A Novel and Effective Multivariate Method for Compositional Analysis using Laser Induced Breakdown Spectroscopy

W Wang\(^1\), B Ayhan\(^2\), C Kwan\(^2\), H Qi\(^1\), S Vance\(^3\)

\(^1\)University of Tennessee, Knoxville, TN, US.
\(^2\)Signal Processing Inc., Rockville, MD, US.
\(^3\)Jet Propulsion Laboratory, Pasadena, CA, US.

E-mail: wwang34@utk.edu, bulent.ayhan@signalpro.net, chiman.kwan@signalpro.net, hqi@utk.edu, Steven.D.Vance@jpl.nasa.gov

Abstract. Compositional analysis is important to interrogate spectral samples for direct analysis of materials in agriculture, environment and archaeology, etc. In this paper, multivariate analysis (MVA) techniques are coupled with laser induced breakdown spectroscopy (LIBS) to estimate quantitative elemental compositions and determine the type of the sample. In particular, we present a new multivariate analysis method for composition analysis, referred to as "spectral unmixing". The LIBS spectrum of a testing sample is considered as a linear mixture with more than one constituent signatures that correspond to various chemical elements. The signature library is derived from regression analysis using training samples or is manually set up with the information from an elemental LIBS spectral database. A calibration step is used to make all the signatures in library to be homogeneous with the testing sample so as to avoid inhomogeneous signatures that might be caused by different sampling conditions. To demonstrate the feasibility of the proposed method, we compare it with the traditional partial least squares (PLS) method and the univariate method using a standard soil data set with elemental concentration measured \textit{a priori}. The experimental results show that the proposed method holds great potential for reliable and effective elemental concentration estimation.

1. Introduction

The techniques for detecting essential elements in materials are of key importance for direct analysis of materials in diverse scenarios, such as agriculture, environment, archaeology, and even space exploration science [1]. In the last decades, numerous analytical techniques have been extensively explored for the compositional analysis. However, most techniques are destructive and a significant amount of the samples are consumed. Furthermore, many of these techniques are expensive and time-demanding due to the sample preparation. Laser induced breakdown spectroscopy (LIBS) is a fast elemental analytical technique that has potential advantages over conventional techniques. The LIBS spectrum can provide a detailed fingerprint of the material analyzed. However, quantitative analysis with LIBS is complicated because of the chemical matrix effects [2], which are directly related to the elemental and molecular composition of the sample and ubiquitously perturbs the LIBS plasma. The existence of matrix effects may also cause the elements in sample to interfere with each other and overshadow the

This work was supported in part by NASA NNX12CB05C.
elements of interest. Many approaches have been developed to compensate for the chemical matrix effects to extract more accurate quantitative elemental concentrations from LIBS spectra [3][4].

A promising method to resolve the challenges is to use multivariate analysis (MVA). As mentioned before, LIBS spectra tend to be very complex and contain much information that might be ignored by univariate techniques. Many studies have applied MVA on LIBS to extract quantitative elemental compositions and determine the type of the sample. Sirven et al. were among the first to apply MVA to LIBS spectra of geological samples, investigating the accuracies of Cr analysis in 30 soil and 30 Kaolinite samples [5]. Their study compared partial least squares (PLS) regression with predictions from neural network analysis (ANN). Martin et al. used PLS to create a calibration model and extract quantitative concentrations of elements used in wood preservatives, including Cu, Zn and As [6]. Gabriela et al. employed a combination of principal component analysis (PCA), ANN and linear discriminant analysis (LDA) to analyze the feasibility for fast identification and classification of mineralized tissues and biomimetal structures among other materials based on LIBS spectra [7]. Some other analysis techniques have also been explored in [8][9].

