Getting the most out of the SuperCam Raman dataset with unsupervised machine learning: Characterization of mineral signatures and their distribution. E. Clavé[1] (elise.clave@dlr.de), G. Lopez-Reyes[2], O. Beyssac[3], O. Forni[4], A. Ollila[5], S. Schröder[1], K. Rammelkamp[1], J. Aramendia[6], K. Castro[6], J.M. Madariaga[6], J.A. Manrique[2], M. Veneranda[2], C.H. Egerland[1], A. Lomashvili[1], R.C. Wiens[7], S. Maurice[4] and the SuperCam Team, [1]DLR-OS, Berlin, [2]ERICA group, University of Valladolid, [3]IMPMC, Paris, [4]IRAP, Toulouse, [5]LANL, NM, [6]University of Basque Country, [7]Purdue University.

Introduction: Since February 18th, 2021, two Raman instruments have been successfully deployed and operated on the surface of Mars, contributing to the characterization of the rocks in Jezero crater and the completion of the scientific goals of the Mars2020 mission [1-4]. These two instruments, SHERLOC [5] (not discussed further) and SuperCam [6,7], are the first Raman instruments to be used beyond Earth.

The SuperCam instrument enables remote analysis of rocks in a radius of several meters around the rover. SuperCam Raman uses a frequency-doubled Nd:YAG laser, firing 532 nm, 4 ns, 9 mJ pulses on a ~1 cm diameter spot on the targets. The light is collected with a Cassegrain telescope, coupled into a 6 m long optical fiber plugged into a transmission spectrometer equipped with a iCCD camera. The light is collected during a 100 ns window centered on the laser pulse. Raman spectra are generally acquired on 3 to 10 different points on a target, accumulating 100 – 400 laser shots. The spectra processing includes wavelength calibration, despiking, dark removal, IRF correction and denoising.

The dataset: During the first 950 sols of the Mars2020 mission, spectra were acquired on 502 points on ~60 geologic targets. Manual inspection of each spectrum led to the identification of six main independent signals in the dataset (Figure 1):

1) A contribution, informally called *fiber bump*, from the *amorphous silica of the optical fiber* between the telescope and spectrometer, excited by laser light backscattered of the sample's surface, of which only ~96% is filtered before injection into the fiber. It consists of a main band below 500 cm^{-1}, with additional contributions around 600 cm^{-1} and 800 cm^{-1}, and is observed in about half the spectra acquired on Mars.

2) The Raman signature of *olivine*, characterized by a doublet around 820 and 850 cm^{-1}, observed in more than 70 spectra [8].

3) The Raman signature of *Ca-sulfate*, characterized by the main ν_1 Raman mode at ~1017 cm^{-1}, and in particular some spectra of anhydrite type II identified based on multiple distinguishable Raman modes. Ca-sulfates were identified in about 15 points [9].

4) C*ontinuum* signal, observed as an upwards slope in the 600 – 2000 cm^{-1} range, the origin of which is still under discussion, and which was observed in ~10 spectra [10].

5) The Raman signature of *Na-perchlorate* mainly identified by the ν_1 mode at 954 cm^{-1}. One point exhibits this mode clearly (together with weaker modes) while it appears weakly in about 5 other points [2,11].

6) The Raman ν_1 mode of *carbonate* – identified as Fe-Mg carbonate with the other SuperCam techniques – around 1090 cm^{-1}, observed in about 20 spectra [12].

Motivation: In multiple spectra, the contribution from the fiber, high noise, and low signals make mineral identification challenging. We are aiming to decompose the data into the individual signal-bearing spectra with statistical strategies to enable an improved and automated characterization of the corresponding mineral phases.

Identifying the different signals in the dataset with unsupervised machine learning methods: We compare two techniques: independent component analysis (ICA) and non-negative matrix factorization (NMF). ICA is a classical method for blind source separation, that is identifying independent signals mixed linearly in a dataset [13]. We use the FastICA algorithm (sickit learn, Python) [14]. NMF is a dimensionality reduction technique, based on matrix decomposition, using matrix with only positive values [15]. These techniques are both considered to provide more physically-meaningful components than, for example, principal component analysis (PCA), as the former rely on source identification and the latter on variance modelling. For both techniques, we used six components, as this is the expected number of independent signals in the dataset.

First model (Spectral range: 200 – 1400 cm^{-1}; 6 components): We first developed models on the typical Raman range (200 – 1400 cm^{-1}). We observe in Fig. 1A-B that the ICA and NMF components are very similar. Four components can be interpreted based on the loadings for both techniques: the fiber, olivine, Ca-sulfate and continuum signal. The interpretation of the other two components is not straightforward, but they are comparable with both techniques; moreover, the signature of neither perchlorate nor carbonate was identifiable in these components.

We observe that the contribution of the optical fiber is visible not only in the "Fiber" component (1st row in Fig. 1A), but also to some extent in other components. Furthermore, it appears to induce amplified noise below 500 cm^{-1}, where the fiber signal is strongest.

