NEW METHOD OF MINERAL ELEMENT ABUNDANCE IDENTIFICATION BASED ON PCA LOADING SPATIAL DISTANCE. Kaichen Guo[1], Zhongchen Wu[1]*, Li Zhang[2], Zongcheng Ling[1], Yun Li[1], MaoCheng Qian[1], Institute of Space Science, Shandong University, Weihai, 264209, China (z.c.wu@sdu.edu.cn); [2]School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209;

Introduction: LIBS (Laser Induced Breakdown Spectrometer) is a power tool for elemental analysis. ChemCam, i.e., the first LIBS system for Martian exploration, has achieve great success. For Martian surface exploration, the most important scientific goal is to detect the elemental makeup of the surface which will help us to better understand the distribution of the surface minerals/rocks, even the Martian geological evolution. In addition, the identification of the elemental abundance of minerals can effectively reduce the Matrix effect and improve the accuracy of quantitative analysis. In this study, the identification strategy for mineral species and the identification method of Mineral based on the published ChemCam 64 calibration data set[1] were reported. Our result show that PCA (Principle Component Analysis) Loading spatial distance can be used to obtain the element information of minerals before quantitative analysis, and then classify minerals according to elemental abundance with an accuracy up to 92.8% in this case.

Data Description: The data of 64 ChemCam preflight calibration standard samples [1], with both confirmed concentration data and LIBS spectral data, was selected for this study. Those standards are mainly composed of igneous materials, some sediment material (such as sulfates, carbonates, and phyllosilicates), and some pure minerals, which covered major species and abundance ranges of typical chemical composition on the Martian surface.

Method of PCA Loading Spatial Distance: Principal component analysis (PCA) is a multivariate technique which is always used to analyze the matrix composed of several inter-related observation variables [2]. In PCA, the principal components(PCs) are obtained from the singular value decomposition (SVD) [3][4] of the observation matrix **X**. The SVD is a generalization of the eigen-decomposition which can be used to analyze rectangular matrices (the eigen-decomposition is defined only for squared matrixes)[5]. The main idea of the SVD is to decompose a rectangular matrix into three simple matrices: Two orthogonal matrices and one diagonal matrix. For a rectangular observation matrix X (m×n), we do SVD decomposition on it[6]:

$$\boldsymbol{X}=\boldsymbol{U}\cdot\boldsymbol{S}\cdot\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{E}=\boldsymbol{P}\cdot\boldsymbol{T}^{\mathrm{T}}+\boldsymbol{E}$$

In which, U is a column orthogonal matrix, satisfied that

$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}_{\mathrm{m}}$, $\boldsymbol{I}$ is identity matrix.$\boldsymbol{V}^{\mathrm{T}}$ is a row orthogonal matrix, satisfied that

$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}_{\mathrm{n}}$, $\boldsymbol{I}$ is identity matrix.

$\boldsymbol{S}$ is a diagonal matrix of the singular values, $\boldsymbol{P}$ is the scores matrix, $\boldsymbol{T}^{\mathrm{T}}$ is the loadings matrix and the matrix $\boldsymbol{E}$ contains the residuals.

The scores and loadings matrix describe the relationship between PCs and observation variables (LIBS data). The vectors of score matrix span a low dimension mathematical space, usually referred to as a score plot, through which the samples could be projected and viewed. The value of loading is highly related to the correlation between variables and PCs.

Data Process : As shown in the right part of Fig.1, standard samples whose concentration is above the threshold value (i.e., average value of each element in Martian soil) were selected as the **verification set**. As shown in the left part of Fig.1, 1-3 high concentration standards from **verification set** were selected for the **training set.** The LIBS emission line assignment of the training standards set were identified by using the Ocean Optics software MaxLIBS, and about 30 stronger emission spectral lines were selected from 6,144 LIBS channels for the analyzed target element. Thus the data size was compressed. The Euclidean squared distance between the distributed variables(identified feature emission lines) in the loadings space and the spatial origin point (i.e. the sum of squares of loadings values corresponding to different PCs) was calculated, to determine the importance of the variables on PCs.

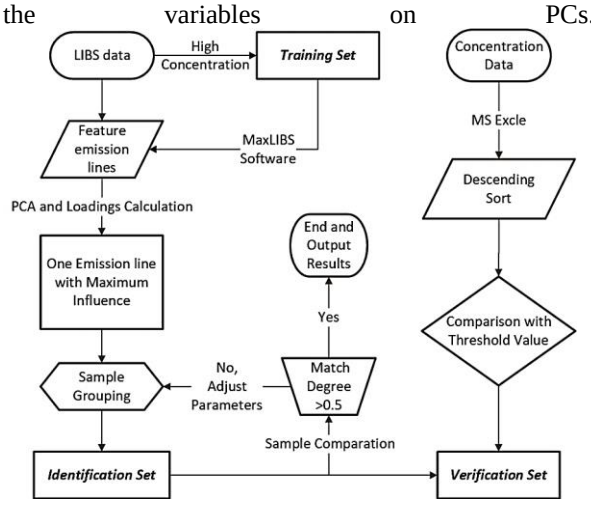

Figure 1. Data processing flow chart plot

Based on this methodology, the critical LIBS emission line which has the greatest contribution to identification of the element will be selected (see next section). Cluster analysis (Sample Grouping, in software Unscramble® X) based on PCA will be reused to select critical emission lines as independent variable to distinguish the samples types without accurate quantitative analysis. The cluster results formed the **identification set**.

