Evaluation of Statistical Methods for Classification of Laser-Induced Breakdown Spectroscopy (LIBS) Data

Michelle Lynn DeVeaux

Advisors:

Professor Ji Young Kim and Professor M. Darby Dyar

Presented to the faculty of Mount Holyoke College in partial fulfillment of the requirements for the degree of Bachelor of Arts with Honors

Department of Mathematics and Statistics

Mount Holyoke College

South Hadley, MA 01075

May 2012

ACKNOWLEDGEMENTS

I would like to thank Professor Ji Young Kim for her guidance and support in all statistical matters and Professor Darby Dyar for opening up the world of NASA, Mars rocks and the rover to me. I am so grateful to you both for providing me with this incredible research opportunity and having faith in me and my abilities throughout the process. I have learned so much.

Working in Professor Dyar's lab has been a wonderful experience. I owe a special thanks to Marco Carmosino for providing me with great resources and sharing his knowledge, and to Elly Breves whose assistance in the lab was always appreciated.

I thank Professor Janice Gifford, not only for serving on my Thesis Committee, but also for her being a wonderful mentor and a wealth of information since my first semester at Mount Holyoke, where she introduced me to the world of statistics.

Professor Margaret Robinson has been a terrific academic advisor and I thank her for her insight during the past four years.

I also would like to thank Professor Audrey St. John whose computer science courses proved very valuable during this project and in all my studies, and whose advice in the past few years has been so welcomed.

I would be remiss not to thank my wonderful family and friends whose love, support, and encouragement make all the difference in my world. I am so grateful to my parents who have been extremely supportive of me in all of my pursuits, big and small. I owe a genuine thanks to my friends Ashleigh Eubanks, Brittany Finder, Sophia Lozada, Liana Simonds, and Lauren Walch, for the wonderful memories and so many laughs we have shared over these four years. You all have enriched my experience at Mount Holyoke in so many ways.

Thank you all for making my experience at Mount Holyoke College one where, as our founder Mary Lyon said, "… if you will jump in you may ride very fast." I jumped in and enjoyed the amazing ride.

ABSTRACT

When NASA's *Curiosity* rover lands in August 2012, the rover will use a laser-induced breakdown spectroscopy (LIBS) instrument to collect data in an effort to understand the chemical composition and geological classification of the rocks on Mars. This is part of a larger endeavor to determine information about the planet's habitability. LIBS is a method used to determine the elemental composition of a given sample. For each rock sample analyzed by the instrument, a LIBS spectrum consisting of over 6,000 different channels is obtained.

In order to prepare for the return of LIBS data from the rover, this project aims to evaluate the accuracy of statistical methods, such as discriminant analysis, support vector machines, and clustering algorithms for categorizing the rock samples into groups with similar chemical compositions based on their LIBS spectra alone. Accurate classification is critical for rapid identification of similar unknown samples, novelty detection, and in the selection of a training set of data for use in the estimation of chemical compositions. Similar studies have been performed; however, they generally fail to use statistical best practices and therefore have wildly optimistic results.

The data used in this project is from the "century set", a suite of 100 igneous rock samples. These 100 samples are the only ones currently available for this project which have both LIBS spectra and known chemical compositions. Having the known chemical compositions allowed the century set samples to be divided into groups with geological similarities based on their Total Alkali-Silica (TAS) classes, and provided a way to evaluate the predictive accuracy of the classification algorithms using *K*-fold cross validation.

The results show that the small sample size and uneven distribution of samples in different TAS classes make classification into many groups difficult, contradicting many of the outcomes displayed in the literature. However, some of the methods explored in this thesis do show promise based on their performance in simpler classification tasks, so the results should be reevaluated once more data is obtained.

LIBS data is scarce, so this thesis also briefly explores the results from one method of simulating a LIBS spectrum based on the sample's chemical composition. Simulated data could be used to examine the effects of sample size on the accuracies of the various classification algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ARI	Adjusted Rand Index
FDA	Flexible Discriminant Analysis
LDA	Linear Discriminant Analysis
LIBS	Laser-Induced Breakdown Spectroscopy
MDS	Multidimensional Scaling
MSE	Mean Squared Error
PCA	Principal Component Analysis
PDA	Penalized Discriminant Analysis
RBF	Radial Basis Function
SLDA	Sparse Linear Discriminant Analysis
SDA	Shrinkage Discriminant Analysis
SVM	Support Vector Machines
TAS	Total Alkali-Silica

INTRODUCTION

The *curse of dimensionality* is a term coined by mathematician Richard E. Bellman in 1961 to describe the challenges of working in high-dimensional spaces (Bishop 2006, 36). This "curse" certainly poses many interesting problems in statistics, as advances in the sciences are causing high-dimensional data to become increasingly prevalent.

