Using Principle Component Analysis of Spectral Mixtures to Analyze Tertiary and Four End-member Mixtures Containing Carbonates and Olivine

David Burnett and Joseph Makarewicz
Olivet Nazarene University Pence Boyce Research Program

Introduction

CRISM images from Mars are expected to contain carbonates such as magnesite [1]. Prior research has been successfully able to determine the approximate percent composition of phyllosilicates in binary lab mixtures using Principle Component Analysis (PCA) [2]. In order to expand this model to work on CRISM images, one of preliminary steps is allowing the algorithm to work on mixtures with more than two components.

Principle Component Analysis

In a previous study, PCA was used to analyze a set spectrum to determine percentages of each end-member in a binary mixture [2]. While principle component analysis was preferable over the alternative methods due to being able to run without user intervention once the model is set up, it still required some user intervention to set up the model. This additional intervention is minimal in a binary mixture, but during testing with a tertiary mixture it became much more difficult to properly identify principle component correspondences. To solve this issue, a method of automatically classifying principle components was required.

Methods and results

Samples: For the sake of this test, two sample sets were used. The first set contains 19 spectra forming a tertiary mixture of magnesite (Mgs), nontronite (Non), and forsterite (Fo) formed by weight percentage [1]. This set was chosen due to its relevance to the predicted elements in the CRISM images [1]. The second set contains 44 spectra forming a four end-member mixture of olivine (OL), plagioclase (PLG), hypersthene (LCP, Low-CA Pyroxene), and augite (HCP, Hi-Ca Pyroxene) [3]. It was chosen due to it having one of the largest number of spectrum available for four end-member mixtures in the Relab Spectral Database [3].

Pre-processing: Before analyzing the spectra, they were pre-processed similarly to the preprocessing done in the previous study [2]. For each of the data sets, instead of using pre-chosen start and end wavelengths, the values were cropped based on the latest starting wavelength and the earliest ending wavelength, ensuring as much

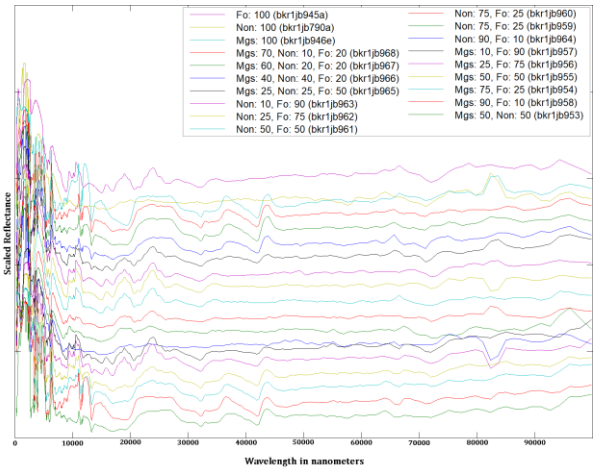

Fig. 1: Relab magnesite, nontronite, and forsterite mixture after processing with offsets.

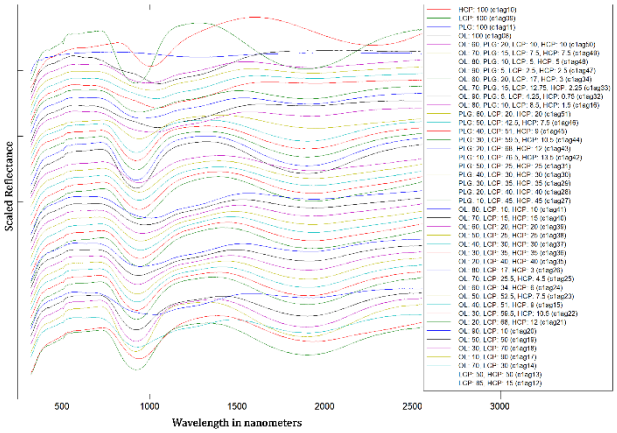

Fig 2: Relab olivine (OL), plagioclase (PLG), hypersthene (LCP), and augite (HCP) mixture after processing with offsets.

available data is used. After that all spectra were resampled using the first wavelength set to ensure they all have the same number of data points which correspond to the same wavelengths. After resampling, each of the spectra were normalized so that the area is 1 to reduce the effect of albedo variations [2].

Procedure: Each of the data sets was processed separately. For each set, a spectra matrix is created using the spectra as the rows of the matrix. Labels were applied to each row in the spectrum matrix corresponding to the weight percentages of the individual end-members; these labels will be used later to correlate the principle

components. The mean of the matrix was subtracted from each spectrum to reduce potential noise of common elements.

