Discovering Compositional Trends in Mars Rock Targets from ChemCam Spectroscopy and Remote Imaging

Diane Oyen, Nina Lanza and Reid Porter
Intelligence and Space Research
Los Alamos National Laboratory
Los Alamos, NM, USA
doyen [at] lanl.gov

Abstract—Onboard the Mars rover “Curiosity”, ChemCam contains two instruments that gather geological data in the form of remote micro images (RMI) for geologic context and laser-induced breakdown spectroscopy (LIBS) for chemical composition. By analyzing the geochemical compositional depth trends of rocks, surface layers are identified that provide clues to the past atmospheric and aqueous conditions of the planet. LIBS produces the necessary data of chemical depth profiles with successive laser shots. To quickly identify these surface layers, we fit a Gaussian graphical model (GGM) to LIBS depth profiles on rock targets. The learned GGM is a visual representation of conditional dependencies among the set of shots making for faster identification of targets with interesting depth trends that warrant more in-depth analysis by experts. We show that our learned GGMs reveal information about the compositional trends present in rock targets that match observations made in more focused studies on these same targets. RMI images provide complementary details about the rock surface. Using RMI and LIBS features, we can cluster similar rock targets by the properties of the rock’s surface texture and depth profile. We present results that show our machine learning methods can help analyze both the breadth and depth of data collected by ChemCam.

I. INTRODUCTION

The Mars Science Laboratory (MSL), or Curiosity, rover landed at Gale Crater on Mars in 2012 and has been collecting a variety of data ever since. Onboard Curiosity, ChemCam’s two instruments gather geological data in the form of remote micro images (RMI) for geologic context and laser-induced breakdown spectroscopy (LIBS) for chemical composition. With over 300,000 spectra already collected, there is more information than can be analyzed manually, offering a challenging opportunity for data scientists to help answer open-ended scientific questions from a rich data source. As ChemCam data continues to be collected from the martian surface during Mars Science Laboratory rover operations, automated methods for analyzing both the depth and breadth of information are needed.

The primary goal of the rover is to determine if Gale crater has been host to a habitable environment. The presence of liquid water is an important component of habitability, as it is currently understood. On Earth, rocks are broken down and altered by interaction with the environment and in particular by interaction with water, leaving distinctive changes in chemistry that provide clues about the nature and duration of aqueous alteration. Materials may be either added (as a coating) or removed (leaving a weathering rind) during the weathering process. In this way rock surfaces record information about their environments over time [1]. In particular, a rock surface records details about the amount and types of fluids with which it has interacted, including the pH and Eh of fluids. Thus, the composition and thickness of coatings and rinds can provide details about both the climate and the presence and abundance of water in the surface environment; e.g. [2].

The depth profiling capability of ChemCam provides information about the presence of chemical gradients in the near surface of rock and soil targets; thus ChemCam can directly observe surface features such as coatings, weathering rinds, and thin stratigraphic layers should they be present [3]. Given the high number of ChemCam observations to date and the high dimensionality of laser-induced breakdown spectroscopy spectral data (~6000 channels) [4], it is not straightforward to evaluate depth trends in composition. We use probabilistic graphical models to identify the relationships between LIBS spectra collected by Curiosity’s ChemCam in Gale crater, and find that the graphical models represent depth trends well. Other statistical methods typically treat each observation as independent from other observations for simplicity. However, such a simplifying assumption discards valuable information about the structure of rocks.

Graphical models provide a visual representation of depth trends. Based on these results, we identify features of interest for characterizing the geology of rocks. We further augment the features extracted from the LIBS data with features extracted from the remote microimager (RMI) data. The result is a collection of features from LIBS and RMI data that are constructed with expert feedback to help quickly characterize rock targets of interest.
II. BACKGROUND

By collecting a sequence of shots performed in a single analysis location, LIBS provides a depth profile of composition that increases in depth with every subsequent shot. Standard statistical analyses treat each of these shots as independent samples. However, the sequence of shots measured by ChemCam on a target are not independent samples, rather they are dependent samples because they are sampling related parts of the rock target. In machine learning we can address these dependencies using probabilistic graphical models that identify dependency relationships among the shots. In a graphical model, the objects are represented as vertices in a graph (or network) with dependencies among objects drawn as edges connecting vertices. Graphical models have been used extensively in bioinformatics applications to discover patterns of interest in protein-protein interactions, gene networks and identifying phylogenetic trees of mutation in pathogens. The questions that they help to answer are about the interrelatedness of a collection of objects. Thus this technique is extremely well suited for answering questions about chemical depth trends.