Laser-induced breakdown spectroscopy (LIBS) is a method used to determine the quantity of various chemical elements in a given sample. The LIBS technique was first performed in 1963 (Miziolek et al. 2006, 5). Since then, the major advantages of LIBS over other similar techniques have caused a surge in LIBS-related research. From its use in biological applications like tissue classification (Yueh 2009) and in the categorization of plastic polymers (Anzano et al. 2010) to extraterrestrial applications, LIBS is becoming a popular technique in the analytical chemistry community.

Some of this activity is due to the presence of a LIBS instrument onboard NASA's *Curiosity* rover currently en route to Mars. Upon its landing in August of 2012, the rover will use the LIBS instrument ("ChemCam") to collect data in an effort to understand the chemical composition and geological classification of the rocks on Mars. Statistical methods will be used to translate the high-dimensional spectral data collected by the instrument into meaningful information

about the chemistry of the rocks on Mars, which will in turn inform an understanding of Martian geology. For each rock sample analyzed with the instrument, a LIBS spectrum consisting of over 6,000 intensities of light at various wavelengths is obtained. Therefore, the limitations related to the *curse of dimensionality* certainly are present in the analysis of LIBS data and restrict the practical application of many statistical procedures in this situation (Duda et al. 2001, 170).

Figure 1: Sketch of *Curiosity* rover

http://www.nasa.gov/mission_pages/msl/multimedia/gallery/pia14156.html

Laser-induced breakdown spectroscopy

Laser-induced breakdown spectroscopy (LIBS) will be employed by the ChemCam instrument on the Mars Science Laboratory rover *Curiosity* to obtain data (atomic emission spectra) about Martian surface rocks and soils. Researchers will use the tools on the rover to study whether the landing region has had environmental conditions favorable for supporting microbial life and for clues about whether life existed (NASA 2011). LIBS is ideal for an extraterrestrial application because it provides real-time analysis and requires no sample preparation (Ukwatta et al. 2012).

