Comparison of Shrunken Regression Methods for Major Elemental Analysis of Rocks Using Laser-Induced Breakdown Spectroscopy (LIBS)

Marie Veronica Ozanne

Thesis Advisor:
Professor M. Darby Dyar

Submitted to the Department of Chemistry at Mount Holyoke College in partial fulfillment of the requirements for a Bachelor of Arts with departmental honor

May 2012

For my parents, who believed in me even when I did not believe in myself

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of my research advisor, Professor Darby Dyar. Her willingness to engage in interdisciplinary research and push me outside my comfort zone has been the greatest gift of my undergraduate career.

I would like to thank Marco Carmosino for teaching me much of what I know about programming and for being a great sounding board.

I am grateful to Elly Breves, who introduced me to this dataset and who has patiently answered my numerous LIBS questions. She is the best lab manager for whom anyone could hope.

Special thanks go out to Dr. Sam Clegg and Dr. Roger Weins of Los Alamos National Laboratory who collected this data, and to NASA, who funded this research through the Mars Fundamental Research Program.

Table of Contents

List of Figures and Tables

INTRODUCTION

Since the early 1800s, scientists have known that specific colors are observed when different elements are burned. When a molecule is heated, it absorbs a photon, which increases its energy; it is promoted to an excited state. As the molecule cools, it emits a photon, which decreases its energy; the lowest energy state is the ground state (Harris, 2010). Photon absorption and emission occur at wavelengths specific to each element.

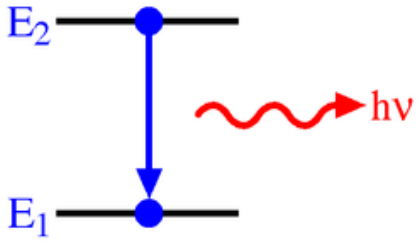

Figure 1: As it decays back down to the ground state, the molecule emits a photon at a particular wavelength.
http://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/AtomicLineSpEm.png/250px-AtomicLineSpEm.png

In a flame test, different chemicals are burned. They produce differently colored flames that are related to the wavelengths at which their component molecules emit photons. These colors range from red and orange to blue and violet. Sodium, for example, produces an orange flame, as shown in Figure 2. In fact, as researchers would later discover, the wavelengths at which absorption/emission occur are unique to a particular element or ion. This

information can be used to identify the composition of an unknown (Cremers *et al.*, 2006).

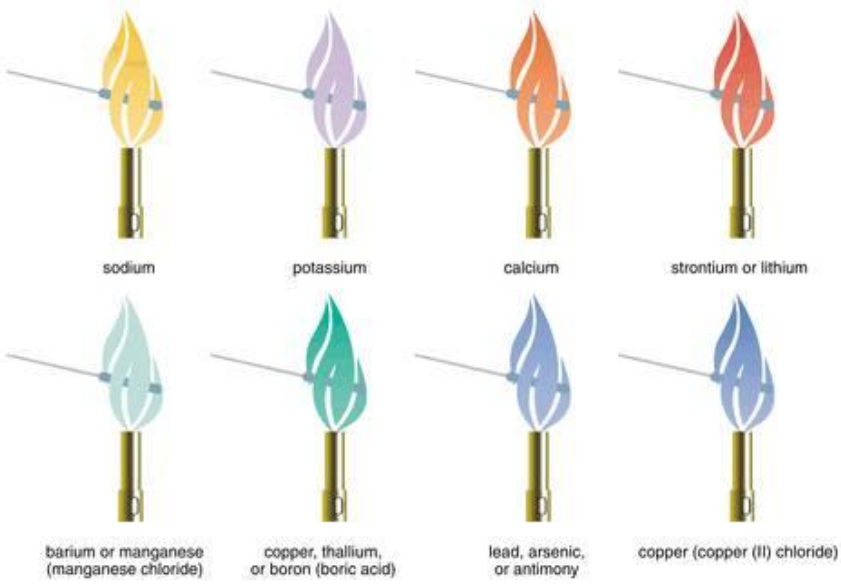

Figure 2: An artist's rendition of a simple flame test. When a given element or ion is burned, it emits photons at specific wavelengths, which correspond to the colors the eye observes. *http://media.tiscali.co.uk/images/feeds/hutchinson/ency/0008n042.jpg*

The absorption and emission properties of light are used in chemical analyses to determine concentrations and compositions of unknowns. The technique that takes advantage of these properties is known as spectroscopy. "Spectroscopy is the study of the interaction of electromagnetic radiation (light, radio waves, x-rays, etc.) with matter" (Harris and Bertolucci, 1978). There are many different types of spectroscopy, each with specific applications.

In general, a spectrophotometer requires a light source and a detector. Where appropriate, sampling can be done through laser ablation, where a laser is

the light source, and a pulsed laser beam is focused on a microscopic region of the sample. Particles, atoms, ions, and electrons are vaporized, creating an evolving plasma. Each laser pulse ablates a few nanograms of the sample, so it is nearly nondestructive. When a laser ablated plasma is analyzed using atomic emission, the technique is called laser-induced breakdown spectroscopy (LIBS) (Harris, 2010). LIBS dates back to 1963 when a laser-plasma was first used on a surface (Cremers *et al.*, 2006). While many forms of spectroscopy require careful sample preparation, LIBS requires none, making it ideal for analyzing rock compositions. Also, a LIBS instrument can operate from a distance, which makes it practical for field research.

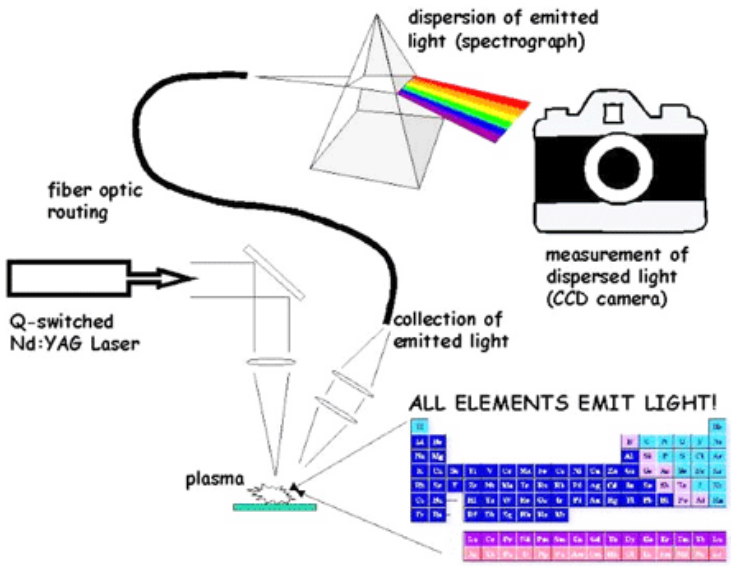

Figure 3: LIBS uses a laser to create a plasma that contains material from the sample of interest. This plasma emits light, which is collected and sent to the detector. A spectrum is drawn of signal intensity versus wavelength. The various intensity peaks correspond to known element emission peaks. *http://upload.wikimedia.org/wikipedia/en/c/c8/Libs.jpg*

A LIBS is one of the two instruments comprising ChemCam, which is part of the Mars Science Laboratory (MSL) on board the Mars rover, *Curiosity* launched in November 2011. This LIBS instrument records spectra for each sample in the ultraviolet (UV), visible (VIS), and visible and near infrared (VNIR) ranges, with signal intensities (elemental emission lines) at 6421 channels (wavelength values) corresponding to elements in the sample of interest.

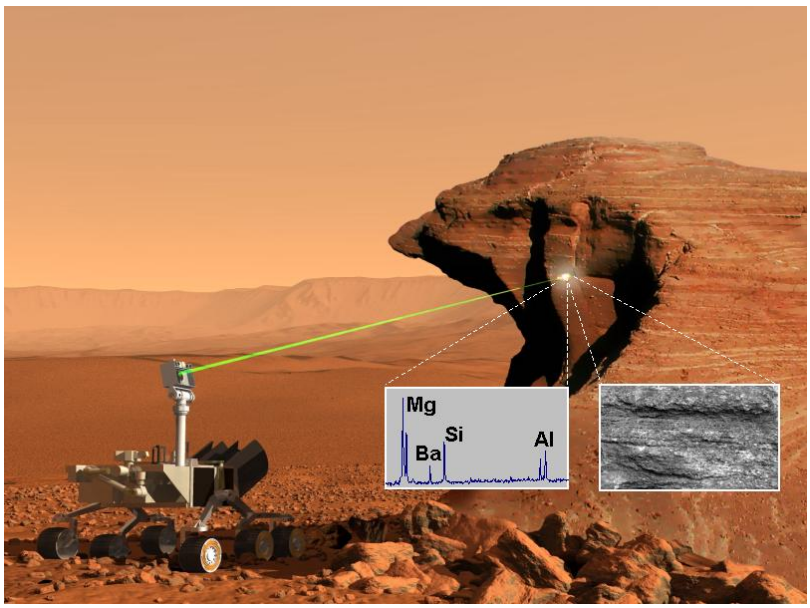

Figure 4: An artist's rendition of ChemCam as it analyzes a rock sample on the surface of Mars. A portion of the possible resulting spectrum is displayed.
http://smsc.cnes.fr/IcMSL/chemcam_operation.png

Mars rocks may have different elemental compositions from those on Earth due to differing conditions under which they were formed. In both places, however, the compositions of the various elements that make up the rock provide valuable information about the chemical evolution of the planet over geologic time. Because ChemCam can only transmit spectra from rock samples, rather than

the rocks themselves, back to Earth for analysis, our understanding of Martian geology depends on our ability to predict elemental compositions of rocks from the 6421 channels of LIBS spectral data.

The conventional method for analyzing spectroscopic data is univariate analysis. This method assumes that one component of the data (usually a single emission line) will adequately explain the behavior of the variable of interest. For example, the concentration of a prepared sample of potassium permanganate can be related directly by the measured absorbance using Beer's Law. However, for more complicated spectral relationships, univariate analysis does not provide such useful results.

Multivariate analysis is needed to account for the covariate interactions that occur. The most straightforward multivariate analysis method is ordinary least squares (OLS). This method assumes that multiple components are needed to adequately explain the behavior of the variable of interest; these various components are included in a statistical model. In order for OLS to provide a model that will yield stable predictions, the model predictors must not be highly collinear. This means that each predictor must contain minimal information about the response variable that is also contained in any other predictor. Due to matrix effects and spectral resolution, which will be explored further in the next chapter, a model in which all predictors are only moderately collinear cannot be constructed for LIBS spectral data of rocks. So, statistical methods that do not

require this feature are needed to provide reliable rock composition predictions from the LIBS data.

RESEARCH GOALS

This thesis explores a set of multivariate analysis techniques known collectively as shrunken regression that have been designed to provide stable models when the data suffer from multicollinearity. Data are multicollinear when two or more variables in the data are correlated and provide redundant information about the response (*Model diagnostics*, 2004).

Partial least squares (PLS) regression is the accepted statistical method in the LIBS community for generating prediction models in cases where the data are highly collinear. Dating from its introduction in the 1960s, PLS is a well-established method and is well-understood by the LIBS community. Due to its popularity, however, other statistical techniques designed to deal with highly collinear data have not been as widely explored to determine if they yield lower model prediction errors than the well-tuned PLS models. PLS results can be used as a benchmark for comparing model prediction errors from other shrinkage methods. Other shrinkage methods of interest are ridge, lasso, fused lasso, elastic net, and sparse partial least squares (SPLS).

In this thesis, specific situations in which any of four methods, including PLS, are most useful will be investigated. The intention of this research is to demonstrate how other shrinkage methods compare to the PLS method. These

techniques will enhance the utility of highly collinear data by providing different ways in which the data can be analyzed. The various models are suited to a range of applications.

BACKGROUND

LASER-INDUCED BREAKDOWN SPECTROSCOPY

Since the early 1800s, scientists have understood that different elements emit different colors. Such emissions have been observed in familiar plasmas like the sun and flames. These colors correspond to various wavelengths (λ) or frequencies ($\nu = \lambda^{-1}$), which are unique signatures for the elements that emit them. Although the colors that are observed represent a small part of the wavelength range, known as visible light (400nm – 700nm), the entire electromagnetic spectrum ranges from γ-rays with wavelengths of 1 pm to extremely low frequency (ELF) waves with wavelengths of 100 Mm.

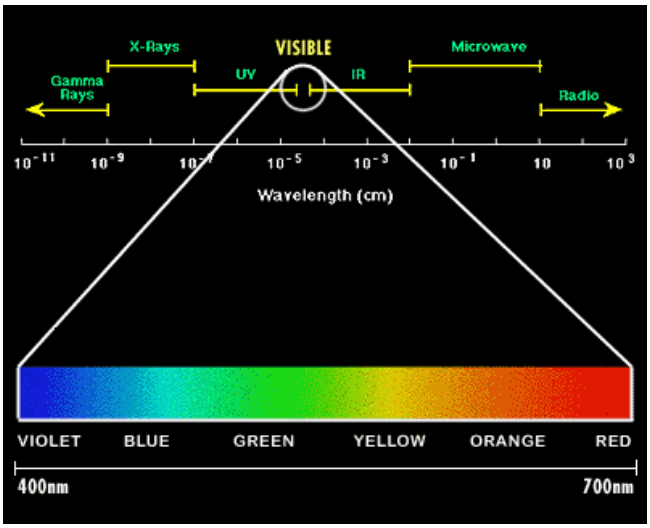

Figure 5: Diagram of the entire electromagnetic spectrum, excluding ELF waves. *http://www.sciencebuddies.org/mentoring/project_ideas/HumBeh_img019.gif*

Laser-induced breakdown spectroscopy (LIBS) owes its existence both to the aforementioned physical phenomenon and to the advent of the laser. The former is inherent in all forms of spectroscopy; sample compositions can be determined using the fact that, when treated with a light source, each element or ion emits photons at distinct and unique wavelengths. LIBS instrumentation belongs to a smaller group of spectroscopic techniques known collectively as laser spectroscopy. With the development of the laser came the realization that it could be used to create plasma that would reveal compositional information about samples of interest. Observations that were possible for the Sun now became possible for materials that did not naturally exist in a plasma state, such as rocks and metals. Lasers allow for increased precision with regard to spectral line frequency measurements.

In the *Handbook of Laser-Induced Breakdown Spectroscopy*, 1963 is cited as the birth year of LIBS. This was the first time a laser-plasma was used on a surface for analytical purposes. This was just three years after American physicist Theodore H. Maiman developed the first pulsed laser at Hughes Research Laboratories in Malibu, California. Closely related to atomic emissions spectroscopy (AES), LIBS measures the intensity of light emitted from a plasma using a series of successive wavelengths to measure the concentration of a particular element in the sample of interest. AES, however, is limited in that it requires sample preparation to dissolve and dilute a sample, and it cannot operate at standoff distances. In both AES and LIBS, thermal energy excites electrons into

higher energy electronic states; photons are emitted as the electrons decay back down to their ground states.