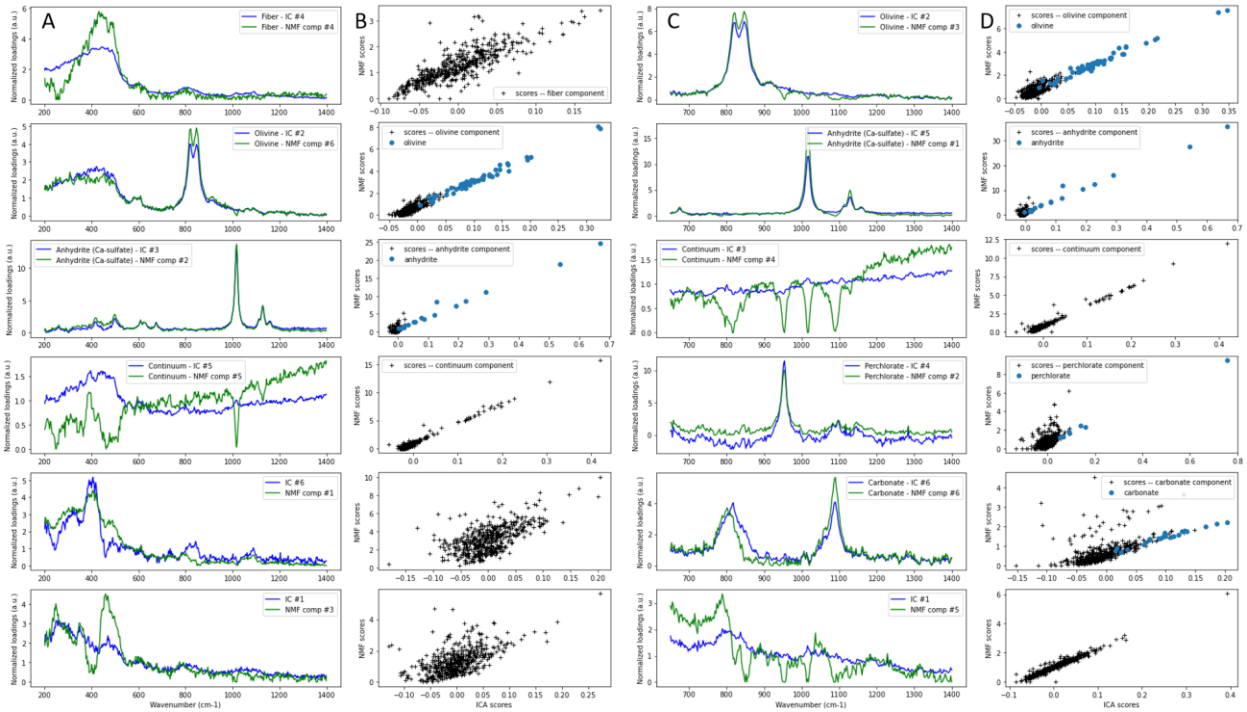

Figure 1 - ICA and NMF models trained (A-B) on the typical Raman range (200 – 1400 cm^{-1}) and (C-D) on a reduced spectral range (650 – 1400 cm^{-1}) using 6 components. The loadings – normalized to the mean – are shown in columns A & C and the scores in columns B & D. The loadings present the characteristic features of our signals of interest: (A) from top to bottom, the fiber bump, olivine, anhydrite, continuum signal; the bottom 2 components don't show clear signatures; (B) from top to bottom, the olivine, anhydrite, continuum signal, perchlorate and carbonate; the component at the bottom doesn't show clear signatures. In columns B & D, we compare the scores of the corresponding ICA and NMF components; the spectra manually identified as bearing the signal of interest are shown in blue.

Second model (Spectral range: 650 – 1400 cm^{-1}; 6 components): We then applied both techniques on a reduced spectral range, to limit the contribution of the fiber bump. In addition to components related to olivine, Ca-sulfate and the continuum signal, we then get components corresponding to perchlorate and carbonate (Fig. 1C-D). Once again, the ICA and NMF results are very similar, and the scores overall correlated. For both perchlorate and carbonate components, we observe points with higher NMF scores compared to the trend, that were not manually identified as either carbonates or perchlorates. No unexpected signature was identified.

Discussion: Overall, the results of ICA and NMF are consistent, identifying most to all expected spectral contributions and features. However, we observe some differences, which we will investigate to determine the best strategy to use these techniques to characterize the Raman spectra acquired with SuperCam on Mars.

Using manually assigned labels, we confirmed that the highest scores for both techniques are overall consistent with the actual mineral signatures for each component. For weaker signatures, the ICA scores are more correlated with our identification. The scores can thus be used to characterize the intensity of a specific feature in the Raman spectra, or look for mineral phase associations. In particular, Ca-sulfates, perchlorates and carbonates were mostly found in distinct points, whereas carbonates and olivine are often detected together. However, for weak signatures, like the carbonates, it is not straightforward to determine the presence of the mineral based solely on the scores; a careful analysis of the spectra is still required (peak position, width, shot to shot data, etc.).

Conclusion: We tested unsupervised machine-learning approaches to characterize the mineral signatures in the Raman dataset acquired on Mars with SuperCam. We show preliminary results highlighting how ICA and NMF can be used to extract the signals present in the dataset, and characterize their distribution.

References: [1] Farley K.A. et al. (2020) *SSR*. [2] Farley K.A. et al. (2022) *Science*. [3] Liu Y. et al. (2022) *Science*. [4] Wiens R.C. et al. (2022) *Sci. Adv*. [5] Bhartia R. et al. (2021) *SSR*. [6] Maurice S. et al. (2021) *SSR*. [7] Wiens R.C. et al. (2021) *SSR*. [8] Beyssac O. et al. (2023) *JGR*. [9] Lopez-Reyes G. et al. (2023) *LPSC*. [10] Clavé E. et al. (2023) *LPSC*. [11] Meslin P.-Y. et al. (2022) *LPSC*. [12] Clavé E. et al. (2024) *LPSC*. [13] Comon (1992) *Signal Processing*. [14] Hyvärinen & Oja (2000) *Neural Networks*. [15] Pauca P. et al. (2006) *Elsevier*