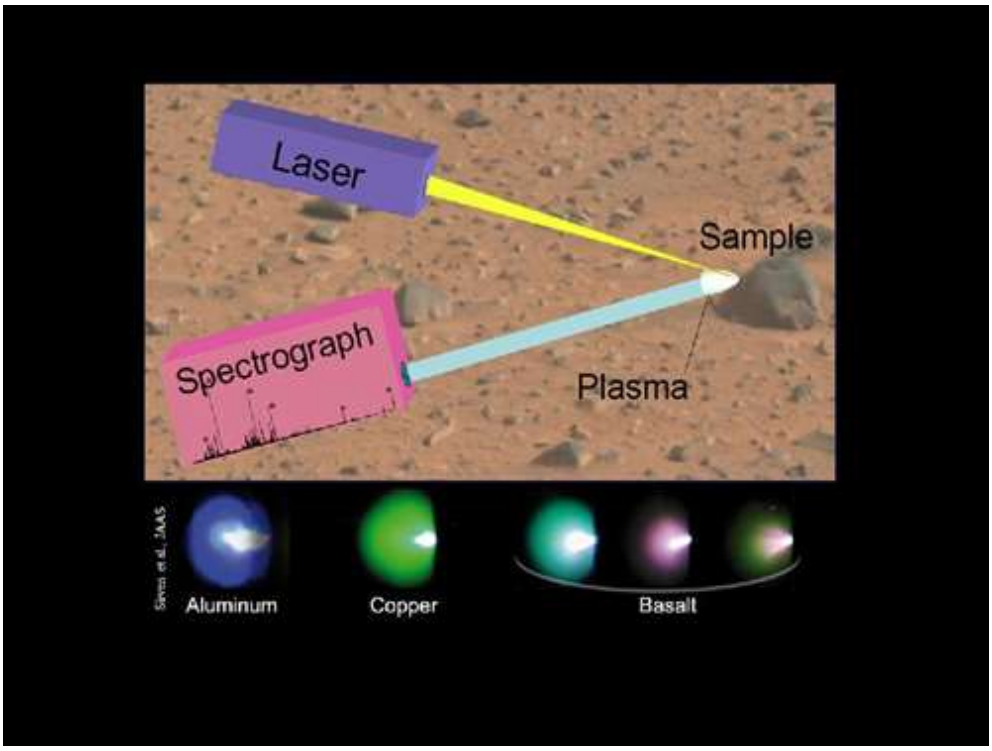

Figure 2: Schematic of laser-induced breakdown spectroscopy

http://www.nasa.gov/mission_pages/msl/multimedia/pia15103.html

The LIBS instrument onboard the rover will fire pulses of a laser at target rock and soil samples from up to seven meters away. Energy from the laser excites a microscopic spot on the target into a glowing, light-emitting plasma (ionized gas). The plasma light is collected by a telescope and focused into the end of a fiber optic cable (NASA Jet Propulsion Lab 2010). The fiber optic cables carry the light to three different spectrometers incorporated into the instrument, one for each of ultraviolet (UV), visible (VIS), and near infrared (VNIR) regions of the electromagnetic spectrum (Lasue et al. 2011). The spectrometers record a spectrum for each sample analyzed by collecting intensities of light emitted at over 6,000 different channels (wavelengths) between 240 and 850 nm, which cover the range of these three regions.

Every chemical element emits a different characteristic wavelength, or color of visible light, as shown in Figure 2 (NASA 2011). Therefore, the peaks found in the spectrum of light emitted by the plasma can be used to identify the chemical elements present in the target sample (NASA Jet Propulsion Lab 2010). Typical rock and soil analyses yield detectable quantities of ten major chemical elements as well as trace amounts of many other minor elements. These different chemical elements interact in the plasma and cause variations in the peak intensities. This is referred to as the chemical matrix effect. Multivariate statistical methods such as partial least squares – a regression method that mitigates the collinearity of the data that results from features of the data – can be used to predict the quantities of each element found in the sample.

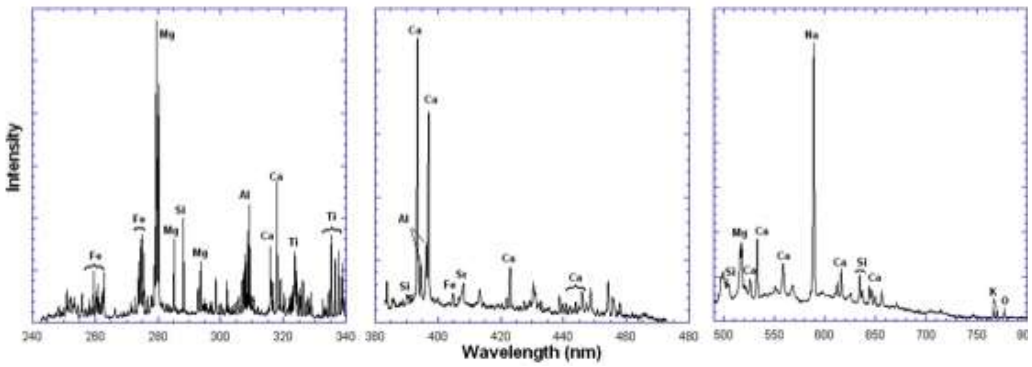

Figure 3: Sample LIBS spectrum with elemental peaks labeled

http://msl-scicorner.jpl.nasa.gov/Instruments/ChemCam/

Overview of Thesis

When the Curiosity rover lands on Mars, it will begin to collect LIBS data that will be sent back to Earth. We will have only photographic information about the target samples being analyzed. In order to prepare to make meaningful conclusions about the returning data, we must use test data sets to investigate and develop statistical methods that are well suited to the analysis of this specific type of data. This thesis contributes to the preparations in two ways.

