias, "Vertex component analysis: A fast algorithl to unmix hyperspectral data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898 – 910, April 2005.

[5] T.W. Anderson, *An introduction to multivariate statistical analysis*, Springer-Verlag, 2nd edition, 1984.