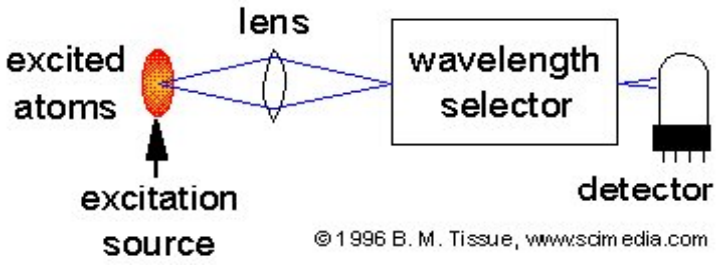

Figure 6: Schematic comparison of AES. AES uses a light source from a spark, flame, or plasma. It requires sample preparation to dissolve and dilute the sample, so it is not well suited to field work.
http://www.chemistry.adelaide.edu.au/external/soc-rel/content/images/aes-expt.png

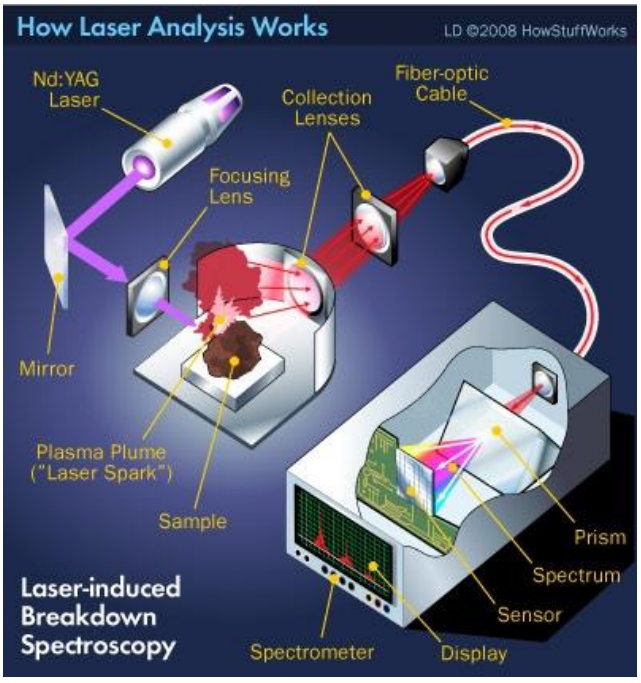

Figure 7. Schematic of LIBS. The light source is a Nd:YAG laser. The detector registers atomic emission to create a spectrum of the sample. Because a laser is used, no sample preparation is needed and the instrument can operate from a distance to analyze a sample. LIBS registers wavelengths ar which photon emission occurs from electronic transitions.
http://technology.automated.it/wp-content/uploads/HLIC/f32cc86353da5e06060fc83bd94e1b83.jpg

In 2004, LIBS was chosen as one of two analytical components in the ChemCam instrument on board the Mars rover, *Curiosity*, which launched on November 26, 2011. This LIBS instrument has the ability to collect LIBS data from standoff distances as great as 7 m. As the excited electrons decay back to the ground state, a light-emitting plasma is produced. From this plasma, a spectrum is obtained that covers three wavelength regions, UV, VIS, and VNIR (223 – 927nm). The spectrum contains a rich array of elemental lines from more than thirty elements likely to be present in the geological sample. These data can be interpreted using multivariate statistical techniques to determine elemental concentrations, like those of the ten major elements, which are by geochemical convention expressed as the oxides SiO_2, Al_2O_3, MnO, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, and Fe_2O_3.

Emission Lines

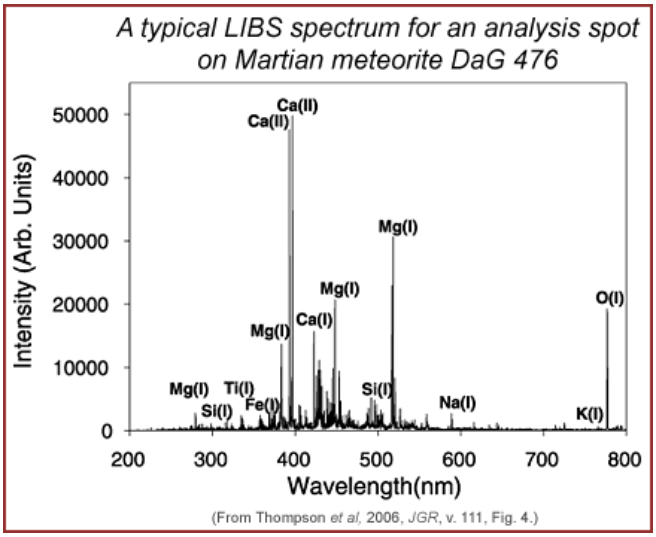

Figure 8. Plot of major elemental emission lines in the LIBS spectrum for a spot on the Martian meteorite DaG 476. As shown here, the emission lines can be matched to known emission lines in a database such as the National Institute of Standards and Technology (NIST) database. *http://www.psrd.hawaii.edu/WebImg/LIBS-DaG476.gif*

LIBS relies on quantized valence-electron transitions that occur when the electrons move to an excited state in the presence of a light source and subsequently decay back down to their ground states, emitting photons. When these transitions are detected by a spectrometer, emission lines are observed at wavelengths that are specific to the elemental or ionic electron source. The National Institute of Standards (NIST) maintains an online database of emission lines for each element/ion in the UV, VIS, and VNIR spectral regions. With proper instrument and spectral calibration, emission lines from the sample spectra can be matched with persistent (stable), observed lines in the NIST database. These emission lines are critical to determining the elemental composition of an unknown sample.

LIBS Challenges for Geological Samples

LIBS is challenging to use for geological sample analysis because peak intensities and areas are influenced by interactions in the plasma that are partially a function of the sample's chemical composition. These interactions are referred to as matrix effects; they are chemical properties of a material that influence the extent to which a given wavelength emission is detected relative to the true abundance of the parent element. The matrix effects are related to the "relative abundances of neutral and ionized species within the plasma, collisional interactions within the plasma, laser-to-sample coupling efficiency, and self absorption" (Tucker, *et al.*, 2010). Matrix effects can cause emission lines not to

be fully resolved (line overlap), which complicates spectral interpretation. Because the data contain so many channels, LIBS also suffers from the "curse of dimensionality", because our intuition about how data will behave breaks down in high dimensions (Hastie, *et al.*, 2009). Fortunately, advanced statistical analysis techniques can tease out relationships that are at first obscured by matrix effects.

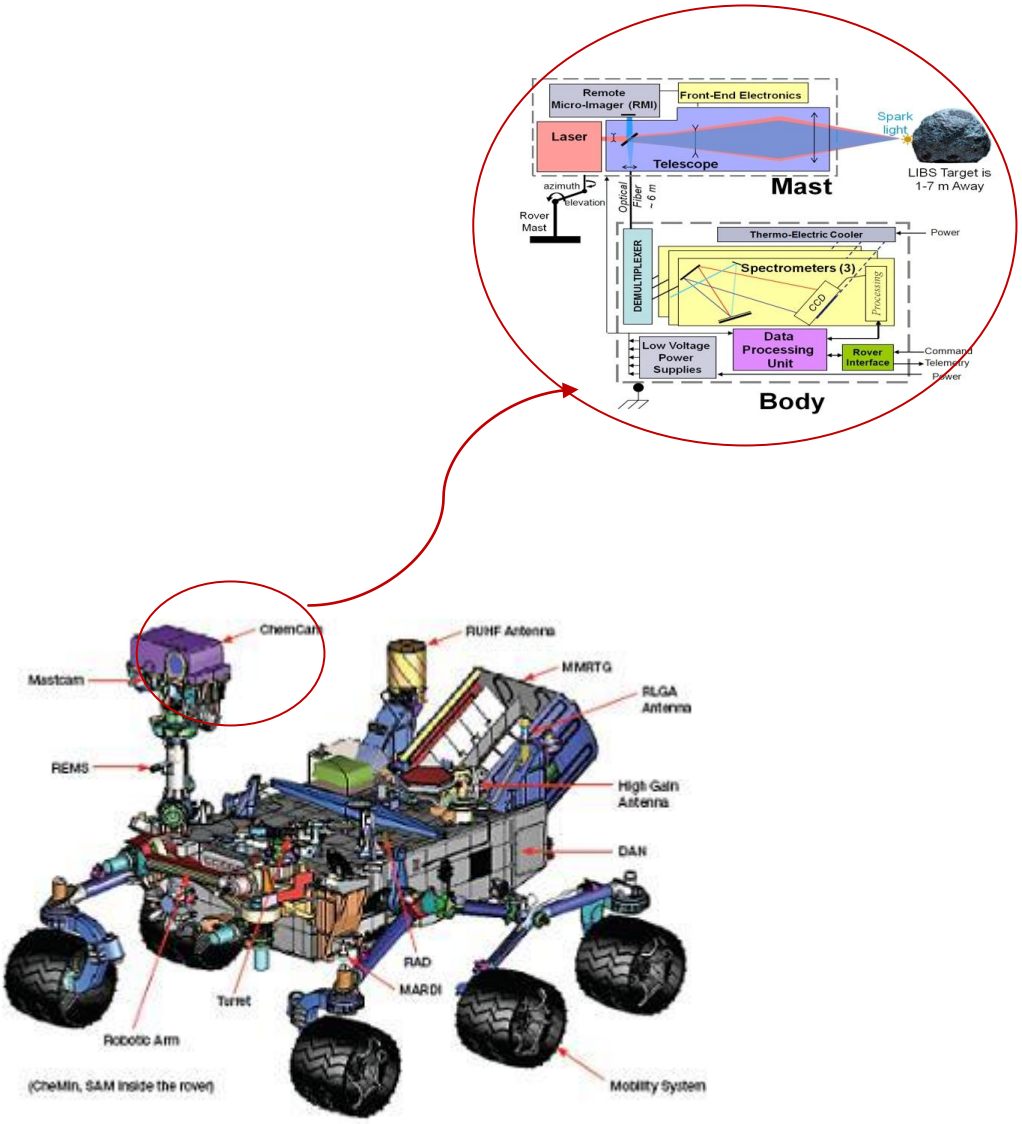

Figure 9: ChemCam schematic on *Curiosity*.
http://msl-scicorner.jpl.nasa.gov/images/ChemCam_block1.jpg
http://www.nasaspaceflight.com/wp-content/uploads/2011/11/D81.jpg

QUATITATIVE ANALYSIS OF EMISSION SPECTRA

What Makes a Good Model

An ideal regression model is sparse[1], interpretable, and robust. Sparsity is important because it implies a simpler model. A simpler model that performs as well as more complex model is always preferable because it is easier to understand. Convenient interpretation is crucial for understanding the chemical makeup of geological samples. Robustness ensures that prediction results are reliable and reproducible.

Univariate Analysis

The conventional method of analysis for spectroscopy data is univariate analysis. The equation has a familiar form:

$$y = \beta_0 + \beta_1 x. \quad (1)$$

In this equation, β_0 is the intercept, β_1 is the slope, x is the independent variable, and y is the dependent variable. This technique can be used to fit a calibration curve for the absorbance (y) versus concentration (x) of a given solution, for example. Such a calibration curve can then be used to interpolate the concentration of an unknown using the absorbance of that solution.

[1] Sparsity refers to the assumption that a smaller subset of the predictor variables is driving the prediction results (Chun and Keleş, 2010).

Multivariate Analysis

Ordinary Least Squares (OLS) uses the Method of Least Squares to determine the best fit for a set of data. It can be used to model the relationship between a response variable and one or more explanatory variables. To fit the model, the residual sum of squares (RSS), which is the distance between all of the responses (data) and their fitted values, is minimized (Ramsey and Schafer, 2002). OLS is a desirable method to characterize the relationship between the response variable and explanatory variables when the explanatory variables are not highly collinear. The general form of an OLS model is:

$$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \quad (2)$$

However, for applications where the explanatory variables (X_i) are highly collinear, OLS performs poorly. The resulting predictions are highly variable, and are thus neither reliable nor reproducible. Shrunken regression methods were developed for use in fields like chemistry where explanatory variables are often highly collinear and thus result in unreliable models. The methods explored here were either developed as direct responses to the OLS difficulties, or were developed to improve on existing shrunken regression techniques.

Partial Least Squares

Partial Least Squares (PLS), also known as projection to latent structures, has its origins in three papers by Wold, dating from 1966-1982. It was developed for use in situations with highly collinear explanatory (*p*) variables that significantly outnumber the number of observations (*N*), such that $p >> N$ (Butler and Denham, 2000). Although there are several variations of PLS, this thesis will only explore PLS2 (hereafter referred to as PLS), which has at least two response variables, because PLS2 has been shown to perform better on LIBS geological data (Dyar *et al.*, 2012) (Rosipal and Krämer, 2006).

PLS has been used to analyze data from a variety of types of spectroscopy, including, but not limited to, near infrared reflectance (NIR) spectroscopy, Fourier transform infrared (FTIR) spectroscopy, and Fourier transform-Raman (FT-Raman) spectroscopy. It has also been the dominant mode of analysis for LIBS spectra.

PLS is commonly employed to predict chemical compositions from a near infrared (NIR) reflectance spectrum (Goutis, 1996). For example, PLS models have been built to predict Ca, Mg, Na, K, P, S, Fe, B, and Mn concentrations in wines. This study revealed some relationships between NIR spectra and elemental concentrations in wines, although more data was needed to produce sufficiently stable models (Cozzolino *et al.*, 2007).

The presence of free fatty acids (FFAs) is one of the most important factors dictating the quality and economic value of olive oil. Because they are more prone to oxidation than triglycerides, their presence in edible oils increases the possibility of producing a rancid product. PLS models were built to determine the FFA concentration in commercial olive oil using FTIR spectra. (Iñón *et al.*, 2003).

PLS has been applied to FT-Raman spectroscopy, where it has been used to "construct highly correlated models relating a petroleum fuel's Raman spectrum to its motor octane number (MON), its research octane number (RON), its pump octane number, and its Reid vapor pressure" (Cooper, *et al.*, 1995). Analyzed petroleum blends contained more than 300 individual chemical species of different concentrations. Researchers found that PLS models could be used to predict MON, RON, pump octane number, and Reid vapor pressure of a fuel from its FT-Raman spectrum.