It is a common challenge for most of the MVA techniques to perform accurate elemental composition analysis which greatly depends on the availability of a comprehensive training set. As for PLS, other than an appropriate number of training samples, parameters like the number of latent variables also need to be properly defined. In this paper, we propose a new elemental compositional analysis method by using the "spectral unmixing technique". We consider the LIBS spectrum from a sample of completely unknown composition, \( x \) (a 1-D signal), as a linear mixture of a limited number of elemental signatures, \( S \) (each column characterizes the profile of an elemental LIBS spectrum). \( S \) includes all possible element signatures that might be contained in the sample under analysis. The formulation can thus be expressed as \( x = S\alpha \) with \( \alpha \) being the elemental concentration abundances. Since the plasma evolution is sensitive to pressure, it is almost impossible to develop a comprehensive database containing all elemental signatures that can be applied to analysis for various scenarios. As a result, we utilize a simple but effective calibration step using one or several samples with elemental composition known \textit{a priori}. The signature library \( S \) is obtained from training samples or standard elemental LIBS spectrum database and the proposed spectral unmixing method achieves comparable or even better performance to the popular PLS method in compositional analysis, but with much easier mathematical formulation, less sensitive to the predefined parameters (e.g., the number of latent variables defined in PLS), and better physical interpretability.

2. Concentration estimation technique

Mixture measurements are frequently encountered in real-world applications. The existence of mixture measurements has brought the decomposition technique to find use in a wide array of applications. For example, in remote sensing, unmixing has been applied to subpixel quantification, mineral identification, etc. [10]. In biological microscopy, the multispectral fluorescence is analyzed to discriminate different co-localized fluorescent molecules with highly over-lapping spectra [11].

Similar to the ubiquitous existence of mixture measurements, LIBS spectra also rarely occur in a pure and isolated fashion. Because of the diverse elements in a sample, characteristic atomic emission lines of the elements will interfere with each other and the measured spectrum would more than likely be a mixture of pure atomic spectrum of the elements that are contained in the sample. Under the above physical assumption, linear unmixing technique has been utilized for compositional analysis due to its effectiveness and simplicity, where the LIBS spectrum measured from sample is given by

\[
x = S\alpha + \varepsilon
\]

where \( x \in \mathbb{R}^l \) is the measured observation vector, \( S \in \mathbb{R}^{c \times l} \) is the elemental signature library whose columns, \( \{s_j\}_{j=1}^c \in \mathbb{R}^l \), correspond to the different characteristic spectrum of pure element, and \( c \) is the number of elemental profiles contained in the library. The coefficient vector is denoted by \( \alpha \in \mathbb{R}^c \). The possible errors and noises are taken into account by the \( l \)-dimensional column vector \( \varepsilon \). Notice that although there is one parameter \( c \) decided by the number of pure elemental signatures that can be
obtained, this parameter does not have much effect on the unmixing results. Once we add one more elemental signature $\{s_j\}_{j=1}^k$ into the library $S$, the corresponding coefficient in $\alpha$ would likely be zero if the sample does not contain the newly added element. Therefore, it is most desirable to include all possible elemental LIBS spectra into the predefined library $S$. Based on this linear model, we are aiming to estimate the elemental concentration in the test samples based on the criteria that the larger values in the unmixed coefficient vector $\alpha$, correspond to the higher concentrations of the corresponding elements in the test sample. One more constraint should be noticed is that since the different elemental atomic responses excited by the high energetic laser pulse should be additive when recorded by the spectrometer, all the entries in the coefficient vector $\alpha$ should be positive, which result in the non-negativity constrained least square (NCLS) linear model as below [12]:

$$x = S\alpha + e, \quad \text{s.t.} \quad \alpha_{j \in x} \geq 0$$

The NCLS aims to estimate $\alpha$ (concentration estimate) from $x$ (LIBS spectrum of test sample) and $S$ (elemental signatures library) and also guarantees that the estimated abundance fractions are all non-negative because of the constraint. On the other hand, since it does not contain a sum-to-one constraint, the summation of the estimated fractions is not necessarily equal to one.

3. Signature library and calibration

To construct a comprehensive elemental signature library $S$, two options are provided. First, we can automatically extract a signature library $S$ by regression if we hold a group of training samples $X = [x_1, x_2, \cdots, x_k]$ with known ground truth of the elemental composition matrix $A = [a_1, a_2, \cdots, a_k]$. Then the estimated $\tilde{S}$ is given by using a pseudo-inverse operation as $\tilde{S} = X \cdot \text{pinv}(A)$. Second, we also can resort to the standard database of LIBS atomic spectroscopy, like the one provided by the National Institute of Standards and Technology (NIST) [13]. However, it only provides information about the characteristic responses of chemical elements at some specific wavelengths. Thus we have to build a signature library including every chemical element of interest based on this incomplete information.