First, this project provides an empirical analysis to evaluate the predictive accuracy of various statistical methods for categorizing the rock samples into groups with similar chemical compositions based on their LIBS spectra alone. Current literature provides some studies presenting results of classification of

LIBS data; however, the majority of sources fail to compare multiple methods, especially methods that span different subsets of statistical algorithms, or they fail to rigorously verify their results using techniques well-known in statistics. This has led to results that seem to be overly optimistic. The goal of this thesis is not to develop new methodology, but rather to apply rigorous error analysis to well-established statistical techniques. Accurate classification is critical for rapid identification of similar unknown samples, allowing us to gain a sense of the distributions of rock types on Mars on a broad scale. Similarly, it can also be used for novelty detection, the identification of a new or unknown sample type. Also, being able to identify samples that are chemically similar to a given sample is important in the estimation of the chemical composition of a given sample.

Data from the "century set", a suite of 100 igneous rock samples, are the primary basis for this thesis. Small-sample size effects can sometimes contaminate the design and evaluation of a proposed classification algorithm, especially in the case of high-dimensional data (Raudys and Jain 1991). The 100 samples in the century set are the only ones currently available for this project that have both LIBS spectra and known chemical compositions. Data on the chemical compositions of many other rock suites are readily available, but LIBS data are scarce. Therefore, the second objective of this project is to examine the possibility of producing a valid model that can be used to simulate the LIBS spectra of a sample for which we only have chemical composition data. A model with good predictive accuracy would allow for the simulation of LIBS data,

which could be used to investigate the effects of sample size on the accuracies of the various classification algorithms.

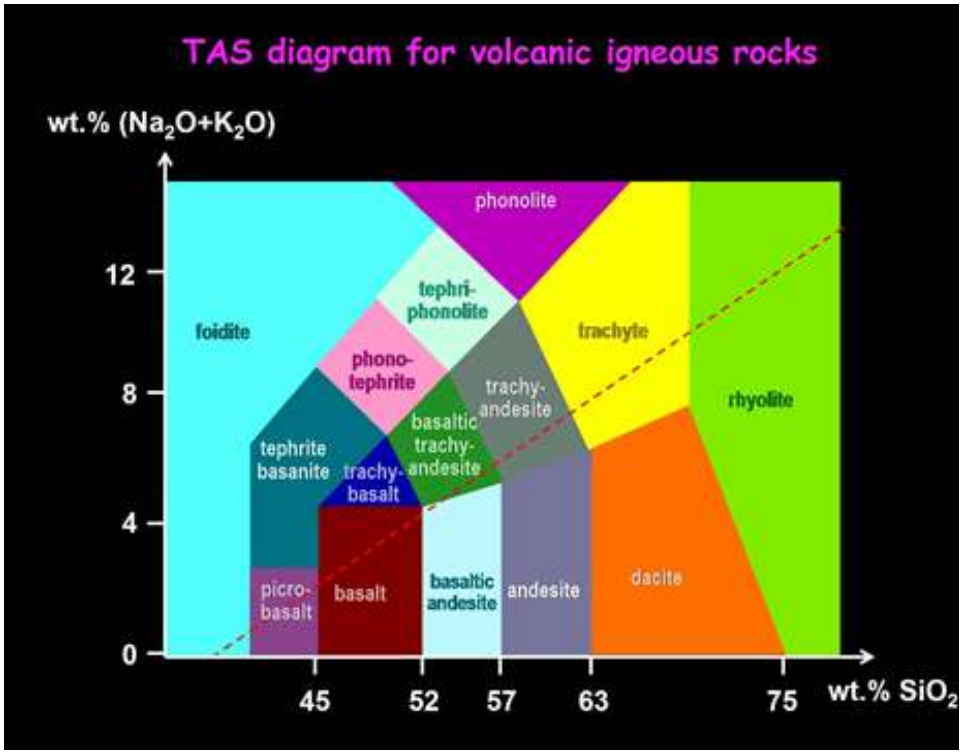

Figure 4: Total Alkali-Silica (TAS) Diagram

http://umanitoba.ca/geoscience/faculty/arc/tas.html

The Data

The century set was analyzed by LIBS at Los Alamos National Laboratory under conditions to mimic the atmosphere of Mars. Additionally, concentrations of major and minor elements were determined by X-ray fluorescence (XRF) using standard procedures at the University of Massachusetts Amherst. Using the data

from XRF, we divided the samples in the century set into classes with similar chemical compositions using the total alkali versus silica (TAS) diagram, as shown in Figure 4. The century set is comprised of samples from 12 of the 15 different TAS classes in the diagram. The majority of samples in the century set are basalts by composition, which makes the distribution of sample sizes in the various TAS classes uneven. More details about the distribution of TAS classes found in the century set are displayed in Table 1.

