Interactive Gaussian Graphical Models for Discovering Depth Trends in ChemCam Data.

D. A. Oyen[1], C. Komurlu[1,2], and N. L. Lanza[1]. [1]Los Alamos National Laboratory, doyen@lanl.gov. [2]Illinois Institute of Technology.

Introduction: Machine learning algorithms are now demonstrating human-level performance on some difficult benchmark problems, such as identifying objects in images. With the quantity of planetary data rapidly increasing, we would like to harness the power of machine learning to further advance planetary science. However, most of the recent successes in machine learning depend on using massive, labeled sets of data to train the algorithms. For many planetary science questions, such labeled data sets may not exist or it may not be possible to label data in such a way that would help to answer open-ended scientific inquiries. However, machine learning can improve the science return of remote sensors by increasing the speed at which scientists discover interesting patterns in their data. Interactive machine learning balances the strengths of machine learning to perform repetitive pattern recognition tasks, while empowering scientists to explore the factors that produce interesting patterns in large sets of data [1].

Spectrometers are increasingly used in remote sensing, yet spectral data can be difficult to analyze due to its high-dimensionality and non-linear mapping to interpretable quantities. As part of the Mars Science Laboratory rover operations, ChemCam's Laser-Induced Breakdown Spectroscopy (LIBS) instrument collects fine-scale atomic spectra from targets up to 7m away [2]. Given the high number of ChemCam observations to date (>400,000) and the high dimensionality of LIBS spectra (~6000 channels), advanced analysis methods are needed. We present an unsupervised machine learning method for discovering surface compositional features on rocks in ChemCam targets [3]. Our approach uses interactive machine learning to (1) give a visualization of shot-to-shot relationships among LIBS observations on a single target, and (2) identify the wavelengths (or elements) involved in the trend. Using the insight that the precision of element abundance is more reliable than accuracy [4], we bypass the quantification of elements, and look directly for patterns of chemical gradients [5, 6]. Additional trends involving different chemistry can then be explored on the same target. We are working to extend this to search the full archive of ChemCam spectroscopy data to find similar geochemical trends among all targets.

Machine Learning for Pattern Discovery: When machine learning is used for pattern discovery, we have data X but do not have labels. Therefore, we use *unsupervised* machine learning which takes the form of probability density estimation, $X \sim P(\theta)$, for a fixed distribution family P and learned parameters θ. The algorithm learns the distribution by inferring the optimal parameters $\hat{\theta}$ from the data, $\hat{\theta} = \arg\max_{\theta}\ [L(X;\theta) - R(\theta)]$, where $L(X; P(\theta))$ is the likelihood of the observed data given the probability distribution $P(\theta)$, and $R(\theta)$ is a regularization term that typically penalizes complex models. The structure of the probability distribution is typically the most interesting aspect because it reveals interesting patterns about the data. Some examples include clustering which assumes that P is a distribution with multiple modes (or centers of clusters); and probabilistic graphical models which assume that P is a multivariate joint distribution that can be factored compactly indicating direct dependencies.

Gaussian Graphical Models: Our previous work demonstrates our method for visualizing shot-to-shot relationships among LIBS observations to discover geochemical trends [5,6]. Here we give some background information to understand the approach before discussing our recent work in identifying the geochemistry involved in discovered trends. Probabilistic graphical models [7], and specifically, Gaussian graphical models (GGM) [8], are unsupervised learning models that assume that each data sample $X = (x_1, x_2, \ldots, x_p)$ is a p-dimensional vector generated by a multivariate joint distribution. Furthermore, the probability distribution can be factored into a compact representation with just a few direct dependencies. The compact representation assumption is a statistical necessity for the robust estimation of a high-dimensional distribution from finite data; and it reveals interesting structure about the dependencies among variables.

To analyze the depth trend of a rock target at a location, we estimate *partial correlations* among spectra using the GGM algorithm. A partial correlation between shot A and shot B 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 A and B is 0 then A and B are conditionally independent. A GGM is estimated from a data matrix $\boldsymbol{X}$, 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, $\boldsymbol{\Sigma}$, is calculated from $\boldsymbol{X}$, then the best sparse approximation, $\boldsymbol{\Theta}$, to the partial correlation matrix for a given sparsity constraint, λ, is estimated. The number of non-zero partial correlations is controlled by the value of λ, which can be any non-negative real number.