Once the matrix is ready, principle component analysis uses the matrix to produce a set of principle component values, a set of matching principle component vectors, and a projection matrix. The projection matrix allows producing later principle component values from other spectrum not found in the matrix.

Principle component correlations: Since there are now more than two end-members in the mixture, it is much more difficult to guess which principle component corresponds to each percentage by inspection; before a single percentage could fully describe the mixture, while now it takes at least two different percentages. Each different principle component needs to be compared to each different end-member's percent values to find the correlations. To simplify the process, a linear fit was constructed for each principle component in relation to each different percent value. Once constructed, the linear fit can be given the same principle component values as inputs to determine the accuracy of the data by means of a graphical comparison.

As shown in figures 3-5, this works well for manually selecting the principle component for each end-member, as the matching principle component is visually distinct from the non-matching components. In order to automatically correlate them, an R^2 score is calculated from the original values and the values calculated using the linear fit. This score is typically between 0 and 1, with higher scores correlating with a better linear fit. From the results (excluding scores below 0.1), magnesite had a R^2 score of 0.14 for PC1 and 0.82 for PC2; nontronite had 0.89 for PC1; and forsterite had 0.32 for PC1 and 0.63 for PC2.

These values match with the principle component that is shown most accurate in the figures above. It is also worth noting that forsterite, which did not get assigned a principle component has comparatively high scores on both PC1 and PC2, as opposed to the two that were matched have one high score and the rest low.

Principle component supplementing: During testing, it was found that all but one of the end-members in the mixture will have a unique principle component. Since the principle components are assumed to represent a real-world value, that means it should not be possible for two end-members to share the same principle component. As a result, in any mixture all but one of the end-members will have a principle component and a linear fit while the final end-member will have no principle component or linear fit.

Since all the end-member's percentages should sum to 100% (essentially based on the sum to one constraint),

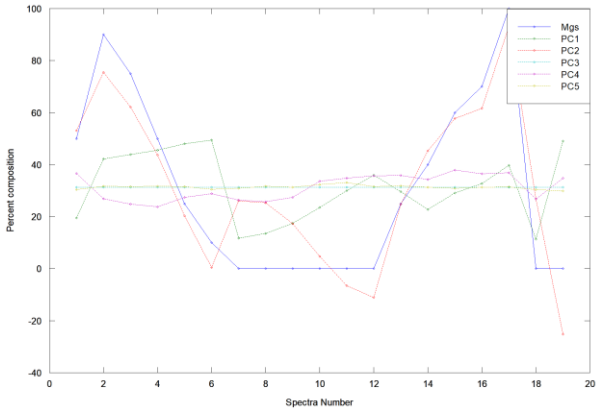

Fig 3: Comparison of the actual magnesite percentages to those calculated by each linear fit. As seen in the figure, PC2 is the closest.

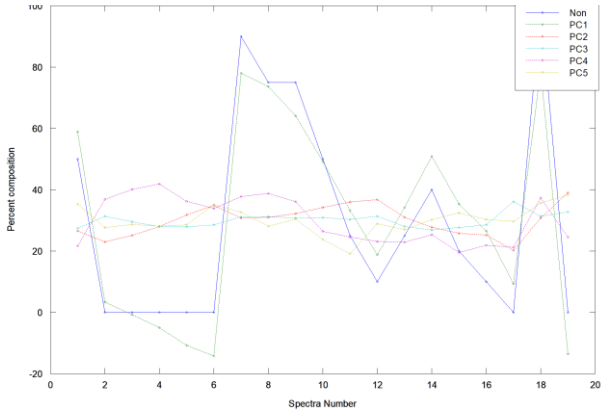

Fig 4: Comparison of the actual nontronite percentages to those calculated by each linear fit. As seen in the figure, PC1 is the closest.

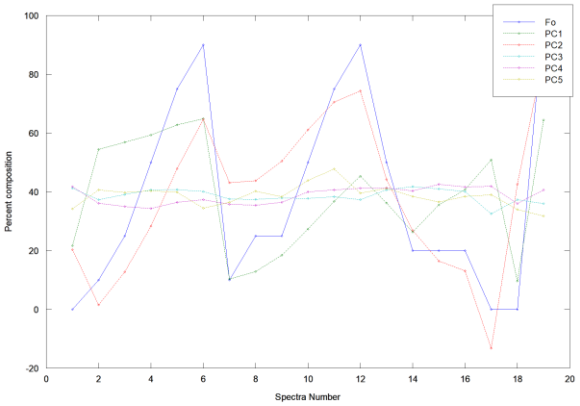

Fig 5: Comparison of the actual forsterite values to those calculated by each linear fit. As seen in the figure, PC1 and PC2 are both the closest, but neither are as close as the fits in the other two images.

the remaining end-member's percentage can easily be calculated simply by subtracting all calculated percentages from 100. This works in a controlled case where there are no other end-members in the mixture but will not work if unknown minerals are present such as in a non-laboratory experiment.