This section provides necessary background on ChemCam data collection and probabilistic graphical models.

A. ChemCam Data Collection

ChemCam targets are selected on a daily basis by a ground team of scientists. Targets include calibration samples, rocks, soils, drill tailings, and other samples. We focus on rock targets in this paper. As Figure 1 illustrates, for each target a typical ChemCam sequence first takes an RMI of the target, then fires several LIBS shots in one location, moves to other locations in a raster pattern to fire several more LIBS shots in each location, and finishes by taking a second RMI in the ending location. Typical raster patterns include the linear raster, as shown in Figure 1, and a grid raster. The data for each target includes two RMI images and a table (wavelength channels × shots) of the spectral response for each location. Pre-processed RMI images are grayscale TIFF images with 1024 × 1024 pixels. The first RMI is taken before LIBS ablation and is centered on the first location of the raster. The second RMI is taken after LIBS ablation and is centered on the last location of the raster. The LIBS spectrometer collects light from plasma over wavelengths of ~200 – 800 nm including UV, VIS and NIR. Wavelength bands are discretized into ~6000 channels.

During normal operations, ChemCam typically obtains 30 shots on a single sampling location, and often multiple sampling locations per target. Each shot ablates ~0.3 – 0.5 µm of material [3], [4]. We analyze all sampling locations on rock targets in the first 587 sols of the mission. We remove data for wavelengths above 850 nm, set all negative values to zero, and normalize the values for each of ChemCam’s three component spectrometers separately by total emission.

B. Gaussian Graphical Models

Using ChemCam LIBS data, each LIBS shot can be treated as an object; then depth trends are modeled as shot-to-shot dependencies [5]. Gaussian graphical models (GGM) [6] provide a visual representation of conditional dependency relationships among a set of objects contained within a data set. The graphical model can be drawn as a graph \( G = (V, E) \) consisting of two sets: the vertices \( V \) and edges \( E \). The vertices represent the objects or variables while the edges represent conditional dependencies among the objects.

The sequence of shots measured by ChemCam on a target are not independent samples of rock, rather they are dependent samples. A correlation analysis would likely reveal correlations among all pairs of shots, yet to discover the structure of dependencies in the data, we need to analyze the partial correlations. A partial correlation between shot \( X \) and shot \( Y \) is the residual correlation after accounting for all other shots. Thus, a partial correlation is an estimate of a direct dependency. If the partial correlation between \( X \) and \( Y \) is 0 then \( X \) and \( Y \) are conditionally independent. When calculating partial correlations from sample data, the value is nearly guaranteed not to be exactly 0, even if there is no true direct dependence between variables, because of sample noise.

![ChemCam LIBS instrument fires a laser at a target, creating a plasma. Depending on its chemical composition, the plasma gives off different wavelengths of light which are measured by the spectrometer, producing an observed spectrum.](image)
Therefore, to estimate a graphical model, a sparsity constraint limits the number of non-zero partial correlations.

A GGM [6] is estimated from a data matrix $X$, where each column $X_j$ is a shot $j$ with spectral values $X_{ij}$ for $i \in \{1, \ldots, n\}$ wavelengths. The sample covariance matrix, $\Sigma$, is calculated from $X$,

$$
\Sigma_{jk} = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)
$$

where $\bar{X}_j$ is the mean of column $j$,

$$
\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}
$$

Then, the best sparse approximation, $\Theta$, to the partial correlation matrix for a given sparsity constraint, $\lambda$, is estimated by maximizing the following objective function:

$$
\hat{\Theta} = \arg \max_{\Theta > 0} \log \det \Theta - \text{tr}(\Sigma^{-1} \Theta) - \lambda \| \Theta \|_1
$$

where $\| \Theta \|_1$ is the L1 norm of matrix $\Theta$,

$$
\| \Theta \|_1 = \sum_{ij} |\Theta_{ij}|
$$

The number of non-zero partial correlations is controlled by the value of $\lambda$, which can be any non-negative real number. If $\lambda = 0$, then all values in $\Theta$ are non-zero. As $\lambda$ increases, the more zeros that $\Theta$ contains.

The likelihood that an edge represents a true direct dependency, or the robustness of the edge, is quantified by estimating several GGMs from the same set of data with different $\lambda$ values of $\lambda$. Edges that appear across more values of $\lambda$ are weighted more heavily, a method known as Bayesian model averaging [7].