By Comparing the final identified samples with the samples in the verification set, the matching degree was calculated which is used to evaluate the recognition ability of elements and output the results.

PCA and Loading Distance Calculation: After emission lines selection of a certain target element using MaxLIBS software, matrix with m (m<30) channels was selected from 6,144 original LIBS channels. The new selected matrix (m×64) was then performed PCA for loading space distance calculation. As a result, only one emission line which has the maximum distance (i.e. the emission line has the greatest influence on sample identification) were selected for subsequent mineral identification. Here, we take the aluminum element as an example to show the loading space distance method, see Table.1, the bolded line is the selected channel. Other elements (i.e. K, Ca, Si, Ti, Na, Mg, Fe, Mn and P) were treated by the same method as above for mineral identification.

Table1. PCA and Loading distance results of elements Al

Wv-Al	PC1	PC2	PC3	**Distance**
624.262	0.006784	0.042292	0.011337	0.001963
308.219	0.175367	-0.24788	0.113079	0.104986
309.26	0.198525	-0.31859	-0.04386	0.142838
281.596	0.023321	0.027046	-0.68219	0.466663
394.398	0.395105	0.571702	0.046776	0.485139
257.487	0.028541	-0.21646	0.683125	0.514328
396.168	0.728368	0.265435	0.093106	0.609644
309.307	***0.491734***	***-0.6248***	***-0.20557***	***0.674439***

Results: Based on the wavelength channel selected through PCA loading distance calculation, the Sample Grouping was performed and the score of PCA was analyzed which corresponds to the calculated distance in the loading space of the analyzed element. Comparison and matching degree of all elements between verification set and identification Samples are shown in Table2.

Discussion: Based on our method, half of the minerals with the major elements in the Martian surface can rightly identified with a matching degree of more than 0.75. And 80% of the elements of minerals have a matching degree better than 50%. The Loading space distance method can basically determine the element information of the samples before the accurate quantitative analysis of the LIBS data. Therefore, the method can be used to assist the identification of substances and the classification of substances to improve the utilization efficiency of LIBS data. Unnikrishnan et.al applied the LIBS technique to identify the four widely used plastics, polyethylene terephthalate (PET), high-density polyethylene (PE), polypropylene (PP) and polystyrene (PS)[7]. Our method may be useful to train and identify the characteristic groups of organic molecules in plastics from LIBS signal directly, so as to improve the classification efficiency. There may be other applications for our approach.

Table2.Comparison and matching condition of all elements

	Boundary (wt%)	**Verification Num**	**Identified Num**	**Match Num**	**Matching Rate**
Al	9.71	43	42	39	0.92
K	0.44	42	40	36	0.90
Ca	6.37	35	28	24	0.85
Si	45.41	39	38	30	0.79
Ti	0.9	22	24	18	0.75
Na	2.73	28	27	17	0.63
Mg	8.35	10	10	5	0.50
Fe	16.73	10	10	5	0.50
Mn	0.33	4	4	0	0
P	0.83	4	4	0	0

Note:
(1) Match Number is the number of samples, which is in the identification set consistent with the samples in the verification set of this element.
(2) Match Rate refers to the ratio of the match number to identified number.

References: [1] Wiens, R. C., et al. (2012). *Space Science Reviews* **170**(1-4): 167-227. [2] Abdi, H. and L. J. Williams (2010). Wiley Interdisciplinary Reviews: *Computational Statistics* **2**(4): 433-459.[3]Gene H. Golub., et al.(1996). *Matrix Computations.* [4]William H. Press., et al. (2002) *Numerical recipes in C.* ISBN 0-521-43108-5 [5] Abdi, H., et al. (2007) Neil Salkind (Ed.). *Encyclopedia of Measurement and Statistics.*[6] Li, B.-Y., et al. (2004). *Analytica Chimica Acta* **514**(1): 69-77. [7] Unnikrishnan, V. K., et al. (2013). *RSC Advances* **3**(48).

Acknowledgement: Sincerely thanks to the ChemCam team for providing 64 standard samples data. **This work was supported by** Natural Science Foundation of China (41573056).