PLS chooses subspaces from the explanatory matrix, **X**, sequentially and projects the response vector, **y**, onto the subspaces of the column spaces of **X** to determine the model coefficients (Goutis, 1996). This applies a correction to the OLS coefficients to generate a model that does not experience the wild variance in predictions from which OLS suffers when features are highly collinear.

PLS involves a two-step process to determine the model coefficients. The first step is the shrinkage step. The shrinkage penalty determines the number of

factors to be included in the regression. This shrinks the coefficients by projecting down from *N*-dimensional space into a smaller *M*-dimensional vector space. In the context of this project, *N* = 6421, the number of channels (wavelength values) at which elemental intensity is measured. This uses linear combinations of previous variables to calculate the model coefficients. The second step completes the regression to produce a PLS model by calculating the RSS; it regresses the response, **y**, on the factors generated in the first step to minimize the sum of squares.

The optimization problem that must be solved to generate the PLS model coefficients is not a nice, tractable problem. It involves information about the variances and covariances of both the explanatory and response variables, which makes the shrinkage calculation complicated (Goutis, 1996).

The PLS formulation of the NIPALS algorithm introduced in Wold (1966) is as follows:

$$w_k = \arg max_w \, (w^T \sigma_{XY} {\sigma_{XY}}^T w), subject\ to$$

$$w^T \left(I_p - W_{k-1} W_{k-1}^{+}\right) w = 1 \ and \ w^T \sum_{XX} w_j = 0 \, . \qquad (3)$$

In this formula, $j = 1, \ldots, k\text{-}1$, σ_{XY} is the covariance of X and Y, I_p is the $p \times p$ identity matrix and W_{k-1}^{+} is the unique Moore-Penrose[2] inverse of $W_{k-1} = (w_1, \ldots, w_{k-1})$ (Chun and Keleş, 2009).

A full PLS model may contain hundreds of coefficients that are linear combinations of the original 6421 channels. Loadings are principal components computed from the spectral matrix (DePalma and Stephen, 2011). They can be used to map model coefficients onto emission lines. The number of coefficients (loadings) in the predictive model is often constrained to a smaller value to avoid overfitting. An overfitted model produces falsely optimistic prediction errors because it is highly customized to its training set; the model will not perform as well when tested on other unseen data.

Ridge Regression

Ridge regression was introduced by Hoerl and Kennard (1970). Ridge regression fits a model to a set of training data by minimizing a penalized RSS. It imposes a penalty, *t*, which explicitly constrains the size of the coefficients as shown in the following equation. This constraint introduces bias into the model. By implementing a bias-variance tradeoff, more stable coefficients are produced in the model.

$$\hat{\beta}^{ridge} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, subject\ to\ \sum_{j=1}^{p} \beta_j^2 \leq t \qquad (4)$$

[2] Independently identified by Moore (1920) and Penrose (1955), the unique Moore-Penrose inverse gives a unique solution if it meets four algebraic constraints (Gill).

The parameter t is always greater than or equal to 0; t controls the degree of shrinkage. In this model, y_i is the elemental abundance in sample i and x_{ij} is the intensity at wavelength j of sample i. Larger t values correspond to greater shrinkage, and thus greater bias. Tuning this parameter properly is crucial to producing the most stable model; an optimal bias-variance tradeoff must be employed.

A ridge regression model involves coefficients from all 6421 channels of the LIBS spectrum. Because it has so many coefficients, it is quite stable, but it does not provide a sparse model. This makes it difficult to discern which channels (and from them, elements) exert the greatest influence on the model prediction values. The model is generally difficult to interpret because it is not parsimonious.

Ridge regression has been used for different types of spectral analysis. For example, it was used to analyze IR spectroscopy to determine secondary structures of proteins. In the context of the study, ridge models were built using IR spectra with corresponding known crystal structures. Part of the intent of the study was to determine which secondary structure elements or other quantities are predictable from IR spectra. Information about the contents of certain types of hydrogen bonds and amino acid composition was extracted. Ridge regression performed well in the context of the study (Rahmelow and Hübner, 1996).

Lasso

Least absolute shrinkage and selection operator (lasso) regression was introduced by Tibshirani (1996). The lasso provides a sparse model by shrinking some coefficients and setting most other coefficients to zero. This has been referred to as the sparsity principle (Chun and Keleş, 2010). Under this principle, it is assumed that a smaller subset of the predictor variables is driving the prediction results. Thus, other coefficients can be excluded from the model (i.e., set to zero) with no significant performance loss. This reduces a large, largely uninterpretable model to a sparse, more interpretable model (Tibshirani, 1996).

The lasso is related to backward-stepwise selection, which "starts with the full model and sequentially deletes the predictor that has the least impact on fit" (Hastie *et al.,* 2009). Although backward stepwise-selection produces a sparse interpretable model, it does not necessarily provide a model with reliable predictive power. This is because coefficients are dropped from the model in an iterative process, and they are not reconsidered for inclusion in the model. Thus, coefficients that are important in the model might be mistakenly dropped early in the process. Ridge regression, on the other hand, is a continuous process that shrinks the coefficients, so it is more stable. As noted previously, ridge does not set any coefficients equal to zero, so it does not produce a sparse model. The lasso combines the desirable features of backward-stepwise selection and ridge regression to provide a sparse, relatively stable model.

The lasso penalty is used to select specific channels (wavelengths) for each element that explain the most variance in its predicted concentration. It shrinks some coefficients and sets other, less influential coefficients to zero. In this model, y_i is the elemental abundance in sample i and x_{ij} is the intensity at wavelength j of sample i (Dyar *et al.*, 2012; Ozanne *et al.*, 2012).

The lasso penalty is defined as

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, subject\ to\ \sum_{j=1}^{p} |\beta_j| \leq t. \qquad (5)$$

The shrinkage parameter t is defined such that $t \geq 0$. As in ridge regression, t controls the amount of shrinkage (i.e., sparsity) in the model. Higher values of t correspond to greater shrinkage.

Although the lasso provides a sparse model, it does not perform well in the $p >> N$ case. Also, for a group of highly correlated variables, the lasso indiscriminately chooses a variable from that group, leaving out potentially valuable information for model predictions.

Lasso regression has been used in a variety of spectroscopy applications. Menze et al. (2004) applied the lasso, among other classification methods, to classify magnetic resonance spectroscopy (MRS) data. Specifically, they were using MRS to detect recurrent tumors after radiotherapy. In vivo magnetic resonance spectra have highly correlated spectral channels, poor signal-to-noise ratios, and fall into the $p >> N$ case. The goal was to find a reliable automated

method to render MRS results accessible to radiologists who lacked the training to interpret the spectra themselves. PLS and ridge regression were also investigated. These regression techniques worked well for this application (Menze *et al.*, 2004). The lasso has also been used to predict elemental compositions for geological samples, as in the context of this thesis (Dyar *et al.*, 2012).

Generalizing the Ridge and Lasso

The ridge and lasso penalties can be generalized as follows

$$\tilde{\beta} = argmin_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|^q \right\} \; for \; q \geq 0. \quad (6)$$

This is known as an L_q penalty. For $q = 0$, this reduces to subset selection, where the penalty restricts the number of non-zero parameters. For $q = 1$, this reduces to the lasso; $q = 2$ corresponds to the ridge penalty. These are the L_1 lasso penalty and the L_2 ridge penalty, respectively. Note that for $q \leq 1$, the constraint region is not convex and the prior is not uniform in direction. This makes solving the optimization problem more difficult (Hastie, *et al.*, 2009).

Given this generalization, it may seem reasonable to estimate q based on the data. Other values of q may work better than $q = 0$, 1, or 2. However, extra variance is generated during this estimation process. The improvement in model fit has not been observed to be great enough to justify the additional variance (Hastie, *et al.*, 2009). Thus, other L_q penalties will not be explored.

Fused Lasso

Fused lasso regression was introduced to address shortcomings of the lasso technique (Tibshirani and Saunders, 2005) while preserving other desirable features. For example, the lasso performs relatively poorly when the number of observations per sample significantly outnumbers the number of samples (i.e. $p >> N$), whereas the fused lasso is particularly useful in the $p >> N$ case. Also, the lasso ignores natural ordering of data features, i.e. it indiscriminately chooses a feature from a cluster of influential features. Thus, the model does not necessarily reflect the features that have the most bearing on compositional predictions. The fused lasso, on the other hand, takes advantage of natural ordering of data features when it selects its coefficients.

The coefficients of the fused lasso model are obtained using the following conditions

$$\hat{\beta}^{fused} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$subject\ to\ \sum_{j=1}^{p} |\beta_j| \leq s_1 and\ \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq s_2. \quad (7)$$

The first constraint ensures that the number of coefficients in the model does not exceed some specific number, and the second constraint ensures that no single coefficient significantly outweighs another coefficient. As in the other models, y_i is the elemental abundance in sample i and x_{ij} is the intensity at wavelength j of sample i.

The fused lasso has been used for feature selection in protein mass spectroscopy (Tibshirani and Saunders, 2005). It has yet to be used in other forms of spectral analysis. Its mathematical relationship to the lasso makes it a reasonable candidate in the context of this thesis. This thesis is the first application of fused lasso to LIBS spectroscopy.

Elastic Net

Elastic net regression is a hybrid of lasso and ridge regression. It retains the sparse properties of lasso regression and the stability of ridge regression in the $p >> N$ case. Much like the fused lasso, it can also select groups of correlated variables. "It is like a stretchable fishing net that retains 'all the big fish'" (Zou and Hastie, 2005).

The model coefficients have the following form subject to choices of α and t constraints; y_i is the elemental abundance in sample i and x_{ij} is the intensity at wavelength j of sample i

$$\hat{\beta}^{elastic.net} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2,$$

$$subject\ to\ \sum_{j=1}^{p} \left(\alpha\left|\beta_j\right| + (1-\alpha)\beta_j^2\right) \leq t\ . \qquad (8)$$

The second term of the elastic net penalty promotes averaging of highly correlated features. Thus, unlike the lasso, which indiscriminately selects a feature from a cluster of highly correlated features to represent in the model, the elastic net retains information about all the features in the cluster by averaging them. The

first term in the penalty controls the sparse nature of the model; it promotes a sparse solution in the feature coefficients (Hastie *et al.,* 2009).

The elastic net penalty is a convex combination of the L_1 and L_2 penalties. It retains the desirable properties of both the lasso and the ridge. Note that when $\alpha = 1$, the penalty becomes a ridge penalty. Similarly, when $\alpha = 0$, the penalty becomes a lasso penalty. The elastic net is viewed as a generalization of the lasso that performs better in the $p >> N$ case and has the ability to execute grouped selection.

The elastic net has been used for classification of bacterial Raman spectra from different growth conditions. "The experiment was designed to be a preliminary investigation of the ability of the Raman system to reliably distinguish between the same bacterial strain grown in different environmental conditions, specifically the presence of chromate in the media" (DePalma and Stephen, 2011). In the context of this study, the cross-validated classification rate ranged from 0.98-1.00 depending on the choice of α (DePalma and Stephen, 2011).

Although the elastic net has been employed for classification in spectroscopy applications, it has not been used extensively in previous spectroscopy analyses for composition predictions. It is a viable candidate, however, because of its mathematical relationship to ridge and lasso regression. Both of the latter techniques have been used extensively for analysis of spectral

data. As a mathematical improvement on them, it is likely that the elastic net will perform at least as well for spectral applications.

Fused lasso versus Elastic Net

At first glance the elastic net and the fused lasso appear very similar. Both the elastic net and the fused lasso give sparse models and perform well in the *p* >> *N* case. They both address the shortcomings of the lasso by selecting groups of highly correlated variables for inclusion in the model instead of arbitrarily choosing a variable from such a group for inclusion in the model.

The first terms in both the elastic net and the fused lasso penalties control the sparsity of the coefficients to provide a parsimonious model. Thus, the bolded part of the elastic net penalty (Equation 8), $\sum_{j=1}^{p}\left(\boldsymbol{\alpha}\left|\boldsymbol{\beta_j}\right| + (1-\alpha)\beta_j^2\right) \leq t$, and the s_1 fused lasso penalty (Equation 7), $\sum_{j=1}^{p}\left|\beta_j\right| \leq s_1$, perform the same function. The elastic net penalty uses two constraints, α and t, where as the fused lasso achieves sparsity according to one constraint, s_1. The α value controls the balance between averaging correlated features and the number of non-zero coefficients (Ozanne et al., 2012).

The second term in the elastic net penalty (Equation 8), $\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right| + (\boldsymbol{1-\alpha})\boldsymbol{\beta_j^2}\right) \leq t$, promotes averaging of highly correlated features. The s_2 penalty in the fused lasso (Equation 7) $\sum_{j=2}^{p}\left|\beta_j - \beta_{j-1}\right| \leq s_2$ controls the magnitude of the differences of the coefficients. These two penalties are quite different. The

elastic net penalty allows the model to select groups of highly correlated features and average them. This ensures that information from the cluster is not lost, but that it is also condensed into one variable coefficient to preserve the sparse nature of the model. The fused lasso penalty does not average the highly correlated features. Instead, it calculates the difference between two adjacent features and shrinks this difference according to some specified penalty s_2. This assumes that there is a natural ordering in the features, and it inherently preserves that natural ordering when it calculates the model coefficients.

The elastic net penalty is more widely applicable than the fused lasso penalty because it does not assume that the features have a natural ordering. However, because it averages highly correlated features, it could significantly reduce the perceived impact of a given feature if the correlated features carry much less weight. This may make model interpretation more difficult.

Sparse Partial Least Squares

Sparse partial least squares (SPLS) regression was introduced by Chun and Keleş in their 2010 paper. This regression technique was motivated by recent advancements in biotechnology. It has gained popularity as a means for analyzing high dimensional genomic data. Because of its dimension reduction capabilities, PLS drew attention first as a means of analysis for this type of data. PLS does not provide a sparse model, however, so it sacrifices straightforward interpretability.

Like the lasso and the elastic net, SPLS operates under the fundamental assumption that some small subset of variables is primarily responsible for determining the response, known as the sparsity principle (Chun and Keleş, 2010). Thus, SPLS also has the ability to perform feature selection, while it retains the stable properties of PLS.