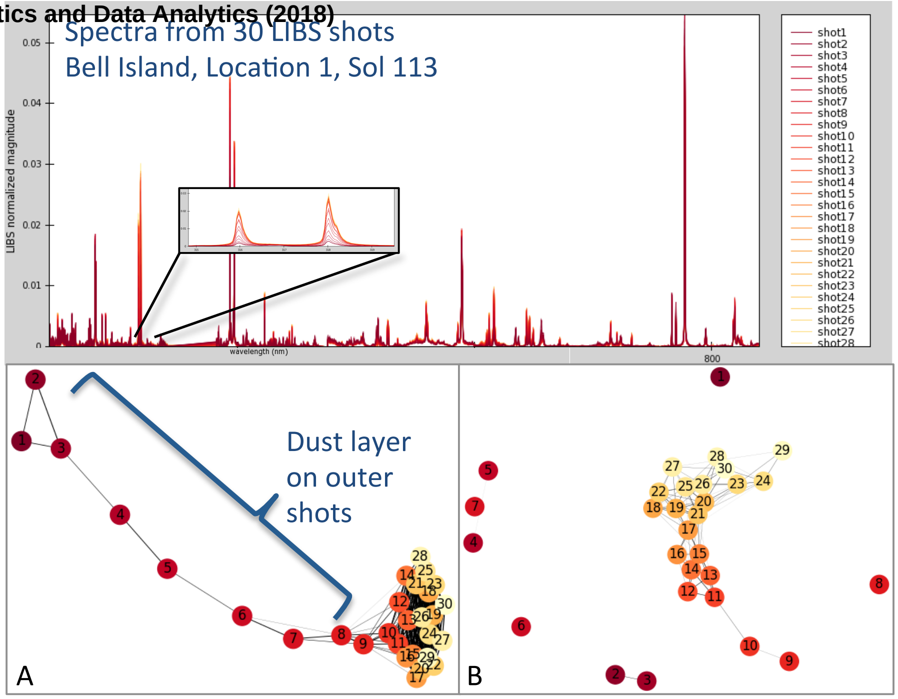

Figure 1 Interactive machine learning takes spectral data from a several LIBS shots (**top**) and learns a GGM (**bottom A**) indicating geochemical trends in ChemCam targets. In this case a clear trend is present in the first several shots at the surface of the target. iGGM then identifies the wavelengths that are responsible for the major structures of the GGM. In this case, when those wavelengths are masked, iGGM learns a GGM without that surface feature (**bottom B**), which can now be used to investigate other geochemical trends in the target.

The resulting GGM is displayed using a spring layout that places strongly correlated nodes near each other as if the correlation weights are springs pulling nodes together in space. If there are no systematic trends, then the non-zero partial correlations will appear on seemingly random pairs of shots, and the displayed GGM will look like an amorphous *blob* (or *hairball* in graph theory terminology). More visually interesting patterns emerge when there are interesting depth trends, such as a chain for systematic decrease/increase in elements, or clusters for sudden change in chemistry (such as a layer). This automated method identifies compositional depth trends associated with varnish and weathering rinds on laboratory samples [5]; and dust layers and thin sulfate veins on Mars targets [6]. We can see in the GGM Figure 1 (A) that there is a surface layer on the ChemCam target. In this case, it is a dust layer which we verified by looking at the decrease in abundance of elements associated with martian dust (e.g., Mg [10]) and the increase in abundance with S and Ca.

Explaining the Geochemistry: The GGM gives a quick visual summary of geochemical trends, but to answer specific science questions, we need to know the geochemistry behind the observed trend. We introduce an interactive Gaussian graphical model (iGGM) algorithm in which the algorithm identifies the wavelengths in the LIBS spectra that are essential for producing the most prominent structures in the learned GGM. This set of wavelengths explains the geochemistry behind the trend. For example, in the Bell Island data, the most prominent structure in the learned GGM is the chain among the first several shots. Our iGGM algorithm identifies the wavelengths that if they were masked from analysis, would make that chain disappear as in Figure 1 (B).

The iGGM algorithm identifies the critical wavelengths by searching though all possible subsets of wavelengths to find a subset of wavelengths that if they were masked would most change the structure of the learned GGM. The gradient of the weighted covariance matrix is calculated with respect to the sample weights. Then a regularization term is placed on the number of weights that can be changed to avoid the trivial solution of masking all weights. The resulting masked wavelengths are those that are critical for producing the trends seen in the GGM. iGGM also displays the newly learned GGM from the masked data as can be seen in Figure 1 which often reveals further geochemical trends in the same target.

Future Work: We plan to extend this work to facilitate quickly searching through the entire data archive of ChemCam observations to find targets with similar geochemical trends.

References: [1] Porter et al. (2013) *Comp. in Sci. & Eng.* [2] Wiens et al. (2012) *Space Sci. Rev.,* 170. [3] Lanza et al. (2015). *Icarus*. [4] Blaney et al. (2014), *JGR*, 119, 2109-2131. [5] Oyen and Lanza. (2015). *LPSC* abstract 2940 [6] Oyen and Lanza. (2017). *LPSC* abstract 1479. [7] Koller and Friedman. (2009). *Probabilistic Graphical Models*. [8] Zhao T. et al (2012) *J. Machine Learning Research*. [9] Oyen et al (2016) *Intl. Conf. Artificial Intelligence*. [10] Lasue et al. (2014). *LPSC*, abstract 1224.