BEHAVIOR OF FEATURE SELECTION IN LIBS SPECTROSCOPY AS A FUNCTION OF VARYING DISTANCE AND DATA PRE-PROCESSING. M. L. Carmosino[1], E. A. Breves[2], M. D. Dyar[2], M. V. Ozanne[2], S. M. Clegg[3] and R. C. Wiens[3]. [1]Computer Science Dept., University of Massachusetts, 140 Governor's Dr., Amherst, MA 01003, mcarmosi@cs.umass.edu, [2]Dept. of Astronomy, Mount Holyoke College, 50 College St., South Hadley, MA 01075, [3]Los Alamos National Laboratory, P.O. Box 1663, MS J565, Los Alamos, NM 87545.

Introduction: ChemCam, a remote sensing instrument package including a Laser-Induced Breakdown Spectrometer (LIBS) and remote micro-imager (RMI), will provide geochemical analyses and context imaging as part of the Mars Science Laboratory (MSL) *Curiosity* rover payload. Statistical analysis is used to extract elemental compositions from LIBS data. Traditional techniques such as simple linear regression of a single peak or multiple linear regression are affected by noise and the highly collinear covariates, e.g., each element has many peaks with highly-correlated intensities. The preferred technique for LIBS data analysis of geological samples is partial least-squares analysis (PLS) [1-3]. Shrunken regression techniques such as the lasso (least absolute shrinkage and selection operator) are also useful because the predictions use fewer (sparse) coefficients.

The goal of this study was to use the lasso approach [4] to examine the importance of varying distance and data pre-processing (baseline subtraction) on results of multivariate analysis of LIBS data. From a statistical perspective, implementing feature selection effectively is necessary as a precursor to applying other types of machine learning to highly-multivariate analysis of LIBS data. From a spectroscopic viewpoint, feature selection allows study of how specific peaks in LIBS spectra behave at different distances and with different numbers of shots/data. This abstract seeks to characterize how LIBS data can vary with distance; a related abstract [5] suggests a method for distance correction that facilitates creation of a distance-independent spectral library (training set).

Experimental Methods: LIBS was used to analyze a suite of 30 pressed pellets of powdered rock standards at Los Alamos National Laboratory. Pellets were placed in a chamber evacuated and filled with CO_2 to a pressure of 7 Torr. A 1064-nm Nd:YAG laser operating at 17 mJ/pulse was used to ablate the samples. Three Ocean Optics HR2000 spectrometers with UV (223-326 nm), VIS (328-471 nm), and VNIR (495-927 nm) wavelength regions were used to collect the optical emission from the sample plasma. Major element compositions of samples came from [2,6,7]. This study used data on nine major elements: SiO_2, Al_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, and Fe_2O_3 for each of 10 scenarios: five different standoff distances (3, 4, 5, 6, and 7 m) with and without the bremsstrahlung continuum ("baseline") removed. Typical variable distance data with the continuum removed are shown for standard sample GBW-07114 in Figure 1.

Data Analysis: Wavelength calibration was performed for each of the three spectral regions and resampled to standardize the wavelength scale. Laser shots were averaged and smoothed. All data were then processed in two different ways: the baseline bremsstrahlung continuum was either fit and subtracted or

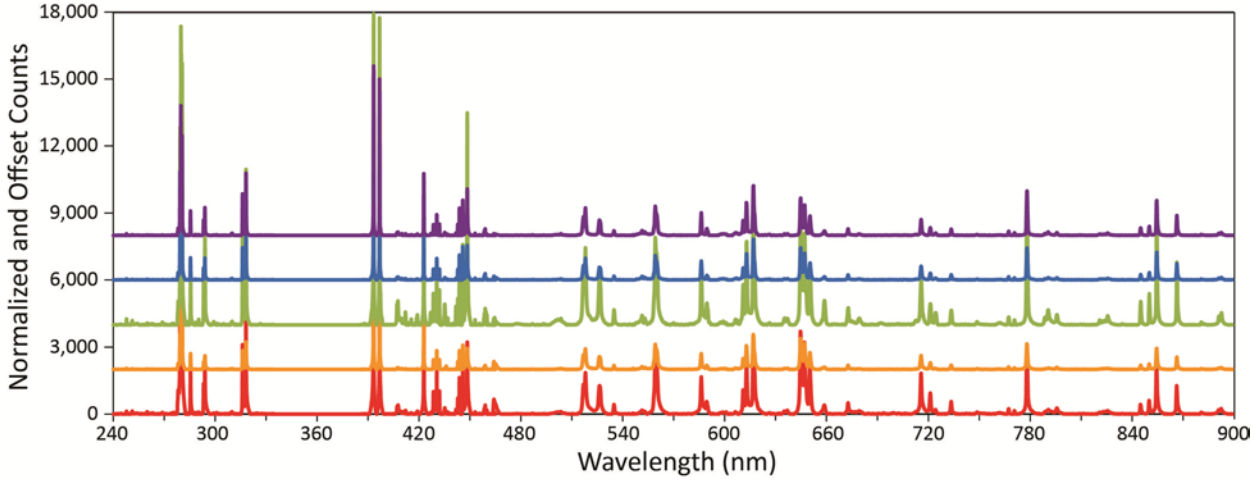

Figure 1. LIBS spectra of sample GBW-07114 acquired at 3, 4, 5, 6, and 7 m stand-off distances. Wavelength calibration was performed for each of the three spectral regions and resampled to standardize the wavelength scale. Laser shots were averaged and smoothed, and bremsstrahlung continuum was removed.