A graph $G = (V, E)$ is created based on the learned $\hat{\Theta}$, such that an edge $e_{ij} \in E$ exists between vertices $X_i$ and $X_j \in V$ iff $\hat{\Theta}_{ij} \neq 0$. In all figures of the learned graphs, vertices represent shots, numbered 1 through 30 (or however many shots exist at the location). Vertices are colored by shot number, starting with dark red at shot 1 and getting lighter to light yellow at shot 30 to aid visualization. The robustness of the direct dependency is represented by the thickness of the edge, with thicker edges being more robust (edge is learned for more values of $\lambda$). Each graph is drawn using a force-directed layout algorithm [8].

III. GRAPHICAL MODELS OF DEPTH TRENDS

We present the graphical models learned from several martian rock targets and discuss in detail how features in the learned graphs represent geochemical depth trends of these rocks. Depth trends are validated by comparing to previously published analysis of the geochemistry of these same targets.

For each location, the data includes spectra from a sequence of several shots in a single location. To analyze the structure of the rock target at this location, we estimate a GGM from the spectral data. Each shot is represented as a vertex in the graphical model. Correlations among vertices are calculated from the spectral data, as described in the previous section. Edges in the graphical model represent the estimated direct dependencies (or partial correlations). If two vertices are not connected by an edge in the graph, then the two vertices are conditionally independent of each other given the other vertices. This means that they may be correlated, but the correlation is indirect and depends on one or more other vertices in the graph.

The graphical models learned and presented in Figures 4–8 demonstrate a few different types of depth trends that are readily apparent in the structure of the graphical model. Each type of graph structure can be labeled as a chain, cluster, or blob; and they correspond to interesting compositional depth trends such as surface layers, stratigraphic layers and homogenous composition.

A. Results from Rock Target “Bell Island”

We first examine the rock target Bell Island from Sol 113 (Figure 3). Figure 4 shows the graph estimated for sampling location 1 on Bell Island (as indicated in Figure 3). It shows that the first several shots are directly dependent on the prior shot and the following shot in the sequence, creating a chain from Shot 1 through Shot 10. This means that, for example, if Shots 4 and 6 are correlated, all of that correlation can be explained by the correlation between Shots 4 and 5 combined with the correlation between Shots 5 and 6. Shots 4 and 6 are conditionally independent given Shot 5. Shots 10 through 30 do not demonstrate the same type of chain structure, indicating that the changes in spectra from shot to shot are less distinct.

The GGM estimated from a sequence of shots on a single location indicate that there are shot-to-shot dependencies that follow expected patterns. Without any prior information about the shot sequence, the GGM is able to reconstruct the shot sequence on the first 10 or so shots, which is consistent with observations that these first shots may differ from the underlying rock. In this case, the surface of Bell Island is extremely dusty, making dust a likely contributor to the spectra of the initial shots [9]. Additionally, the first sampling location of Bell Island appears to be a thin, sulfate-rich vein [10]. Our results suggest that in its 30 shot analysis at this location, ChemCam sampled through surface dust into the sulfate vein below. Examination of the spectral data confirm that the elements enriched in dust (for example, magnesium [9]) decrease with depth (shot number) while sulphur and calcium increase.

B. Results from Depth Study on Target “McGrath”

The target McGrath (sol 234, ccam02234) consists of raised, linear features that appear more resistant to erosion than the surrounding rock (Figure 5a). These raised features also appear to contain several layers of material within them (Figure 5b). A number of ChemCam analyses were performed on this target, including both 30- and 150-shot depth profiles. Layers of both high and low magnesium were observed that showed trends with depth [11]. These trends were ascertained by direct examination of the spectra. Based on these analyses, these
features have been interpreted as fracture-filling cements that record variability in the depositing fluid composition [11].

Our GGM analysis of the McGrath target data clearly shows that the composition of the target changes with depth (Figures 6–8). For each location, the data includes spectra from a sequence of 150 shots at that single location.

Figure 6 shows the graph estimated for sampling location 1 on McGrath-5 on sol 234, (Figure 5, ccam02234). In this learned graph the vertices are connected mostly in sequential order even though the algorithm is not given any information about the ordering of the shots. It shows that the most shots are directly dependent on the prior shot and the following shot in the sequence. This indicates that sequential shots are directly dependent on each other, as we would expect. The more interesting feature of this graph is the curve formed near the top. Shots 10 and 28 have a strong direct dependency despite not being sequential shots. This tight dependency pulls the early shots (before 10) closer to the later shots (after 28) creating the gently curving structure at the top of the graphical model. The tight bunching of all vertices after about shot number 70 reflect that there is little compositional change at this depth.