SPLS adds an L_1-constraint to the formula given for PLS in the previous section. In PLS, "typically all or a large portion of the variables contribute to the final direction vectors which represent linear combinations of the original predictors" (Chun and Keleş, 2010). Chun and Keleş argue, however, that a large number of irrelevant variables actually contribute noise to the PLS prediction. Applying an L_1 (sparsity) constraint eliminates these variables from the model to improve predictions. SPLS computes the model coefficients using the following formulation.

$$min_{w,c}\{-\kappa w^T M w + (1-\kappa)(c-w)^T M(c-w) + t_1|c|_1 + t_2|c|_2^2\},$$

$$subject\ to\ w^T w = 1. \qquad (9)$$

In this formula, $M = X^{\mathrm{T}}YY^{\mathrm{T}}X$. SPLS has four tuning parameters (κ, t_1, t_2, and K). The L_1-penalty (t_1) encourages sparsity on c where c is a direction vector; this is the thresholding parameter. The L_2-penalty (t_2) addresses the potential singularity (inability to invert) in M while solving for c. The K parameter controls the number of hidden components. The κ parameter is responsible for the starting value for the SPLS algorithm (Chun and Keleş, 2010).

SPLS (in addition to the elastic net) was evaluated for classification of bacterial Raman spectra from different growth conditions. It returned a cross validated average classification rate of 0.809 ± 0.032. It did not perform as well as the elastic net did for this experiment. Also, interpretation of important features was more complicated than for the elastic net case because, as with PLS, SPLS does not use the feature matrix directly, but rather principal components computed using the matrix (DePalma and Stephen, 2011).

Like elastic net regression, SPLS is a relatively new technique, so it has not yet had a chance to be tested in spectroscopy for prediction of elemental compositions. Given its relationship to PLS and its ability to produce parsimonious models, SPLS is a reasonable choice for spectral analyses.

Table 1. Summary table of the various regression methods explored in this thesis

Method	**OLS**	**PLS**	**LASSO**	**ELASTIC NET**	**SPLS**
Summary		Chooses subspaces from explanatory matrix, **X**, sequentially projects response vector, **y**, onto the subspaces of the column spaces of **X** to determine the model variable coefficients.	Shrinks some coefficients and sets others equal to zero in accordance with shrinkage parameter, *t*. Provides a sparse model that can be used for both feature selection and composition predictions.	Generalizes the lasso. Shrinks some coefficients and sets others equal to zero; averages highly correlated features and shrinks averages. Provides a sparse model that has more terms than the lasso and can be used for feature selection and composition predictions.	Adds an L_1-constraint to the PLS algorithm to impose sparsity. Performs feature selection yet retains stable properties of PLS.
Penalty	None	$w^T(I_p - W_{k-1}W_{k-1}^+)w = 1,$ $w^T\Sigma_{XX}w_j = 0 \; for \; j = 1, \dots, k-1$	$\sum_{j=1}^{p}\lvert\beta_j\rvert \le t$	$\sum_{j=1}^{p}\left(\alpha\lvert\beta_j\rvert + (1-\alpha){\beta_j}^2\right) \le t$	$w^Tw = 1, \lvert w\rvert \le t$
Tuning Parameters	None	*k*, # of dimensions	*t*	*t*, α	κ, t_1, t_2, *K*
Advantage(s)		Provides a stable model.	Provides an interpretable model, selects subset of predictors with the strongest effects on the response variable. Can be used for feature selection when less data are available.	Performs well in the $p >> N$ case. Provides an interpretable model that is more stable than the lasso. Useful for feature selection.	Performs well in the $p >> N$ case.
Disadvantage(s)	Does not provide a stable model when the variables are correlated; predictions suffer from wild variance so they are unreliable.	Provides a complex model in which all coefficients are linear combinations of the original channels. Involves a complex optimization problem with no simple, closed-form representation.	Does not perform as well in the $p >> N$ case. Arbitrarily chooses one covariate from a group of highly collinear covariates to use in the model and discards the rest.	Cannot be used for feature selection in situations when less data are available because it overwhelms the data with too many model variables.	Shrinkage properties not yet well understood.

Model Selection

Model selection is critical in statistical analyses. One important point of comparison is the accuracy of the model. A good model minimizes the prediction error. Mean squared error (MSE) is often used as a measure of the overall size of the measurement error (Rice, 2006). MSE for a model coefficient has the following form:

$$MSE = E\left[\left(\hat{\beta} - \beta\right)^2\right] = Var\left(\hat{\beta}\right) + bias\left(\hat{\beta}\right),$$

$$where\ bias\left(\hat{\beta}\right) = E\left(\hat{\beta}\right) - \beta. \quad (10)$$

In these formulae, $\hat{\beta}$ is the calculated model coefficient and β is the true parameter.

Bias can be thought of as the extent to which the model "memorizes" the training set. Generally, bias has two components: model bias and estimation bias. Model bias is the difference (error) between best-fitting linear approximation and the true function. Estimation bias is the difference (error) between the average estimates of the model components ($x\hat{\beta}$) and the best-fitting linear approximation. For shrunken regression techniques, there is an additional estimation bias because the models are not best-fitting linear approximations; they are good approximations (Hastie *et al.*, 2009). Variance describes the flexibility that the model has to deviate from the training set when it calculates the model coefficients.

When the MSE is minimized, so are the bias and the variance of the model. Thus, a reasonable bias-variance tradeoff can be achieved for each of the various shrunken regression methods. Prediction error results for shrunken regression methods are often reported as root mean squared errors of prediction (RMSEP) because these have the same units as the original measurements of sample compositions, which in this thesis are expressed as wt% oxides.

Although point-value RMSEPs are good first approximations of model performance, they do not take into account their variability. Both model stability and MSE must be taken into account when identifying a superior method (Hothorn *et al.*, 2005). Both of these components can be examined using benchmark experiments. "Benchmarking" has its origins in land surveying, where the original meaning is: "A benchmark in this context is a mark, which was mounted on a rock, a building, or a wall. It was a reference mark to define the position or the height in topographic surveying or to determine the time for dislocation" (Patterson, 1992). Similarly, the performance of an algorithm can be assessed by "standing" on a reference point (the point-value RMSEP) and assessing the variability from that point.

In "real world" applications where there exists a single learning sample *L* with a distribution $\hat{\mathbb{Z}}_n$with no dedicated independent test sample, benchmark experiments require several steps. Models are trained on samples from $\hat{\mathbb{Z}}_n$, which is the distribution of empirical data and model performance is assessed using

samples from $\hat{Z}$. Thus, for each model fitted on a bootstrap sample (sample generated from the original data through resampling), the original learning sample *L* is used as a test sample. This approach leads to overfitting because the same sample(s) can be present in both test and training sets. This can be corrected using out-of-bootstrap observations, where the test sample can be defined in terms of L/L^b, where L^b *(b = 1, …, B)* is a bootstrap sample. This can also be handled using cross-validation techniques, where each bootstrap sample L^b is divided into *K*-folds and the performance is defined as the average of the RMSEP of each fold (Hothorn *et al.*, 2005; Eugster *et al.*, 2008). This method operates as described in the cross-validation section of this thesis.

For this model comparison, the null hypothesis is that there is no difference between the models being compared. In the context of this thesis, the null hypothesis states that there is no difference in the performance of the PLS, lasso, elastic net, and SPLS models. Their RMSEP values and corresponding variances are assumed to be equivalent. For each comparison, a *p*-value is computed. Assuming that the null hypothesis is true, the *p*-value is the probability of getting a test statistic at least as extreme as the observed value. Prior to completing calculations, the researcher defines an α value, which determines the level of significance for the *t*-test. Common α values are 0.05 and 0.01. When the *p*-value is less than the defined significance level, the null hypothesis is rejected (Rice, 2007). The Bonferroni correction may be used to counteract the problem of multiple comparisons, which happens when a set of statistical tests are considered

simultaneously, all with significance level α. This significance level is not valid for the set of all comparisons even if it is appropriate on an individual level; it is lowered to α/n, where n is the number of statistical tests being performed (Weisstein, 2012).

When choosing a model, interpretability, sparsity, and stability are also points for consideration. The ridge penalty, for example, provides a robust model. This is important because it means the prediction results will be consistent over time. However, the ridge penalty does not yield a sparse model, which makes model interpretation challenging at best. PLS is similar in that it produces a robust, but saturated model. In contrast, the lasso, elastic net, and fused lasso penalties yield sparse outputs by performing automatic feature selection. From a statistical point of view, a parsimonious model is always desirable because it is more easily interpreted. The latter two penalties also result in relatively stable models.

The amount of data available is also an important consideration. The lasso and elastic net penalties are both useful for automatic feature selection, but model choice is dependent on the amount of data available. When data are plentiful, the elastic net is preferable because it includes more information from the data and the model is more stable. Because of the averaging effect in the elastic net penalty, more nonzero coefficients are produced in the model, which leads to greater model stability. In cases where the amount of data is lacking, the lasso can be used perform feature selection without overwhelming the data with too many

model variables because it provides a sparser model than the elastic net (Ozanne *et al.*, 2012).

Cross-Validation

Because the goal of this thesis is to draw conclusions about the usefulness of various statistical methods to LIBS, it is necessary to develop meaningful ways to compare and quantify differences among models. For this purpose, cross-validation is one of the simplest and most widely used methods for estimating prediction error. This method directly estimates the expected extra-sample error, which is the average generalization error when the method $\hat{f}(X)$ is applied to an independent test sample from the joint distribution of *X* and *Y* as shown below (Hastie *et al.*, 2009).

$$Err = E\left[L\left(Y, \hat{f}(X)\right)\right] \tag{11}$$

$E[Z]$ calculates the expected value, where $Z = L\left(Y, \hat{f}(X)\right)$, which is the loss function. Use of the expected value of a loss function is a standard way to calculate prediction error.

Ideally, models would be trained on some portion of the data, known as the training set, and then tested on a held-out, unseen portion of the data, known as the test set. In many cases, however, there is not enough data to have two completely separate data sets. In such instances, *K*-fold cross-validation is used. *K*-fold cross-validation splits the data set into *K* approximately equal-sized parts.

When models are being fit for a sample in K_i, the other K-1 folds (all K_j folds, $i \neq j$) are used to train the model and the K_i fold is used to test the model.

The formula for the cross-validation estimate of prediction is as follows, where κ: $\{1, \ldots, N\} \rightarrow \{1, \ldots, K\}$ is an indexing function that shows the fold to which observation i is allocated through randomization (Hastie et al., 2009).

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right) \quad (12)$$

For "leave-one-out" cross-validation, $\kappa(i) = i$.

Common K values are five and ten (Hastie et al., 2009). In the special case when $K = N$, this is known as "leave-one-out" cross-validation. In this case, the model fit is computed using all of the data except the i^{th} observation. Through this method, there is enough data to train the model sufficiently, but there is no danger of validating the model using the same data on which it was trained, which would yield an artificially low model prediction error.

For example, in a data set of 100 samples, using 10-fold cross-validation means that there will be ten folds, each containing ten samples (because ten divides evenly into 100). Suppose a model is being built to predict the composition of ten samples and suppose that the data for these samples are contained in the tenth fold. The model will be trained on the data from the first nine folds, and then it will be tested on the data from the tenth fold to generate model predictions for the compositions of these ten samples. This process will be

repeated for the remaining nine folds. Composition predictions for all 100 samples will be calculated.

In such "leave-one-out" cross-validation, the cross-validated estimator is unbiased for the true prediction error, which is shown in the following formula.

$$Err_{\tau} = E_{X^0,Y^0}\left[L\left(Y^0, \hat{f}(X^0)\right)|\tau\right] \quad (13)$$

Although the estimator is unbiased, it can have high variance because the N "training sets" (individual samples) are very similar to one another. If N is sufficiently large, "leave-one-out" cross-validation can also be computationally expensive (Hastie *et al.*, 2009).

In a data set of 100 samples, using "leave-one-out" cross-validation means that there will be 100 folds, each containing a single sample. When the model is being built to predict the first sample, for example, the first sample will be left out of the training set, which will contain the other 99 samples. Then the model will be tested on the first sample to generate a model prediction for the composition of that sample. The same process will be repeated until predictions have been generated for all 100 samples.

For 5-fold and 10-fold cross-validation, a bias-variance tradeoff is employed. The variance is lower than in "leave-one-out" cross-validation. However, depending on how the model training varies as a function of training set size, bias could be a problem. Expected error (*Err*) shown in the following

formula is well estimated using 10-fold cross-validation. Expected error is calculated by averaging the true error over the training sets, τ.

$$Err = E_{\tau}E_{X^0,Y^0}\left[L\left(Y^0, \hat{f}(X^0)\right)|\tau\right] \quad (13)$$

There is too much variation in the true error, Err_{τ}, across training sets for it to be estimated effectively. This variation gets averaged out in the expected error to provide reliable prediction error values for the models in question.

If data are sufficiently plentiful, using a test set is preferable to cross-validation because this would eliminate a source of variance (cross-validation procedure) for calculating model error. In situations where data are insufficient, however, use of a test set instead of cross-validation can increase model error. This distinction is undoubtedly application dependent and "plentifully large" has not yet been defined for LIBS. It is conservative to err on the side of cross-validation, as is done in this thesis.

Feature Selection

Shrinkage penalties that result in sparse models are useful for feature selection. Feature selection (also known as variable selection) is a machine learning term that refers to choosing a subset of variables that exert the most influence in making model predictions. High dimensionality (i.e., $p >> N$) results in extremely complex models where overfitting of the data is a concern (Hastie et al., 2009). Model overfitting gives overly optimistic prediction results. An

overfitted model will give less accurate predictions when tested on unseen data. Variable selection mitigates the “curse of dimensionality” and results in more interpretable models. Several shrunken regression methods, including lasso, fused lasso, elastic net, and SPLS can be used to perform feature selection.

The ability to perform feature selection is an important aspect of a desirable model. It pertains to two of the three qualities previously mentioned: interpretability and sparsity. In the context of this thesis, feature selection ensures that the model coefficients have physical significance; they correspond to wavelength channels. It also promotes sparsity because only the most influential features are included in the model; others are driven to zero. Thus, a parsimonious, physically meaningful model results.

METHODS

Samples and Experimental Methods

A suite of 100 igneous rocks was analyzed using LIBS at Los Alamos National Laboratory. Approximately 150 g of sample was crushed to particle sizes of <45 µm. This was about ten times smaller than the LIBS beam diameter. This was done to minimize sample inhomogeneity and equalize grain size and porosity to allow the most accurate composition measurements. For LIBS sample preparation, a few grams of each sample were pressed into pellets in an aluminum cup using 35 tons of pressure to further mitigate inhomogeneity and equalize porosity (Tucker et al., 2010).

Because atmospheric pressure exerts known effects on LIBS spectra, samples were placed in a chamber filled with 7 torr CO_2 to simulate the Mars surface pressure. They were ablated from a standoff distance of 9 m using a 1064-nm Nd:YAG[3] laser operating at 17 mJ/pulse; 50 laser shots were taken per sample. The optical emission from resultant sample plasma was collected using three Ocean Optics HR2000 spectrometers with UV (223-326 nm), VIS (328-471 nm), and VNIR (495-927 nm) wavelength regions (Tucker et al., 2010). A resultant spectrum is shown in Figure 10. Elemental oxide concentrations for the ten major oxides were measured from samples prepared as noted above using the

[3] Nd:YAG stands for neodymium-doped yttrium aluminum garnet.