Realizing that the atomic response intensity recorded on LIBS spectrum always varies because of environmental factors, like pressure, temperature, etc., [1] not only the elemental signatures might not be consistent to the sampled LIBS spectrum of the testing sample, but that all of the elemental signatures in the library might not be homogeneous since they might be measured under different environments. We thus utilize a simple calibration process that using one or several standard samples $x_{\text{calib}}$ that the elemental composition $\alpha_{\text{calib}}$ is known a priori. With the extracted signature library $S$, a normalization step is performed firstly, that the spectral magnitude is rescaled such that the peak value of the elemental signature becomes 1. This normalization has been repeated for each of the extracted signatures in $S$. Then, we estimate the abundance vectors $\alpha$ for the calibration samples $x_{\text{calib}}$ based on the aforementioned NCLS algorithm. After that, the rescaling mapping matrix $T$ for calibration can be estimated with the estimated $\alpha_{\text{estimate}}$ and the ground truth $\alpha_{\text{calib}}$ that $T_{\alpha_{\text{calib}}} \cdot \alpha_{\text{calib}} = \alpha_{\text{calib}}$.

Then the signature library can be calibrated by $T$ as $\tilde{S} = S \cdot T$. Notice that this calibration only needs to be used on the manually extracted signatures from standard database, since the signatures automatically extracted by training samples are naturally calibrated in the regression process. In this paper, the NCLS algorithm and the Calibration process are combined to yield the proposed algorithm, referred to as non-negativity constrained least square with calibration (NCLSC).

4. Experiments

4.1. Test data and performance measures

In order to evaluate the performance of the proposed NCLSC, a standard LIBS spectral data set that consists of 12 soil samples with different elemental Cu and Zn concentrations is used [14]. The ground truth concentration in the 12 soil samples were determined with inductively coupled plasma-optical
emission spectroscopy (ICP-OES), shown in the Table 1. In the following experiments, we compare the concentration estimation results of the NCLSC to the traditional PLS method.

Two performance measures have been used in the analysis. One is the normalized root mean square error of prediction (NRMSEP) and the other is the correlation coefficient \( \gamma^2 \). A model with high \( \gamma^2 \) and low NRMSEP is the preferred one [14]. The NRMSEP measures the relative average prediction error. Suppose \( \alpha_i \) is the predicted concentration and \( \pi_i \) is the actual concentration of sample \( i \) where there is a total of \( n \) samples. Then, the NRMSEP is expressed as below, where \( \pi_{\text{max}} \) and \( \pi_{\text{min}} \) is the maximum and minimum concentration in samples, respectively.

\[
NRMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\alpha_i - \pi_i)^2}{n}} \left( \frac{\pi_{\text{max}} - \pi_{\text{min}}}{100}\right)
\]

The correlation coefficients measure \( \gamma^2 \) is defined as below:

\[
\gamma^2 = 1 - \frac{\sum_{i=1}^{n} (\pi_i - \frac{\sum_{i=1}^{n} \pi_i}{n})^2}{\sum_{i=1}^{n} (\pi_i - \alpha_i)^2}
\]

4.2. Experiment analysis

Cross validation is used in experiments for both of the PLS and the proposed NCLSC. In PLS, the number of latent variables is optimally set to 3 and the model is built by the information extracted from 11 of the 12 soil samples and the one left out is used as the test sample to estimate the Cu and Zn concentrations. Similarly in the NCLSC, in the first test, referred to as NCLSC-ASL (automatic signature library), the library \( \bar{S} \) is automatically estimated from training samples, then \( \bar{S} \) is used to estimate the abundance of the left out one. In the second test, referred to as NCLSC-MSL (manual signature library), the signature library \( \bar{S} \) is manually extracted from the standard element LIBS database and calibrated by two samples. All of the other ten soil samples are used as test sample for validation.