In Figure 7, McGrath-5 location 3, the graph forms two different groups of vertices, where there are many edges within each group but few between the two groups. In graph theory, these groups are called clusters. Vertices 1-60 form a cluster that is slightly separated from the cluster formed by vertices 60-150, suggesting a relatively sudden change in composition at this depth. Note that in social-network analysis, graph clusters are typically much more separated, often containing no edges at all between the two clusters. Such a complete separation of clusters will not happen in LIBS data because of the conical shape of the target ablation, guaranteeing that each sample contains some overlap with the previous sample.

In Figure 8, McGrath-5 location 5 demonstrates a common feature found on targets in Gale crater; which is a chain. It shows that the first several shots are directly dependent on the prior shot and the following shot in the sequence, creating a short chain from Shot 1 through Shot 5. This is likely showing the composition of a thin surface layer of dust. Shots 8-150 cluster closer to one another but show a very clear trend from first shots to last as the laser penetrates the rock surface, pointing to a constantly changing composition with depth.

C. Results from Rock Target “Coronation”

In contrast to the examples on McGrath-5, if there were little compositional change with depth, the graph would be expected to show no structure by shot, as seen for the very first ChemCam target, Coronation, in Figure 9. Beyond the first two shots, there is no apparent ordering or structure to the vertices in the graph. This indicates that the graph learning algorithm cannot distinguish a closer dependency between sequential
Fig. 5. (a) Mastcam image of McGrath target, sol 166 (mcam00888, image 0166MR0888019000E1). (b) RMI of McGrath5, sol 234 (ccam02234) with 5x1 raster locations with 150 shots each.

 shots than any other pair of randomly selected shots. Such a graph is often referred to as a blob or hairball in graphical model literature. These blobs are indicative that there is little shot-to-shot compositional change in the target and so it still provides us with some information about depth trends (or lack thereof) in the target.

IV. CLUSTERING IMAGES BY SURFACE CHARACTERISTICS

We use computer vision to automatically analyze salient features of materials in images. Based on prior work in the analysis of microscopy images relevant to nuclear material forensics [12]–[14]; we apply this expertise to the RMI images. Feature vectors are constructed from raw pixel data that help to describe human-interpretable qualities of the imaged material, such as texture, granularity, size of nodules, prevalence of nodules, etc. Our feature vectors include classical methods such as co-occurrence matrices and morphological granulometries. We then determine the features that are most relevant for identifying coatings and other surface features indicative of compositional characteristics on martian rock targets through interactive clustering of images.

To help choose the most relevant features for the RMI images and surface category types of interest, we use interactive machine learning. We investigate image clusterings using a human-in-the-loop to determine which features provide the greatest separation of clusters. Figure 10 illustrates results obtained in this way. Results indicate that a number of different surface morphologies of interest can be characterized, and differentiated, using a number of standard texture descriptors.
Fig. 8. Graphical model of shot-to-shot depth trends from Mcgrath-5 location 5, sol 234 (ccam02234). The first several shots are directly dependent on the prior shot and the following shot in the sequence, creating a chain from Shot 1 through Shot 5. This is likely showing the composition of a surface layer of dust. Shots 8-150 cluster closer to one another but show a very clear trend from first shots to last as the laser penetrates the rock surface, pointing to a constantly changing composition with depth.

Fig. 9. Graphical model of shot-to-shot depth trends from Coronation, sol 13 (ccam02013). Beyond the first two shots, there is no apparent depth trend. The graph learning algorithm cannot determine whether shots 30 and 18 (for example) are any more dependent on each other than shots that are sequential, therefore there is no apparent structure in the edges of the graph. This uninformative structure-less shape is often called a blob or hairball in graphical model literature.

V. QUANTIFIED FEATURES OF ROCK TARGETS

From the RMI images, we have a collection of features that represent surface texture descriptors that help to characterize rock targets. In addition to these features, we want depth characteristics extracted from the LIBS data to help scientists to quickly explore the large set of rock samples from ChemCam. Based on observations of graphical models learned from rock targets on Mars, we identify several features of interest.