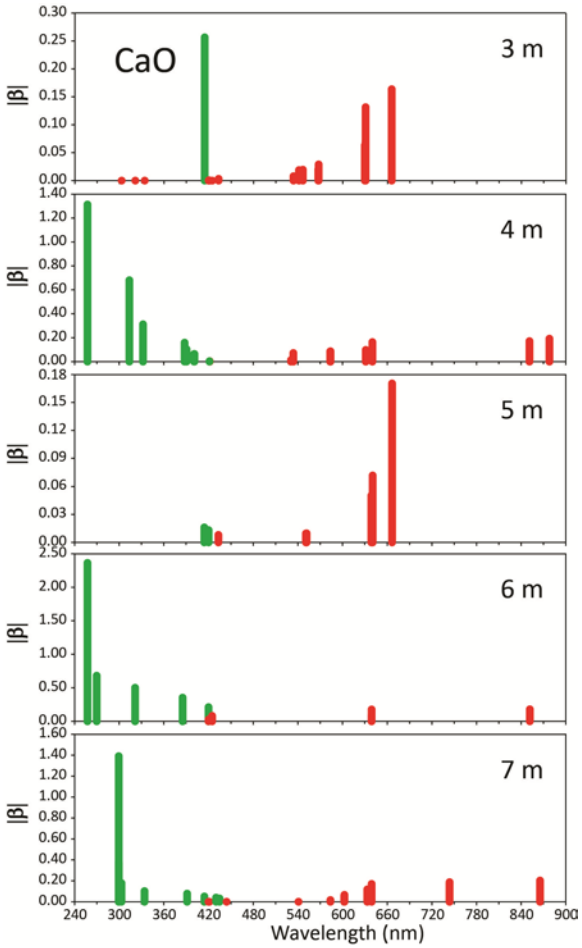

Figure 2. β values for CaO models of baseline-subtracted LIBS data. Positive β's are shown in green and negative coefficients in red. The position of the line shows the wavelength where the feature occurs; the height of the line is the magnitude of β at the wavelength on the *x* axis.

this step was skipped, as noted above. Lasso models were built using the R statistical software package glmnet [6]; details of the procedure are given in [7,8]. The parameter *t*, an index of model complexity somewhat analogous to the number of components in a PLS model, was chosen for each model using a heuristic that selects the component number with a mean-squared error (MSE) within one standard error (SE) of the minimum.

Results of Variable Distance: Varying the distance between the telescope and the target results in qualitative differences in peak intensities and the overall shape of LIBS spectra as seen for the example in Figure 1. Plots of the β coefficients chosen by the lasso models at each distance show that the models are truly selecting different peaks at each distance, showing the response of the plasma to changes in beam area and acceptance angle [5]. Figure 2 shows values of β coefficients chosen for CaO at varying distances. We expected that the prominent emission line at 422.67 nm would be utilized in all models. However, while it does appear, its magnitude varies greatly. The lasso-selected lines yield fundamental insights into which ionization states dominate at variable distances, and these variations are not smoothly linear with distance. The models make heaviest use of channels in the UV and VIS regions of the spectra where the strong lines from neutral and low ionization states fall. This result may also be caused by differences in spectral responses of the different detectors, which have not been corrected for in this study (but see [5]).

Results of Baseline Removal: To assess the effect of bremsstrahlung continuum subtraction on feature selection, the percentage of features (selected wavelength bins) that occur more than once for each element across all distances was calculated:

$$Overlap = \frac{total\ \#\ features - unique\ \#\ features}{total\ \#\ features} \times 100.$$

Table 1. Overlap Comparison

	B-sub	No B-sub
SiO_2	6.90	14.89
Al_2O_3	3.85	12.24
MgO	14.58	12.82
CaO	6.15	20.59
Na_2O	10.71	10.00
K_2O	9.30	13.79
TiO_2	5.26	8.70
P_2O_5	2.78	23.33
Fe_2O_3T	6.52	0.00

B-sub = baseline subtracted data;
No B-sub = no baseline subtraction.

Overlap is thus an effective representation of whether the lines chosen as β's by the model for each wavelength are similar or not. This overlap metric does not count very close wavelengths (within 0.5 nm) as overlapping, so it gives a useful measure of how similar models are with and without baseline subtraction. Results are shown in Table 1. Overlaps on non-baseline-subtracted data are much higher than for baseline-subtracted data, but actually there is very little feture-sharing among distances for either set of models. This result underscores the importance of continuum removal, as also demonstrated by [2,5].

Acknowledgments: Research supported by NASA MFRP grants NNG06GH35G and NNX09AL21G.

References: [1] Clegg S. M. et al. (2009) *Spectrochim. Acta B*, 64, 79-88. [2] Tucker J. M. et al. (2011) *Chem. Geol.*, 277, 137-148. [3] Dyar M. D. et al. (2012) *Chem. Geol.*, 294-295, 135-151. [4] Tibshirani R. (1996) *J. Royal Stat. Soc. B, 58*, 267-288. [5] Clegg S. M. et al., this conference. [6] Fabre et al. (2011) *Spectro Acta B*, 66, 280. [7] Vaniman et al. (2012) *Space Sci Rev.*, submitted. [8] Ozanne, M. V., this conference. [9] Dyar M. D. et al. (submitted) *Spectrochim. Acta B.*