Result validation

Initial inputs: Once the principle components are correlated and results calculated using the fits and supplementation, the original labels are graphed against the calculated values for result validation, as shown in figures 7-8.

Training and testing sets: For further validation, the data can be split into a training and a testing set. This was difficult to do with the tertiary mixture set due to only having 19 available spectra but could easily be done with the four end-member mixture set since it contained 44 different spectra.

The validation procedure essentially removes a portion of the spectra (testing set) from the data set (now training set) during the principle component analysis. Principle component analysis is run using the training set as input where principle components are automatically assigned to values and fits automatically generated. The resulting projection matrix and linear fits are used to determine values in the testing set, which can be analyzed using an R^2 value.

For the tertiary mixture set, 7 spectra were set aside to form the testing set, leaving the other 12 as the training set. Due to the difference in the spectra in the testing set, different features were made more prominent causing the algorithm to automatically correlate PC1 as nontronite and PC2 as forsterite; this leaves the magnesite percentages to be calculated from the other two end-members. Testing the remaining spectrum gave the results shown in figure 9, with an R^2 score of 0.91 for magnesite, 0.80 for nontronite, and 0.75 for forsterite.

The four end-member mixture set was much larger, allowing 17 spectra to be set aside as the testing set while still having 27 spectra in the training set. This lead to PC1 as LCP, PC2 as HCP, PC3 as PLG, and OL being calculated from the other three percentages. Testing the spectrum gave the results shown in figure 10, with an R^2 score of 0.94 for OL, 0.88 for PLG, 0.91 for LCP, and 0.80 for HCP. This correlation is in general closer than the tertiary set.

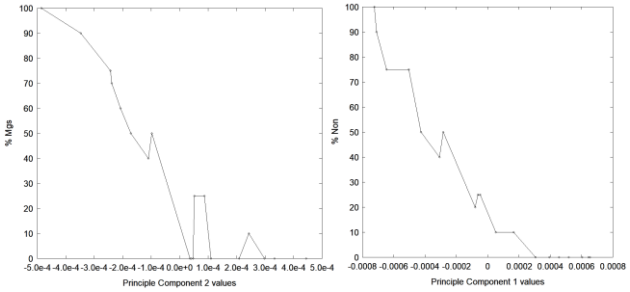

Fig. 6: Scatter point plots comparing the principle component values to the original percentages. Compares percent magnesite to PC2 (left) and percent nontronite to PC1 (right). Values are not fully linearly correlated as they were in the previous study of binary mixtures.

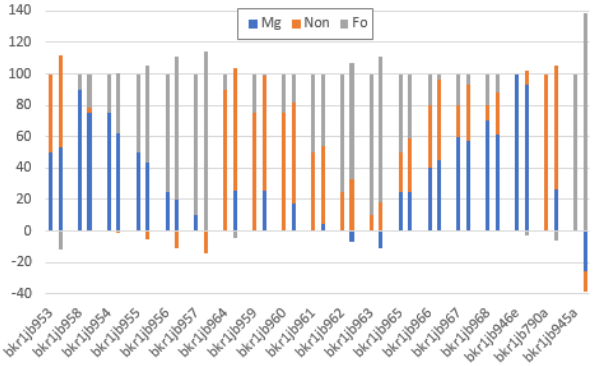

Fig. 7: Results of principle component decomposition for the tertiary mixture set. Actual values are shown with a dashed line and calculated with a solid line.

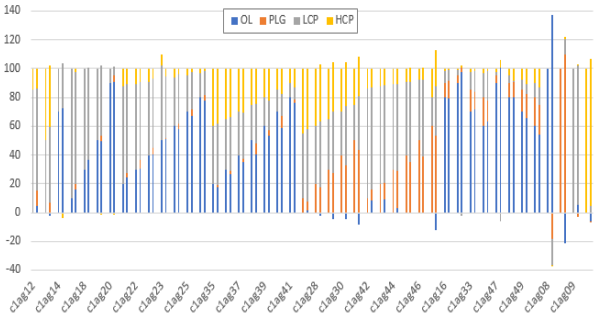

Fig. 8: Results of principle component decomposition for the four end-member mixture set. Actual values are shown with a dashed line and calculated with a solid line.