X-ray Fluorescence (XRF) Laboratory at the University of Massachusetts, Amherst, supervised by Dr. Michael Rhodes (Rhodes and Vollinger, 2004).

Statistical Analysis

Wavelength calibration was performed for each of the three spectral regions using the NIST database. In each of the three wavelength regions (UV, VIS, and VNIR), 10-15 prominent major elemental/ionic emission lines were selected from the NIST database. These were matched to spectral peaks in the sample spectra, which were identified in terms of pixel number (pixel number is

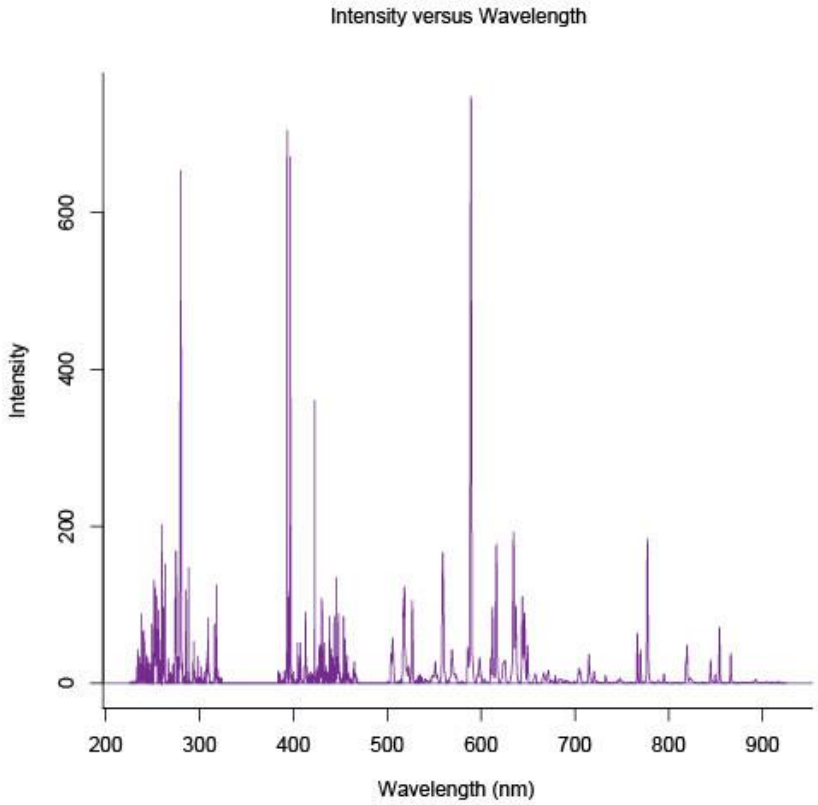

Figure 10. Averaged and baseline subtracted sample spectrum from the 100 sample suite of igneous rocks. Wavelength regions span UV, VIS, and VNIR.

proportional to wavelength). The NIST wavelength value was plotted against the pixel number using a quadratic fit (Tucker, *et al.*, 2010). A more accurate fit would use a third order polynomial of the form:

$$\lambda_p = I + C_1 p + C_2 p^2 + C_3 p^3 \quad (14)$$

In this equation, λ is the wavelength value of pixel *p*, *I* is the wavelength of pixel 0, C_1 is the first coefficient (nm/pixel), C_2 is the second coefficient (nm/pixel^2), and C_3 is the third coefficient (nm/pixel^3). This calibration was necessary because the wavelength for all spectrometers drifts slightly as a function of time and environmental conditions (Ocean Optics: Installation and Operation Manual).

The 50 spectra for each sample were averaged and smoothed, and the baseline (A-D offset, and ambient light background) was modeled and subtracted using the R package "Peaks" (Ozanne *et al.*, 2012) (Morhac, 2008). Spectra were averaged to remove noise. They are smoothed to match up the wavelength axes among the 100 spectra. These preprocessing steps were performed using the CRAN R package "hyperSpec" (Beleites and Sergo, 2012). This package deals specifically with data for spectroscopic analysis.

The R package "caret" was used to create one set of CV folds that were used to generate a list of training sets to build each of the models. This ensured that all of the models were being trained and tested on exactly the same data folds. This package was also used to create grids of tuning parameters for each of the regression techniques. As the models were built, they cycled through these tuning

parameters and ultimately selected the best model according to the one SE model selection criterion described in the next section. For example, ten possible *t* values were tested for the lasso, ranging from 0.1 to 0.9. For the SiO_2 model, the model with $t = 0.633$ was selected.

The "caret" package was also used as a wrapper package for the various CRAN R packages needed to construct the shrunken regression models. This permitted functional code to be written in the same form for all the regression techniques. Within this wrapper code, the PLS models were constructed using the CRAN R package "pls" (Mevik *et al.*, 2011). The CRAN R package "glmnet" was used to create the elastic net models (Friedman *et al.*, 2010). The lasso and SPLS models were built using the "elasticnet" and "spls" packages, respectively (Zou and Hastie, 2012) (Chung *et al.*, 2012). Sample code is available in the appendix.

Finally, "caret" can be used to compare various models using *t*-tests. This is a built in function of the package, and *t*-tests are a statistically robust test for whether significant differences in model performance are present. "Caret" completes these *t*-tests. Pair-wise model comparisons were made. For each pair-wise comparison, p-values were calculated. The level of significance used in this thesis was $\alpha = 0.05$.

Model Selection Heuristics

There are two model selection heuristics that are commonly used to select the best number of components (coefficients) for the shrunken regression model. "Best" is determined by MSE and, in relevant cases, sparsity of model coefficients. The same model selection heuristic is used across all models regardless of whether they are sparse to make the methods comparable.

A "global minimum" heuristic can be used to choose the number of components for the model. The global minimum is defined as the point at which the model MSE is at an absolute minimum. According to the definition of MSE, this is the point at which bias and variance of the model are minimized, so this should provide the best predictive model. This does not, however, ensure that the model will be as parsimonious as possible and will still yield good predictions.

In order to minimize MSE and make the models as parsimonious as possible, a "one standard error" (SE) heuristic was employed to choose the number of components for each of the models. This heuristic selects the component number that yields a mean square error (MSE) of prediction within one SE of the global minimum, as illustrated in Figure 11 (Ozanne et al., 2012). The intent of the one SE heuristic is to select a model with smallest number of components without sacrificing model performance (i.e., small MSE). This employs Occam's razor, which is a scientific rule that requires that the simplest of competing theories be preferred to the more complex theories (Merriam-Webster).

This also avoids overfitting, which is a potential danger of using the “global minimum” selection heuristic.

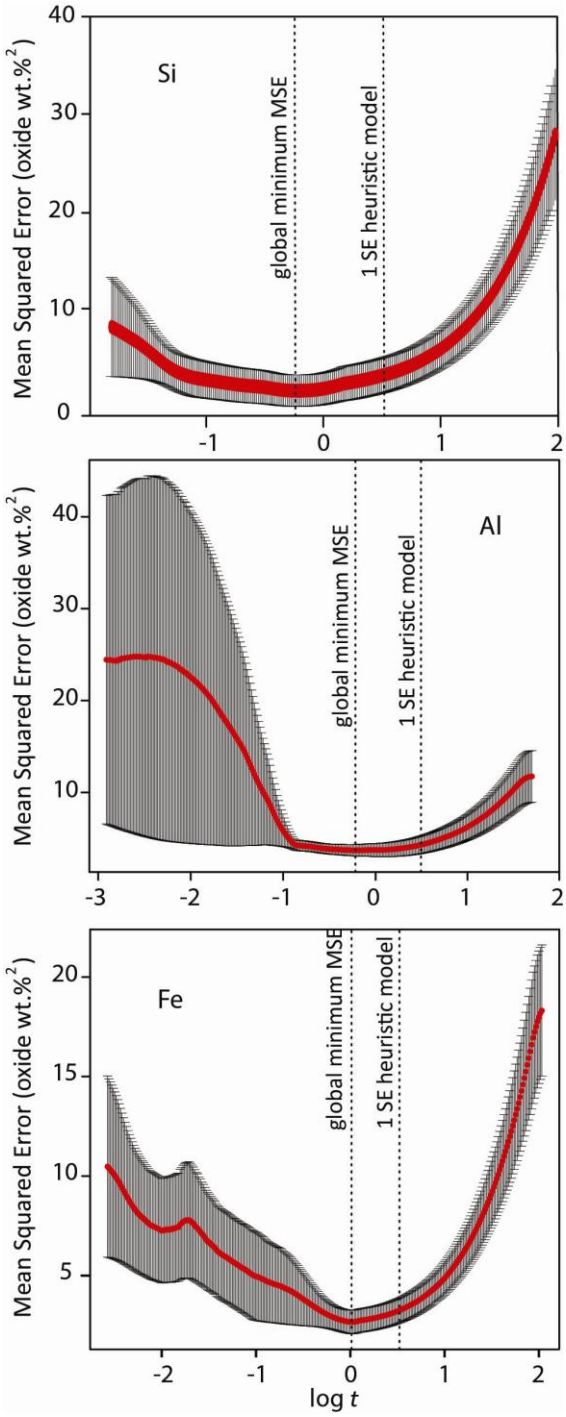

Figure 11. Mean square error (in units of wt% oxides squared) plotted against the log of the t value (the shrinkage parameter) for Si, Al, and Fe. The global minimum MSE on each plot indicates the value of *t* for which the smallest prediction error was obtained. The 1 SE heuristic model is the value of *t* used for predictions because this value of *t* strikes a balance between accuracy and model simplicity (Ozanne et al., 2012).

RESULTS

OVERVIEW

RMSEP values were used to compare model prediction accuracies for the PLS, lasso, elastic net, and SPLS regression models. Pair-wise comparisons of model RMSEP values using Student's *t*-test reveal that there is no statistical difference in the prediction accuracies of the four models for this suite of 100 igneous rocks. These results are shown in Table 7, which gives both the *p*-values and the absolute differences in RMSEP for the model comparisons.

MODEL SELECTION

Partial Least Squares

The number of components in the PLS model was constrained to 15 to prevent overfitting. The selected models for the ten major elemental oxides all used fewer than 15 components, as shown in Table 2. MgO had the most complex model with 14 components, while Na_2O had the most parsimonious model with 3

Table 2. Number of components chosen for the PLS models

Elemental Oxides	Number of Components
SiO_2	7
Al_2O_3	6
TiO_2	10
Fe_2O_3	12
MgO	14
MnO	12
CaO	7
K_2O	9
Na_2O	3
P_2O_5	10

A maximum of 15 components was possible. The number was constrained to prevent overfitting. The number of components was chosen using the one SE heuristic.

components. The other elements had components numbering from 6 to 12, with SiO_2 and Al_2O_3 on the lower end of that range, and TiO_2 and Fe_2O_3 on the upper end.

Lasso

The tuning parameter, *t*, was constrained to $0.1 \leq t \leq 0.9$ using the CRAN R package *caret*. Larger values of *t* would have permitted models with more coefficients. *Caret* constrains to a maximum value of 0.9 because larger values do not improve the predictions sufficiently to compensate for the lost model simplicity. These *t* values, shown in Table 3, minimize the number of components in the model and the model RMSEP values according to the one SE heuristic.

Table 3. Lasso: chosen tuning parameters

Elemental Oxides	***t***
SiO_2	0.633
Al_2O_3	0.278
TiO_2	0.544
Fe_2O_3	0.456
MgO	0.633
MnO	0.456
CaO	0.544
K_2O	0.544
Na_2O	0.189
P_2O_5	0.544

The range of the tuning parameter was constrained to $0.1 \leq t \leq 0.9$ using the CRAN R package *caret*. These *t* values minimize the number of coefficients in the models and minimize the MSE values, according to the one SE heuristic.

Elastic Net

For the elastic net models, the tuning parameters were constrained such that $0.1 \leq t \leq 3.0$ and $\alpha = \{0.1, 1.0\}$ using the CRAN R package *caret*. Models were selected according to the one SE heuristic. Most elemental oxide models selected $\alpha = 0.1$, which represents 10% L_1-penalty and 90% L_2-penalty. The

exception was SiO_2, which selected $\alpha = 1.0$. This is equivalent to a lasso penalty because it is 100% L_1-penalty. The MnO model had the smallest t value ($t = 0.1$); Al_2O_3, Fe_2O_3, MgO, and CaO had the largest t value ($t = 3.0$). The latter four models had the most coefficients.

Table 4. Elastic Net: chosen tuning parameters

Elemental Oxides	***t***	***α****
SiO_2	0.68	1.0
Al_2O_3	3.0	0.1
TiO_2	0.68	0.1
Fe_2O_3	3.0	0.1
MgO	3.0	0.1
MnO	0.1	0.1
CaO	3.0	0.1
K_2O	0.68	0.1
Na_2O	2.42	0.1
P_2O_5	0.68	0.1

The ranges of the tuning parameters were constrained to $0.1 \leq t \leq 3.0$ and $\alpha = \{0.1, 1.0\}$ using the CRAN R package *caret*. These values were chosen according to the one SE heuristic. * In the case were $\alpha = 1.0$, this model is equivalent to the lasso model with $t = 0.68$.

Sparse Partial Least Squares

For the SPLS models, tuning parameters were constrained using *caret*: $\eta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $K = \{1, 2, 3, 4, 5\}$. The η parameter is a combination of the t_1 and t_2 parameters; it controls the sparsity of the model. The sparser models are for Al_2O_3, TiO_2, and MgO. TiO_2 has the largest number of hidden components ($K = 4$). The third parameter was $\kappa = 0.5$, which simply dictates the starting point for the SPLS algorithm.

Table 5. SPLS: chosen tuning parameters

Elemental Oxides	***η****	***κ***	***K***
SiO_2	0.9	0.5	3
Al_2O_3	0.7	0.5	3
TiO_2	0.7	0.5	4
Fe_2O_3	0.9	0.5	2
MgO	0.7	0.5	3
MnO	0.9	0.5	2
CaO	0.9	0.5	3
K_2O	0.9	0.5	2
Na_2O	0.9	0.5	3
P_2O_5	0.9	0.5	2

The possible values of the tuning parameters were constrained to $\eta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\kappa = 0.5$, and $K = \{1, 2, 3, 4, 5\}$. These values were selected according to the one SE heuristic. *The η parameter is a combination of the t_1 and t_2 parameters discussed in the SPLS background section.