Although the LIBS spectra provide us enough channels for analysis, many channels in the data are not useful for prediction, and even may degrade the accuracy of the results. Thus feature selection can be used to isolate meaningful regions of the spectra before applying multivariate analyses. Manual feature selection has been employed with some success by [15]. In [14], two frequency ranges have been considered for analysis. In addition, we have searched for other frequency regions to obtain better estimation accuracy. From Table 2, it can be seen that the searched frequency range 315-340nm yields better estimation accuracy.

| Metal | Spectral Range nm | PC | \( \gamma^2 \) | NRMSEP % |
|-------|-------------------|----|-------------|----------|
| Cu    | (200-600)         | 3  | 0.90        | 10.44    |
|       | (300-350)         | 3  | 0.94        | 7.90     |
|       | (315-340)         | 3  | 0.96        | 6.34     |
| Zn    | (200-600)         | 3  | 0.90        | 10.62    |
|       | (300-350)         | 3  | 0.94        | 8.14     |
|       | (315-340)         | 3  | 0.95        | 6.53     |

Table 3 shows the performance of frequency ranges using the NCLSC method with signature library \( \bar{S} \) estimated automatically from prior known elemental compositions and the LIBS spectra of training samples. It also shows that the selected frequency range 315-340 nm provides more accurate estimation. More importantly, when Table 3 is compared to Table 2, it is obvious that the NCLSC-ASL provides more accurate predictions on all three frequency ranges. More comparison results between the PLS and the NCLSC-ASL for the selected frequency range of 315-340 nm are shown in

| Metal | Spectral Range nm | \( \gamma^2 \) | NRMSEP % |
|-------|-------------------|-------------|----------|
| Cu    | (200-600)         | 0.9206      | 6.33     |
|       | (300-350)         | 0.9412      | 7.43     |
|       | (315-340)         | 0.9685      | 5.43     |
Figure 1, from which, we can find that the NCLSC-ASL method achieves better performance than the PLS method. Furthermore, the NCLSC methods do not require defining any parameter, like the number of latent variables that depends on the major elements in the testing sample. The results from this study indicate that NCLSC-ASL holds reliable quantitative analytical ability of LIBS and has great potential for constructing robust calibration model for prediction of elemental concentration.

Figure 1. Comparison of concentration prediction accuracy between PLS (top) and NCLSC (bottom, automatically extracted $\bar{S}$).

Figure 2. Concentration estimation based on manually extracted signature library $\bar{S}$.

Similar to using training samples to build the PLS model, using samples with elemental concentrations known a priori to build the signature library $\bar{S}$ might be a limitation for the NCLSC methods. Thus, the manual construction of the signature library by using the standard elemental LIBS spectrum database might be an alternative. The database NIST provides the intensity of elemental atomic responses at only some specific wavelengths on the LIBS spectrum for each element, and a lot of the atomic responses on other wavelengths are not available. To form a complete spectrum, we assume the unavailable responses on the other wavelengths are zero. To make the signature library more diverse, we include all 13 major and trace metal elemental signatures and the calibration is done by two randomly chosen samples. Figure 2 shows the concentration estimation results based on the calibrated $\bar{S}$. The quantitative evaluation results are shown in Table 4, from which we can find that the performance of the NCLSC-MSL is not as good as the NCLSC-ASL. This might be caused by the fact that the elemental signatures that we manually extracted in $\bar{S}$ are not accurate enough because of the incomplete elemental signature information, resulting in a bit worse concentration estimation. If we can obtain a more accurate database with complete elemental signatures, the manually built $\bar{S}$ with calibration might be able to give much more accurate concentration estimates. Table 4 compares the NCLSC-MSL results with the univariate model. The results demonstrate the advantages of NCLSC-MSL over
the univariate model, though with some inferiority when compared to NCLSC-ASL. Therefore, we conclude that the pure elemental signatures play a very important role in elemental composition analysis. The automatically estimated signature library from training samples performs better than the manually constructed signature library from a standard database with incomplete information, and there is a trade off between the estimation accuracy and the training sample requirement.