Discussion

The techniques from this study will be essential in later analysis of hyperspectral images. Since the valid range of wavelengths along with the spectrum relevant to an image may vary, it is essential to be able to be able to automatically assign principle components rather than needing to manually create each projection matrix and linear fits.

Similarly, since this works with both a tertiary and a four end-member mixture, it is likely the algorithm can be expanded to analyze a mixture of any number of end-members provided enough unique spectrum are provided. This is essential to be able to analyze non-laboratory spectrum, such as CRISM or OMEGA hyperspectral images, as they rarely will contain just two components.

There does seem to be a small breakdown of correspondence between the values. In the previous study, values tended to hold a strong linear relationship with few unexplained deviations [2]. As shown in figure 6, the principle component values are not fully linear, leading to the occasional dip in the percent to principle component value graph. The issue was especially common when the mixture had 0 percent of the element in question. The issue can be partly attributed to similarities between different spectrum classes, which could possibly be avoided by using a larger training set or more varied spectrum classes. The issue did not appear to be more common in the four end-member mixture than in the tertiary mixture, which leads to the assumption that it will not significantly impact results provided enough spectrum are provided in the training set.

Summary and conclusion

Expanding the technique of principle component analysis to mixtures of more than two end-members, including a tertiary mixture of magnesite, nontronite, and forsterite; and a four end-member mixture of olivine, plagioclase, hypersthene, and augite. Principle components are automatically correlated to end-member percentages allowing minimal user involvement. For both studies, results show that a linear relationship still gives reliable results for mixtures containing more than two end-members, and that principle component analysis produces reliable results on mixtures containing carbonates and olivine in addition to the previously found pyroxene [4] and phyllosilicates [2].

Future work will involve applying this technique to less closely matched mixture sets and to remotely sensed data. It would also be ideal to test with a method that works with a smaller number of spectra such as bootstrapping.

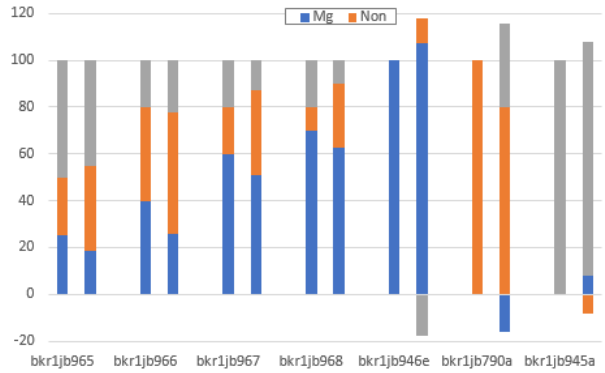

Fig 9: Results of validating the principle component algorithm on seven members of the tertiary mixture set.

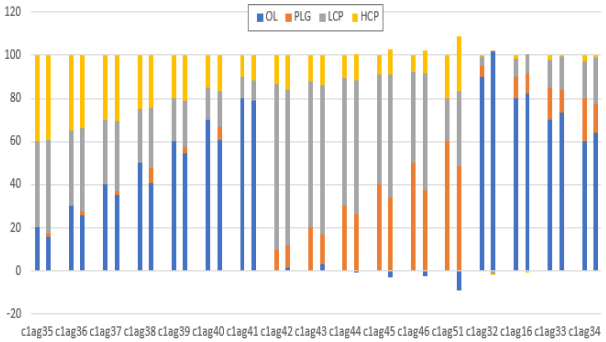

Fig 10: Results of validating the principle component algorithm on 17 members of the four end-member mixture set.

References

[1] J. L. Bishop *et al.*, "Coordinated spectral and XRD analyses of magnesite-nontronite-forsterite mixtures and implications for carbonates on Mars," *J. Geophys. Res. Planets*, vol. 118, no. 4, pp. 635–650.

[2] J. S. Makarewicz, H. D. Makarewicz, and J. L. Bishop, "Spectral Mixture Modeling Using Principle Component Analysis Applied to Nontronite-Ferrihydrite and Kaolinite-Montmorillonite Mixtures," presented at the Lunar and Planetary Science Conference, 2018, vol. 49, p. 1378.

[3] "RELAB disclaimer." [Online]. Available: http://www.planetary.brown.edu/relabdocs/relab_disclaimer.htm. [Accessed: 03-Jul-2018].

[4] J. S. Makarewicz and H. D. Makarewicz, "Spectral mixture decomposition using principal component analysis applied to pyroxene mixtures," in *2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2013, pp. 1–4.