Table 6. Prediction errors for tuned models

	PLS		Lasso		Elastic Net		SPLS	
	RMSEP	SE	RMSEP	SE	RMSEP	SE	RMSEP	SE
SiO_2	2.39	0.80	**2.03**	0.70	2.05	0.65	2.50	0.61
Al_2O_3	**1.64**	0.41	1.87	0.71	1.84	0.71	**1.64**	0.64
TiO_2	**0.33**	0.09	0.35	0.12	0.36	0.12	0.37	0.14
Fe_2O_3	**1.30**	0.42	1.54	0.55	1.46	0.55	1.50	0.46
MgO	1.72	0.35	1.81	0.59	1.76	0.52	**1.70**	0.42
MnO	0.02	0.01	**0.02**	0.01	0.03	0.01	0.02	0.01
CaO	**0.80**	0.17	0.89	0.25	0.94	0.27	0.85	0.18
K_2O	0.45	0.19	0.47	0.16	**0.42**	0.19	0.50	0.22
Na_2O	0.66	0.27	0.72	0.24	0.69	0.25	**0.57**	0.12
P_2O_5	0.18	0.07	0.17	0.09	0.17	0.08	**0.17**	0.06

RMSEP and SE are in units of wt% oxides. Lowest RMSEP values for each element are in bold.

51

Table 7. Pair-wise comparisons of RMSEP differences and *p*-values

	RMSEP differences (upper diagonal) and *p*-values (lower diagonal)				
		PLS	Lasso	Elastic Net	SPLS
SiO_2	PLS		0.36	0.34	-0.11
	Lasso	1.00		-0.02	-0.47
	Elastic Net	1.00	1.00		-0.45
	SPLS	1.00	0.27	0.23	
Al_2O_3	PLS		-0.23	-0.20	0.00
	Lasso	0.64		0.03	0.23
	Elastic Net	1.00	1.00		0.20
	SPLS	1.00	0.29	0.498	
TiO_2	PLS		-0.02	-0.03	-0.04
	Lasso	1.00		-0.01	-0.02
	Elastic Net	1.00	1.00		-0.01
	SPLS	1.00	1.00	1.00	
Fe_2O_3	PLS		-0.24	-0.17	-0.21
	Lasso	0.72		0.07	0.03
	Elastic Net	1.00	0.42		-0.04
	SPLS	0.29	1.00	1.00	
MgO	PLS		-0.10	-0.04	0.01
	Lasso	1.00		0.06	0.11
	Elastic Net	1.00	0.93		0.05
	SPLS	1.00	1.00	1.00	
MnO	PLS		0.00	0.00	0.00
	Lasso	1.00		0.00	0.00
	Elastic Net	0.90	0.08		0.00
	SPLS	1.00	1.00	0.35	
CaO	PLS		-0.09	-0.13	-0.04
	Lasso	0.61		-0.05	0.04
	Elastic Net	0.07	0.57		0.09
	SPLS	1.00	1.00	1.00	
K_2O	PLS		-0.01	0.03	-0.05
	Lasso	1.00		0.04	-0.03
	Elastic Net	1.00	0.85		-0.08
	SPLS	1.00	1.00	0.38	
Na_2O	PLS		-0.06	-0.03	0.09
	Lasso	1.00		0.03	0.15
	Elastic Net	1.00	1.00		0.12
	SPLS	1.00	0.39	0.84	
P_2O_5	PLS		0.01	0.00	0.01
	Lasso	1.00		-0.08	0.00
	Elastic Net	1.00	1.00		0.01
	SPLS	1.00	1.00	1.00	

None of the p-values are smaller than the specified significance level $\alpha = 0.05$. The null hypothesis is not rejected; in the context of this data set, there is no statistically significant difference in performance among these four regression methods.

DISCUSSION

OVERVIEW

Despite the lack of statistical significance in the difference between models, some models do appear to give smaller errors depending on the element for which the models are constructed. This begs the question of whether the model choice should be made on an elemental basis. While this might improve prediction errors slightly, it would also complicate data analysis because the models would not be of the same type across elements. Also, because the perceived differences between models are so small, it is possible that the models that provide the smallest RMSEP values for this dataset would not apply to a different dataset.

However, the extent to which these conclusions can be generalized may depend on the size and composition of the data set. Although samples with major element weight percent oxide values ranging from 0-100% were chosen by geologists, there was no a priori reason why this suite of 100 igneous rocks should provide a comprehensive set of compositions. It is possible that results would change if a larger data set with even greater compositional variation was used, for example.

SiO_2

Although there is no statistical difference in model performance among the four regression methods explored here, the elastic net and lasso models seem to perform slightly better for SiO_2 composition predictions than PLS and SPLS do, as shown in Figure 12. Because Si emission lines are few but distinct in the UV-VNIR wavelength range, it is possible that the process by which PLS and SPLS

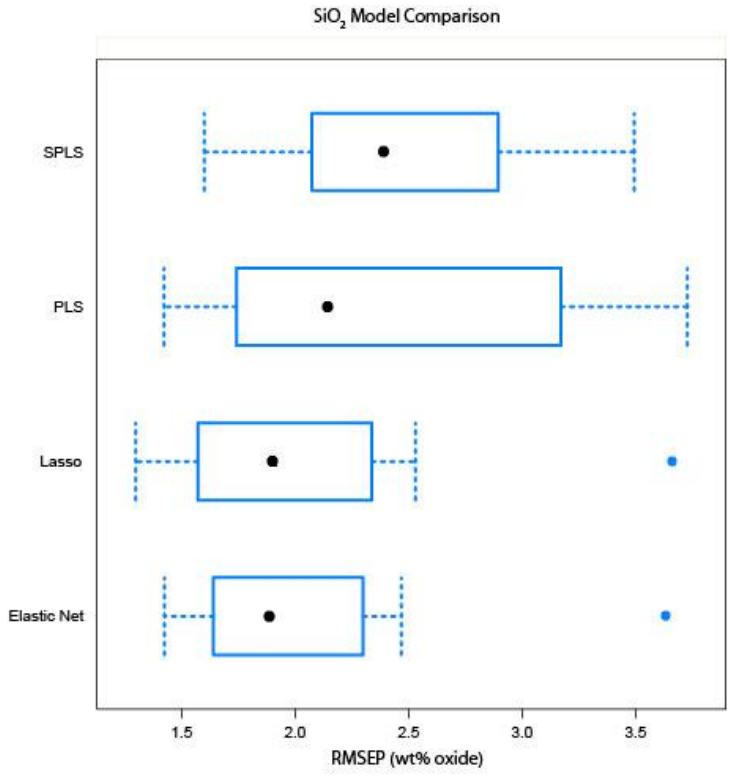

Figure 12. Box-whisker plot of the RMSE values for SiO_2 models. The elastic net model and the lasso model give the smallest prediction errors; the elastic net yields the smallest spread. The linear combinations seem to hurt the prediction capabilities for SiO_2 composition predictions for the PLS methods.

remove collinearity (taking linear combinations of channels) overwhelms the information inherent in the Si emission lines. In contrast, the lasso and the elastic

net methods select relevant channels according to their tuning parameters and drive others to zero. Presumably these models are selecting the Si emission lines, and these lines will not be overwhelmed by other irrelevant information.

Al_2O_3

All four models are statistically equivalent in their prediction accuracies. For Al_2O_3, SPLS and elastic net perform slightly better in terms of absolute RMSEP value, while SPLS and PLS perform better in terms of error spread, as

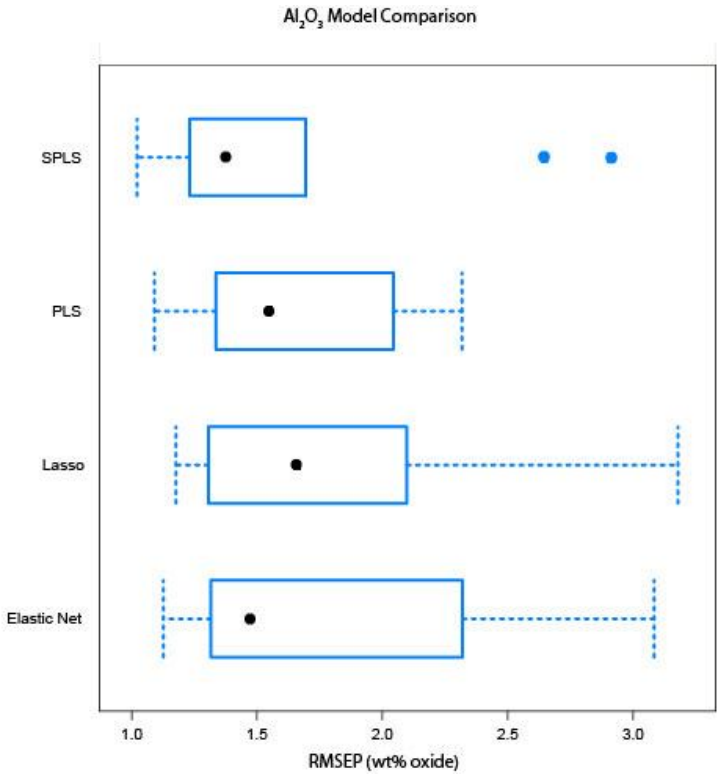

Figure 13. Box-whisker plot of the RMSEP values for Al_2O_3 models. In terms of RMSEP value, SPLS and elastic net perform slightly better than PLS and lasso. However, SPLS and PLS have smaller spreads for possible RMSEP values. All models are statistically equivalent.

shown in Figure 13. Like Si, Al does not have many emission lines in the UV-VNIR wavelength range. This may be why SPLS and elastic net perform slightly

better for Al. Also, Al emission lines may be highly correlated with other emission lines.

MnO

Although there is no statistically significant difference between any of the models for MnO predictions, there appears to be a substantial difference in RMSEP value for the lasso and elastic net models, as shown in Figure 14. It

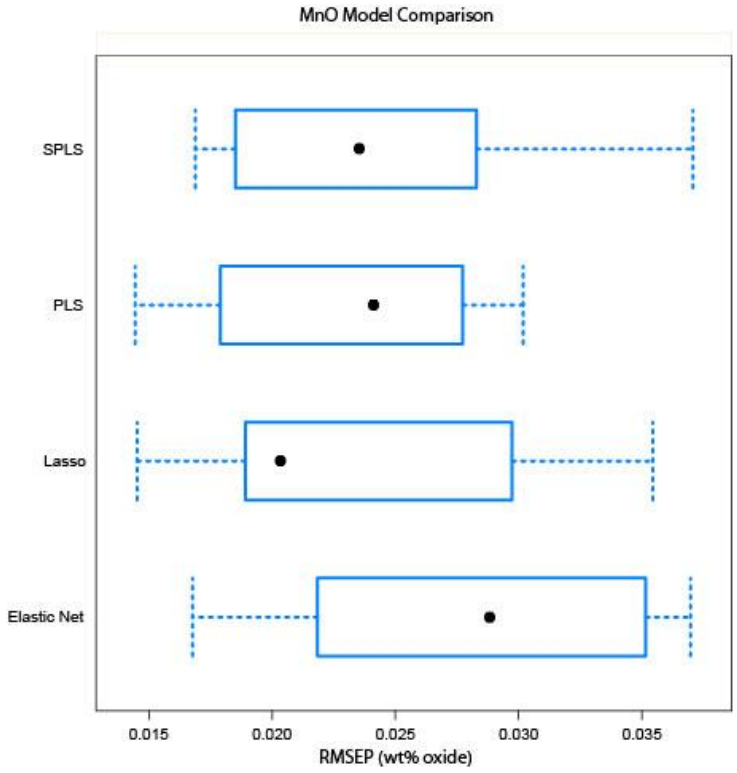

Figure 14. Box-whisker plot of the RMSEP values for MnO models. Although it may appear that there are large differences in the RMSEP values on this scale, the difference are not statistically significant. The error spread appears to be larger for sparse models.

should be noted that the *p*-value for this pair-wise comparison is 0.08 (Table 7), which is not significant at $\alpha = 0.05$, but it is still a much smaller *p*-value than those obtained for the other three pair-wise comparisons for MnO. At a slightly

more liberal significance level ($\alpha = 0.1$) the difference would be statistically significant. Geological samples tend to contain less MnO than some other major elemental oxides. The sparse nature of the lasso (which has fewer coefficients than the elastic net) may allow MnO emissions to be more prominent in the model and thus provide lower a lower RMSEP value. Given the RMSEP spread for all four models, however, it is difficult to say for certain that this is the case.

MgO

For the four MgO models, the RMSEP values are comparable, as shown in Figure 15. The model performances are statistically equivalent based on RMSEP

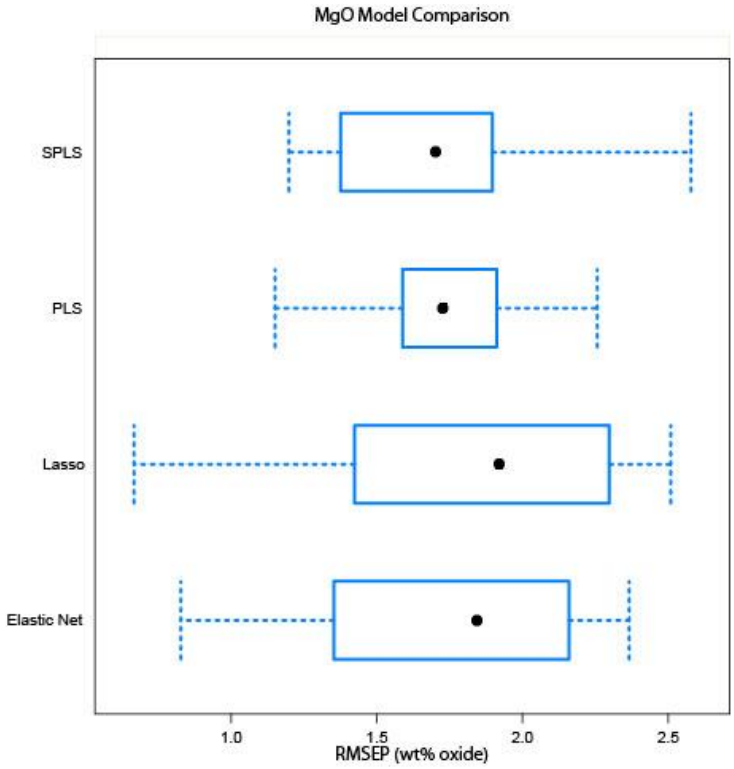

Figure 15. Box-whisker plot of the RMSEP values for MgO models. The error spread appears to be larger for sparse models. There is no statistical difference in performance among the four models.

values. It appears that the spread is a little tighter for PLS than for the other, sparse models. This suggests that some information is being lost when coefficients are dropped from the models, which is being retained by the linear combinations in PLS to give less variability in the RMSEP value.