Table 1. Cu and Zn concentrations ground truth of 12 soil samples (mg/kg) [14].

| Sample No. | 1 | 2 | 3 | 4 | 5 | 6 |
|------------|---|---|---|---|---|---|
| Cu         | 26| 27| 29| 33| 36| 39|
| Zn         | 66| 72| 83| 100| 117| 135|

| Sample No. | 7 | 8 | 9 | 10 | 11 | 12 |
|------------|---|---|---|----|----|----|
| Cu         | 43| 48| 54| 59 | 70 | 81 |
| Zn         | 152| 180| 209| 238| 295| 352|

Table 4. The analytical results of the NCLSC-MSL with wavelength ranges (200-600 nm)

| Metal | Method       | $\gamma^2$ | NRMSEP% |
|-------|--------------|------------|---------|
| Cu    | NCLSC-ASL    | 0.9228     | 8.52    |
|       | NCLSC-MSL    | 0.8091     | 13.39   |
|       | Univariate M | 0.6213     | 18.81   |
| Zn    | NCLSC-ASL    | 0.9206     | 8.63    |
|       | NCLSC-MSL    | 0.7400     | 15.62   |
|       | Univariate M | 0.4522     | 22.81   |

5. Conclusion
In this study, a novel and high performance multivariate method for compositional analysis based on LIBS spectra was proposed. We utilized a simple but effective linear mixture model to describe the systematics of the physical formation of LIBS spectrum, and used a calibration step based on training samples to obtain standard pure elemental signatures that satisfies the homogeneity with the testing sample. The experiments with an elaborately prepared soil dataset demonstrated that the proposed analysis framework holds great potential for reliable quantitative elemental concentration estimation. Compared with traditional analysis methods, the proposed method further holds advantages in areas such as less sensitive parameter setting requirement and better interpretability.

6. Acknowledgment
We would like to thank Ningfang Yang, Prof. Neal S. Eash and Prof. Jaehoon Lee for providing us the standard soil data set for the experimental validation and also thank Prof. Nicole Labbe and Prof. Madhavi Marin for the insightful discussions regarding on various multivariate analysis techniques.

Reference
[1] Miziolek, Palleschi V, Schechter I 2006 Laser induced breakdown spectroscopy (LIBS): Fundamentals and Applications (NewYork: Cambridge University Press)
[2] Cremers D and Radziemski L 2006 Handbook of Laser-Induced Breakdown Spectroscopy (England: West Sussex)
[3] Thompson J, Wiens R, Barefield J and Vaniman D 2006 J. Geophys. Res. Planets 111 E05
[4] Salle B, Lacour J, Mauchien P, Fichet P and Maurice S 2001 Spectrochim. Acta Part B 61 301
[5] Sirven J, Bousquet B, Canioni L and Sarger L 2006 Anal. Chem. 78 1462
[6] Martin M, Labbe N, Rials T and Wullschleger S 2005 Spectrochim. Acta Part B 60 1179
[7] Vtkova G, Novotny K, Prokes L, Hrdlicka A and Kaiser J 2006 Spectrochim. Acta Part B 6 113
[8] Colao F, Fantoni R, Ortiz P, Vazquez M, Martin J, Ortiz R and Idris N 2010 Spectrochim Acta Part B 65 688
[9] Dyar M, Carmonis M, Breves E, Ozane M and Wiens R 2012 Spectrochimica Acta Part B 70 51
[10] Chang C and Heinz D 2000 IEEE Trans. Geosci. Remote Sensing 38 144
[11] Dickinson M, Bearman G, Tille S and Lansford R 2001 BioImaging 31 1272
[12] Kwan C, Ayhan B, Chen G, Wang J, Ji B and Chang C 2006 IEEE Trans. Geosci. and Remote Sensing 44 409
[13] N. Physical Meas. Laboratory, URL http://physics.nist.gov/PhysRefData/Handbook/Tables.htm
[14] Yang N, Eash N, Lee J, Martin M, Zhang Y, Walker F and Yang J 2010 Soil Science 175 447
[15] Tucker J, Dyar M, Schaefer M, Clegg S and Wiens R 2010 Chem. Geol. 277 137