As we have seen, rapidly changing chemistry at the surface is often represented as a chain in the GGM. Additionally, the entire collection of samples in a location may be more or less chain-like. Therefore, for each target, we learn a GGM and then from that GGM calculate a chain score for the surface, and the interior. The chain score is based on how closely the learned graph matches a perfect chain in which sequential shots are connected to each other, but non-sequential shots are not directly connected. We define a similarity function, \( S(\hat{\Theta}, G^*) \) between the learned GGM, \( \hat{\Theta} \), and an ideal solution, \( G \):

\[
S(\hat{\Theta}, G^*) = 1 - \frac{2}{p(p-1)} \sum_{i=1}^{p} \sum_{j=i+1}^{p} (j - i) \left| \hat{\theta}_{ij} - g_{ij} \right|
\]

where \( p \) is the number of vertices in the graph. This similarity function weighs the differences between the ideal graph and the learned graph by how far apart the nodes are in the sequence 1, 2, \ldots, \( p \). The ideal solution, \( G^* \), is a graph of the perfect chain. Therefore, if learned edges appear between shots that are far apart sequentially, this lowers the score more than edges between nodes that are not so far apart sequentially.
Based on observations of graphical models learned from the targets, we are also interested in graphs that cluster, indicating a sharp change within the rock. We calculate a clusteredness score, which calculates the cost of creating a graph cut at each vertex in the graph. A low graph-cut cost produces a high clusteredness score. We use both the score and the vertex number at which the maximum score was achieved as features.

For the clusteredness score, we imagine that a compositional layer change could occur at any shot, \( k \), in the sequence \( 1, 2, \ldots, p - 1 \), so that the graph would form two clusters, with one cluster containing vertices \( \{ X_1, X_2, \ldots, X_{k-1}, X_k \} \) and the other cluster containing the rest of the vertices \( \{ X_{k+1}, X_{k+2}, \ldots, X_{p-1}, X_p \} \). To calculate the clusteredness score, for each \( k \), we compare the learned GGM, \( \hat{\Theta} \), against the ideal clustered graph, \( G^*(k) \), which is a binary block-diagonal matrix, such that the entries are defined as:

\[
g_{ij}^*(k) = \begin{cases} 
1 & \forall \ i \leq k, j \leq k \\
1 & \forall \ i > k, j > k \\
0 & \text{otherwise}
\end{cases}
\]

The clusteredness score is therefore \( \max_k S(\hat{\Theta}, G^*(k)) \). We also use the \( \arg \max \) value of \( k \) as a feature indicating where the layer change occurs.

Another feature of interest to geochemists is knowing which wavelengths contributed most to the trends seen in the graphical model. Therefore, we construct features from the wavelengths that have the highest variance (and therefore contribute the most to the \( \hat{\Theta} \) values) across the shots with edges connecting them:

\[
C = \frac{1}{|E|} \sum_{e_{ij} \in E} \hat{\theta}_{ij}(X_i - \bar{X}_i)(X_j - \bar{X}_j),
\]

where \( C \) is a vector of length \( n \), the number of wavelengths. The highest-value elements of \( C \) correspond to the wavelengths that contribute most to the learned graphical model. We use the wavelengths associated with the top ten element values of \( C \) as features.

Altogether, this is a total of 14 features derived from the Gaussian graphical model. These features were selected as features to help distinguish targets with depth trends of interest.

### VI. CONCLUSIONS

The GGM estimated from a sequence of shots on a single location indicate that there are shot-to-shot dependencies that follow expected patterns. Without any prior information about the shot sequence, the GGM is often able to reconstruct the shot sequence on the first several shots as a chain, which is consistent with observations that these first shots may differ from the underlying rock. In our primary example of this, the surface of Bell Island is extremely dusty, making dust a likely contributor to the spectra of the initial shots (e.g. [9]). In other examples, particularly on the depth analyses of McGrath, we see that a general depth trend from the outer to inner shots on a target follow a chain-like structure. Deviations from the chain, such as an s-curve or clustering of vertices in the graph are indicative of interesting geochemical compositional trends, such as thin stratigraphic layers. These depth trends can be visually interpreted through graphical models. These interesting features in the graphical models can then be quantified and combined with features extracted from surface images to enable scientists to quickly explore the large set of rock targets sampled by ChemCam.

We have identified quantifiable features that can be automatically extracted from ChemCam LIBS and RMI data. We plan to further exploit these features in learning algorithms to automatically classify rock types and to predict depth trends from surface images.

### ACKNOWLEDGMENTS

Work sponsored by Los Alamos National Laboratory’s LDRD program. We thank the ChemCam team for providing publicly available data in NASA’s online archive1.

### REFERENCES


1http://pds-geosciences.wustl.edu/missions/msl/chemcam.htm