CaO

The RMSEP values for the four models for CaO are comparable. PLS and SPLS appear to have slightly smaller error spreads, as shown in Figure 16. Although there is no statistical difference among the models, it is worth noting that the *p*-value for the PLS – elastic net pair-wise comparison was 0.069, which

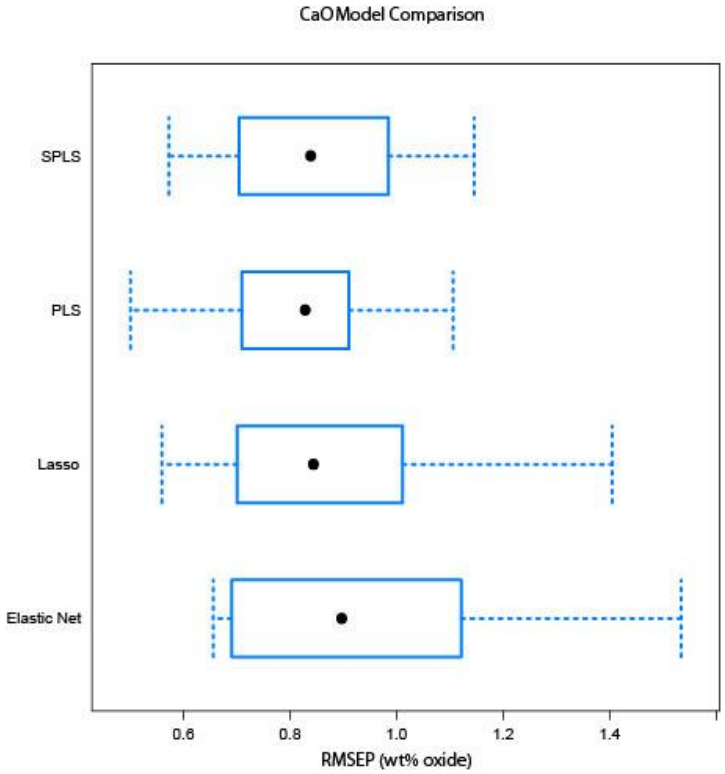

Figure 16. Box-whisker plot of the RMSE values for CaO models. The error spread appears to be slightly larger for sparse models. All models are statistically equivalent according to pair-wise comparisons using Student's t-test.

is close to the significance level of 0.05. This would be an interesting comparison to make with different samples. It is possible that PLS performs better than the sparse models because Ca has more emission lines so the lines retain weight in linear combinations of channels. Ca emission lines might be dropped or obscured in the other models due to high correlation with other lines or averaging effects.

Na_2O

For the prediction models for Na_2O, the sparse models seem to have less variability in their RMSEP values than PLS even though all four models are statistically equivalent. SPLS gives the smallest RMSEP, followed by elastic net

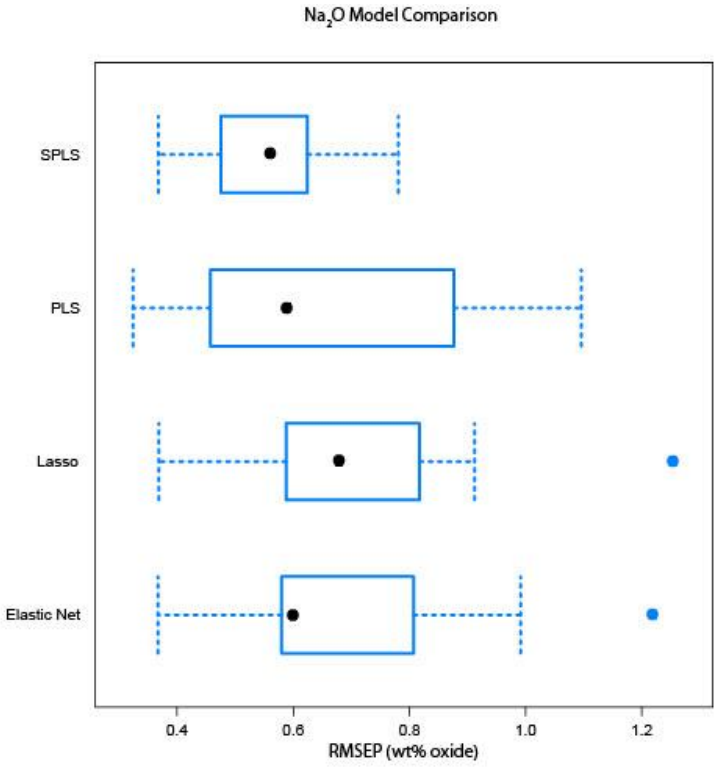

Figure 17. Box-whisker plot of the RMSEP values for Na_2O models. The sparse models appear to give slightly tighter spreads for the RMSEP values. All models are statistically equivalent.

and lasso, as shown in Figure 17. It is reasonable to assume that information from Na emission lines is retained in the sparse models and is more influential because these models have fewer coefficients to obscure the information and lead to increased RMSEP variability.

K_2O

All four regression models for K_2O are statistically equivalent. The error spreads are comparable as well, as shown in Figure 18. K has many emission lines, which may be why PLS performs well; the K emission lines remain dominant in the linear combinations that form the model coefficients. These results suggest

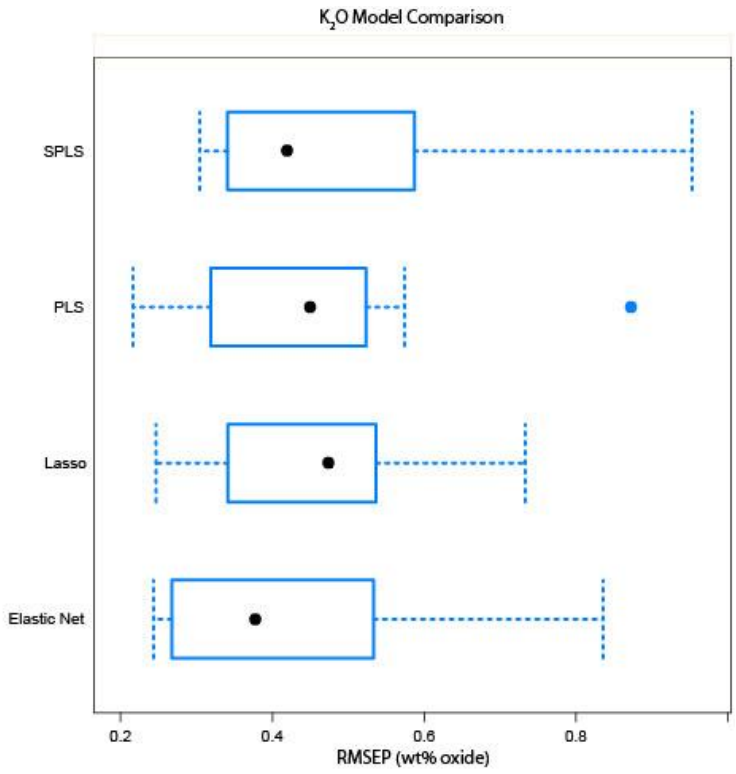

Figure 18. Box-whisker plot of the RMSEP values for K_2O models. There are no statistically significant differences between models. The RMSEP variability is relatively uniform across models.

that K lines are also sufficiently distinct to be selected for model inclusion by sparse techniques. This also implies that K lines are not highly correlated with other emission lines because the lasso performs comparably to PLS, so highly correlated lines do not appear to be indiscriminately dropped from the model. The lasso appears to retain K channels.

TiO_2

The four models for TiO_2 are statistically equivalent. In terms of RMSEP value, they are almost identical, but the PLS model has a much smaller spread than the sparse models, as shown in Figure 19. This suggests that Ti emission

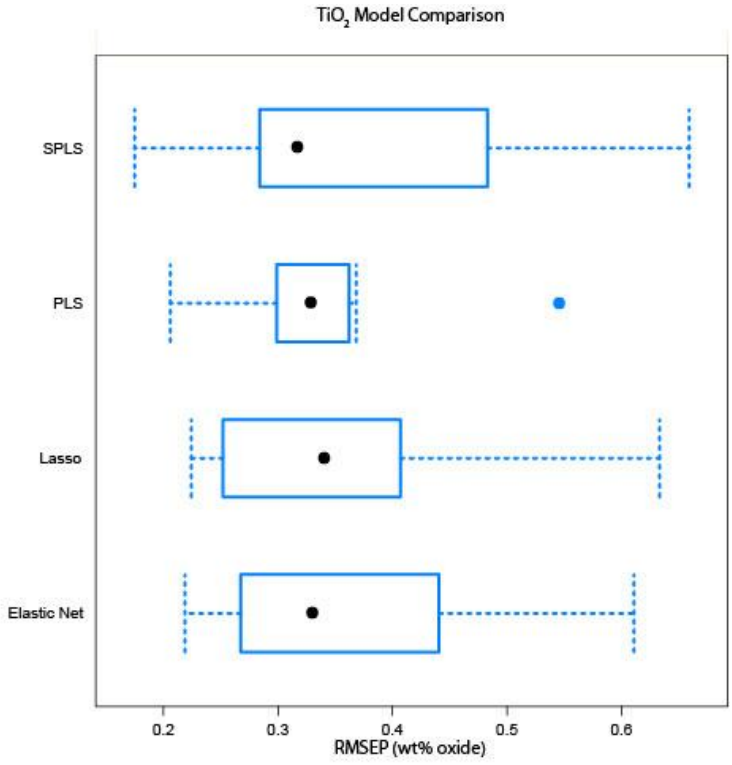

Figure 19. Box-whisker plot of the RMSEP values for TiO_2 models. Although all four models are statistically equivalent, the PLS model appears to have a tighter spread than the other, sparse models.

lines retain weight in the linear combinations of coefficients so as to produce a small RMSEP and reduce the variablity in RMSEP.

P_2O_5

RMSEP values and spreads are comparable for all four models for P_2O_5, as shown in Figure 20. All four models are statistically equivalent. These results suggest that P has plentiful, distinct emission lines that retain prominence both in the PLS model and the sparse models.

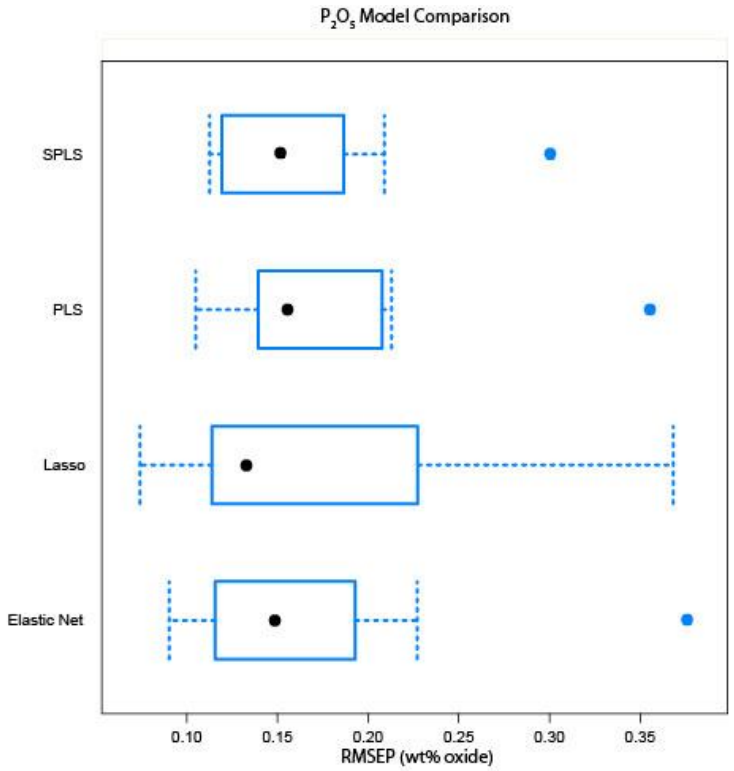

Figure 20. Box-whisker plot of the RMSEP values for P_2O_5 models. There are no statistically significant differences between the models. Model spreads appear to be comparable.

Fe_2O_3

As with the other major element oxides, the four regression models have statistically equivalent accuracies. The SPLS and PLS models, however, appear to have smaller RMSEP spreads than the lasso and elastic net models, as shown in Figure 21. PLS appears to perform slightly better for Fe than the other models,

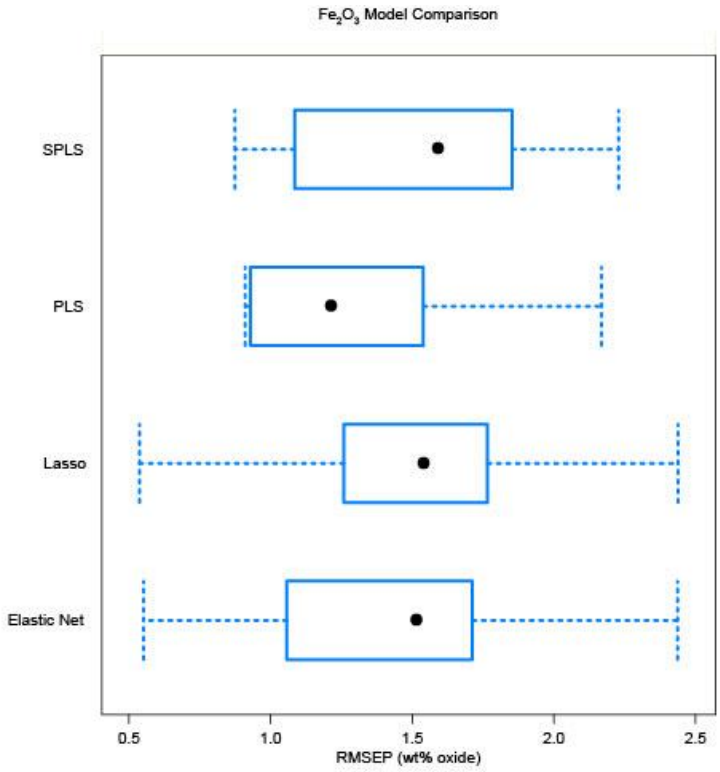

Figure 21. Box-whisker plot of the RMSEP values for Fe_2O_3 models. All models are statistically equivalent. The SPLS and PLS models appear to have smaller RMSEP spreads than the lasso and elastic net models. Fe lines are plentiful in the emission spectrum, so they may have more prominence in the linear combinations that form the coefficients in the PLS model.

both in terms of RMSEP value and RMSEP spread. Fe emission lines are plentiful, so they may be more prominent in the linear combinations that form the coefficients for the PLS model. This may also explain the smaller spread of the SPLS RMSEP. In contrast, lasso may pick lines from clusters of highly correlated

variables and may miss the Fe emission line(s), which could explain the increased variability in RMSEP. Similarly, the elastic net may average out the effects of the Fe emission lines, which could lead to greater RMSEP variance.

CONCLUSIONS

On an elemental basis, the PLS, lasso, elastic net, and SPLS models are statistically equivalent. No model significantly outperforms the others. Thus, models may be chosen based on their other qualities, such as ease of expression, sparsity, and physical interpretability.

In LIBS analysis, PLS is the conventional form of analysis. It is a logical choice because it is familiar to researchers in the field. It has two major drawbacks, however. First, the shrinkage algorithm does not have a nice, closed form expression, so it is difficult to discern exactly what is happening during the shrinkage. This has been a subject of debate (Butler and Denham, 2000; Rosipal and Krämer, 2006). Second, because the model coefficients are made by taking linear combinations of channels from the original data set, these coefficients do not have explicit physical meaning, though they can be mapped onto emission lines using loadings plots. However, this matchup is not as convenient as the direct mapping from model coefficients to emission lines inherent in lasso and elastic net: it is not sparse.

The lasso, elastic net, and SPLS regression methods all improve on the PLS method by enabling feature selection. This indicates that the model coefficients for these three techniques have explicit physical meaning. From a statistical point of view, these models are preferable to PLS because they are

parsimonious. From a chemical perspective, these models are preferable because they are interpretable; coefficients have physical significance so interactions between elements can be better understood. Based on this work, however, one does not find any clear preference for one model over the others. With more data, it is likely that a clear "best" performer (or performers) will emerge.

The flight model on board MSL contains many more channels than the LIBS used to collect the data used in this thesis. This means that more data will be contained in the spectra, so there will be more features available for selection. Subject to good training set from which to build the models, the increased feature availability may lead to more accurate models. Thus, all the models could perform better for spectra obtained on Mars because they will have more emission lines from which to build their models.

The ultimate goal of this thesis is to determine the best method(s) for determining the elemental compositions of Mars rocks from LIBS spectral data. SPLS, lasso, and elastic net not only provide models with accuracies comparable to PLS model accuracies, but they also give interpretable models. These methods can improve researchers' understanding of how Mars evolved by painting a clearer picture of elemental compositions and interactions in rocks.

FUTURE WORK

DATA PREPROCESSING: AVERAGED VERSUS UNAVERAGED SPECTRA

Spectra from each of the 50 shots for each sample were averaged in this thesis, which resulted in 100 averaged spectra that were used for analysis. This was done to average out noise. This is permissible because the samples are homogenous (rock powders). However, the spectra from MSL will be from rocks, which by their nature are heterogeneous. Each shot could contain distinct compositional information. In the context of this application, information about the compositions of the samples could be lost if the shots were averaged before analysis. Moreover, the effect of training set size is untested, although it seems likely that larger data sets will produce improvements in accuracy. Therefore, models fitted using unaveraged spectra (5000 spectra) should be investigated, where CV folds are controlled such that all 50 spectra for a given sample are contained in one unit of analysis (one fold).

ADDITIONAL TECHNIQUES

Due to time constraints, not all shrunken regression techniques of interest were explored for this dataset. Other promising candidates include ridge and fused lasso, both of which were discussed in the background of this thesis. A subset of shrunken regression, known as bridge regression is also of interest because it

contains some sparse methods. These techniques will be pursued in the coming months.

BENCHMARK EXPERIMENTS

Advances in benchmark experiments for comparison of shrunken regression models have been fairly recent. The resampling process used in these experiments must be explored in further detail to determine which method yields the most accurate analysis of model superiority. Resampling techniques to be compared include the bootstrap and cross-validation (Eugster *et al.*, 2008).

AUTOMATIC LINE ASSIGNMENT TO KNOWN PEAKS

Although the lasso and elastic net perform feature selection and select variables that have physical significance, the process by which they select variables does not guarantee that the same emission lines will be chosen each time. In the case of the lasso, for example, it indiscriminately picks a component from a group of highly correlated variables to be included in the model. This choice varies each time the model is run. Also, assignment of the included channels to known emission lines by hand is time consuming, so manual assignment for each element each time the model is run is not feasible.

This process must be automated to fully exploit the strengths of these sparse shrunken regression techniques. Work on this automation has been started.

LIBS ON MARS: GOING THE DISTANCE

One of the numerous challenges that exists with collecting compositional data on Mars relates to the standoff distance at which the sample is collected. When spectra are transmitted from MSL back to Earth, they will include the standoff distance. Sparse shrunken regression models may provide valuable input for other methods used to better interpret the relationship between standoff distance and peak intensity. Features from these sparse models can be used in more advanced regression methods such as generalized additive models (GAMs) that require significantly lower dimensional data than the original LIBS spectral data.

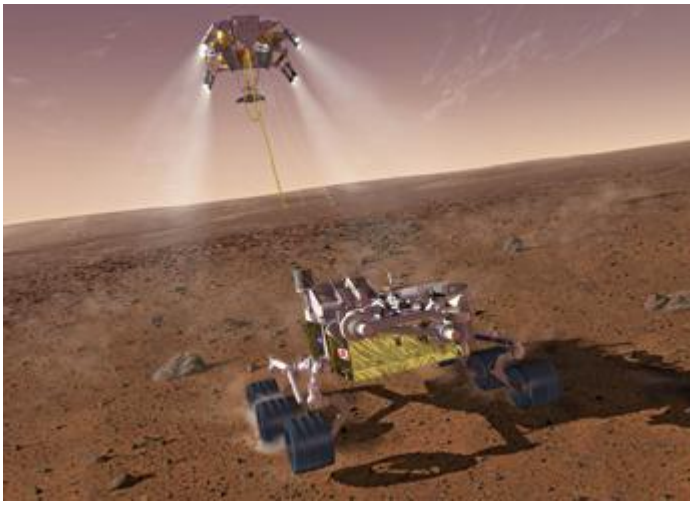

Figure 22. Artist's rendition of MSL landing on Mars. MSL will land on Mars on August 5, 2012 at 10:00 pm US Pacific

REFERENCES

Beleites C., (2012) Interface for hyperspectral data, i.e. spectra + metal info (spatial, time, concentration,…). URL http://hyperSpec.r-forge.r-project.org/.

Butler N. A. and Denham M. C., (2000) The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society. Series B.* 62, 585-593.

Chung D. *et al.,* (2012) Sparse Partial Least Squares (SPLS) Regression and Classification. URL http://www.stat.wisc.edu/~chungdon/spls/.

Chun H. and Keleş S., (2010) Sparse partial least squares regression for simultaneous dimension reductions and variable selection. *Journal of the Royal Statistical Society. Series B.* 72, 3-25.

Cooper J. B., *et al.,* (1995) Determination of Octane Numbers and Reid Vapor Pressure of Commercial Petroleum Fuels Using FT-Raman Spectroscopy and Partial Least Squares Regression Analysis. *Anal. Chem.* 67, 4096-4100.

Cremers D. A. and Radziemski L. J., (2006) Handbook of Laser-Induce Breakdown Spectroscopy. John Wiler & Sons, Ltd: Chichester, pp 1-22.

Dyar M. D. *et al.,* (2012) Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochim. Acta* B (submitted).

Eugster M. J. A. *et al.,* (2008) Exploratory and Inferential Analysis of Benchmark Experiments. Technical Report Number 030, Dept. of Statistics, University of Munich.

Friedman *et al.,* (2010) Regularized Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Goutis C., (1996) Partial Least Squares Algorithm Yields Shrinkage Estimators. *The Annals of Statistics*. 24, 816-824.

Harris D. C. (2010) Quantitative Chemical Analysis 8th Ed. W. H. Freeman and Company: New York, pp. 493-495.

Harris D. C. and Bertolucci M. D., (1978) Symmetry and Spectroscopy: An Introduction to Vibrational and Electronic Spectroscopy. Dover Publications, Inc.: New York, pp. 1-4.

Hastie T. *et al.,* (2009) The Elements of Statistical Learning. Springer: New York, pp 57-82, 219-233, 649-651, 658-668.

Hoerl A. E. and Kennard R. W., (1970) Biased Estimation for Nonorthogonal Problems. *Technometrics.* 12, 55-67.

Hothorn T. *et al.,* (2005)The Design and Analysis of Benchmark Experiments. *Journal of Computational and Graphical Statistics.* 14, 675-699.

Iñón F. A. *et al.,* (2003) Selection of calibration set samples in determination of olive oil acidity by partial least squares – attenuated total reflectance – Fourier transform infrared spectroscopy. *Analytica Chemica*. *Acta* 489. 59-75.

Kuhn, M. (2012) caret: Classification and Regression Training. URL http://caret.r-forge.r-project.org/.

Merriam-Webster's Collegiate Dictionary, Eleventh Ed., Merriam-Webster, Inc: Springfield, p. 857, 2003.

Mevik *et al.,* (2011) Multivariate regression methods Partial Least Squares Regression (PLSR), Principal Component Regression (PCR) and Canonical Powered Partial Least Squares (CPPLS). URL http://mevik.net/work/software/pls.html.

Model diagnostics (2004). Stat 328. URL http://www.public.iastate.edu/~alicia/stat328/Model%20diagnostics.pdf.

Morhac, M. (2008) Spectrum manipulation: background estimation, Markov smoothing, deconvolution and peaks search functions. URL http://cran.r-project.org/web/packages/Peaks/.

Ocean Optics: HR2000+ High-speed Fiber Optic Spectrometer Installation and Operation Manual (2010) Ocean Optics, Inc., p. 13.

Ozanne M. V. *et al.,* (2012) Comparison of lasso and elastic net regression for major element analysis of rocks using laser-induced breakdown spectroscopy (LIBS). LPSC XLIII Abstract 2391.

Patterson J. G. (1992). *Benchmarking Basics*. Crisp Publications Inc.: Menlo Park.

Rahmelow K. and Hübner W., (1996) Secondary Structure Determination of proteins in Aqueous Solution by Infrared Spectroscopy: A Comparison of Multivariate Data Analysis Methods. *Analytical Biochemistry*. 241, 5-13.

Ramsey F. and Schafer D., (2002) The Statistical Sleuth. Duxbury Press: Pacific Grove, pp. 123-124.

Rice J. A., (2007) Mathematical Statistics and Data Analysis. Brooks/Cole: Belmont, p. 335.

Rhodes J. M. and Vollinger M. J., (2004) Composition of basaltic lavas sampled by phase-2 of the Hawaii Scientific Drilling Project: Geochemical stratigraphy and magma types. *Geochem. Geophys. Geosyst.* 5, Q03G13.

Rosipal R. and Krämer N., (2006) Overview and Recent Advances in Partial Least Squares. *LNCS 3940*, 34-51.

Tibshirani R., (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B.* 58, 267-288.

Tibshirani R. *et al.*, (2005) Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B.* 67, 91-108.

Tucker J. M. *et al.*, (2010) Optimization of laser-induced breakdown spectroscopy for rapid geochemical analysis. *Chem. Geo.* 277, 137-148.

Wagaman A. S., (2008) Topics in High-Dimensional Inference with Applications to Raman Spectroscopy. <dissertation – Michigan, dept. of stat.>

Weisstein E. W., (2012) Bonferroni Correction. From *MathWorld*—A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniCorrection.html.

Zou H. and Hastie T., (2012) Elastic-Net for Sparse Estimation and Sparse PCA. URL http://www.stat.umn.edu/~hzou.

Zou H. and Hastie T., (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B.* 67, 301-320.

APPENDIX

REMOVING NA VALUES FROM DATA

```
killNA ← function(comps.xm.hs, element.name){
        print(element.name)
        my.elt ← !is.na(comps.xm.hs@data[,element.name])
        comps.xm.hs[my.elt]
}
```

FIXING CROSS-VALIDATION FOLDS

```
model.folds ← function(elements, x.hs){
        sapply(elements, function(y){
                createFolds (killNA(x.hs, y)@data[[y]], k=10, list=TRUE,
                returnTrain=TRUE)},
                USE.NAMES=TRUE, simplify=FALSE)
}

cv.trainControl ← function(elements, cv.folds){
        sapply(elements, function(y){
                trainControl(method= "cv", selectionFunction= "oneSE",
                index=cv.folds[[y]])},
                USE.NAMES=TRUE, simplify=FALSE)
}
```

CREATING TUNING GRIDS FOR MODEL PARAMETER(S)

```
enet.grid ← createGrid(method = "glmnet", len=2)
pls.grid ← createGrid(method = "pls", len=15)
lasso.grid ← createGrid(method = "lasso", len=10)
spls.grid ← createGrid(method = "spls", len=5)
```

BUILDING MODELS

```
MakeLassoModels.caret ← function(elements, x.hs, tuning.grid, trc.list){
        mapply(function(y, tc.y){
                train(as.data.frame (killNA(x.hs, y)[[]]), killNA(x.hs, y)@data[[y]],
                method = "lasso",
                USE.NAMES=TRUE, SIMPLIFY=FALSE)
}
```

```
MakeElasticNetModels.caret ← function(elements, x.hs, tuning.grid, trc.list){
        mapply(function (y, tc.y){
                train(as.data.frame(killNA (x.hs, y)[[]]), killNA(x.hs, y)@data[[y]],
                method = "glmnet",
                USE.NAMES=TRUE, SIMPLIFY=FALSE)
}

MakePLSModels.caret ← function(elements, x.hs, tuning.grid, trc.list){
        mapply(function (y, tc.y){
                train(as.data.frame (killNA(x.hs, y)[[]]), killNA(x.hs, y)@data[[y]],
                method = "pls",
                USE.NAMES=TRUE, SIMPLIFY=FALSE)
}

MakeSPLSModels.caret ← function(elements, x.hs, tuning.grid, trc.list){
        mapply (function(y, tc.y){
                train(as.data.frame (killNA(x.hs, y)[[]]), killNA(x.hs, y)@data[[y]],
                method = "spls",
                USE.NAMES=TRUE, SIMPLIFY=FALSE)
}
```

EVALUATING DIFFERENCES BETWEEN MODELS

Collecting Resampling Results

```
resamps ← resamples(c(SPLS = MakeSPLSModels.caret, PLS =
        MakePLSModels.caret, LASSO = MakeLassoModels.caret, ENET =
        MakeElasticNetModels.caret)
```

Creating Box-whisker Plots (Example)

```
bwplot(resamps, metric = "RMSE", models = c("SPLS.SiO2", "PLS.SiO2",
        "LASSO.SiO2", "ENET.SiO2"))
```

Computing t-tests (Example)

```
difValues.SiO2 ← diff(resamps, models = c("SPLS.SiO2", "PLS.SiO2",
        "LASSO.SiO2", "ENET.SiO2"))

summary(difValues.SiO2)
```