Century Set (n = 100)	
TAS Class	**Frequency**
Andesite	4
Basalt	42
Basaltic andesite	2
Basaltic trachyandesite	9
Basanite tephrite	4
Dacite	10
Foidite	7
Phonotephrite	4
Picrobasalt	3
Rhyolite	6
Trachyandesite	1
Trachybasalt	8

Table 1: Frequency of TAS classes in the century set

CLASSIFICATION AND CLUSTERING METHODS

This section describes the methods used in a general sense and then specifies how each method is used in the scope of this project. The next section displays the results from applying these methods to the century set.

Machine Learning Algorithms

Statistical data often can be naturally divided into two or more groups, where the reasoning behind the division is known in advance. The goal of statistical classification is to establish rules, or classifiers, on the basis of objects with known group memberships. These can be reliably used for predicting the group membership of new observations, as well as evaluating the performance of the rules (Varmuza and Filzmoser 2009, 197).

Statistical methods for data classification fall into the larger category of machine learning algorithms. A machine learning approach uses a data set of samples, called a training set, to tune the unknown parameters of an adaptive model (Bishop 2006, 2). The known categories of the samples within the training set can be expressed in a vector of labels, **t**. For classification of the century set, the TAS classes assigned to each sample will be used as labels. The resulting model can be expressed as a function $y(\mathbf{x})$ that takes in vectors of input features, **x**, and generates an output vector, **y**, encoded in the same way as the label vector, in this case in the form of a predicted TAS class for each sample (Bishop 2006,

2). The exact form of the model reduces the error on the training set. This trained model can then be used to predict the label of new observations. Algorithms in which class labels are provided for each sample in the training set fall into the category of supervised machine learning.

In other pattern recognition techniques, the training data consists only of input features, **x**, without a vector of labels, **t** (Bishop 2006, 3). Such methods are considered part of unsupervised learning. In unsupervised techniques, algorithms form clusters or natural groupings of the input patterns (Duda et al. 2001, 17).

In order to evaluate prediction error of the supervised classification models, *K*-fold cross-validation is used. Ideally, if we had enough data, we could train our prediction model with a training set of data and then use different set of data, a validation set, to assess the model's performance for prediction. Unfortunately, LIBS data are scarce so other methods like *K*-fold cross validation can be used. *K*-fold cross validation allows one data set to be divided into test and validation sets and rotates the partition so that every sample is included in the test set. The original sample is randomly split into *K* roughly equal-sized parts. For $k = 1, 2, \ldots K$ the *k*th part is used as a validation set, and the other $K - 1$ parts are used as the training set to fit the model. A chosen measure of error is calculated for predicting the classes of the samples in the validation set (Hastie et al. 2009, 241 – 242). The *K* results can then be averaged to produce a single error estimation.

Methods for Dimension Reduction and Data Visualization

Multidimensional Scaling

High-dimensional data are impossible to visualize without the help of some form of dimension reduction technique. Multidimensional scaling (MDS) is one of these techniques. It projects data points to a lower-dimensional space so that the dissimilarities between original data points are represented as the distances between the points in the lower-dimensional space (Duda et al. 2001, 573). Classical multidimensional scaling uses the Euclidean distance as a measure of dissimilarity.

Principal Component Analysis

Principal component analysis (PCA) is another method used for dimension reduction. It can transform a group of many highly-correlated *x*-variables into a smaller set of uncorrelated latent variables (a variable that is not observed, but rather inferred) that can be used in place of the original variables (Varmuza and Filzmoser 2009, 59). The data are first centered and scaled. The latent variables are determined by finding the directions in the variable space that best keep the relative distances between the objects. In other words, the latent variables best preserve the variance of the data values. The variable that preserves the maximum variance of the data is called the first principal component (PC1). For a data matrix with p variables, its first principal component is defined by the linear combination of the variables $X_1, X_2, \dots, X_p$

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ \ldots + \ a_{1p}X_p$$

with a loading vector of unknown coefficients $a_1 = \left(a_{11}, a_{12}, \ldots, a_{1p}\right)^T$ with normalized length such that ${a_1}^T a_1 = 1$ (Holland 2008; Varmuza and Filzmoser 2009, 69 -70).

The kth principal component is the linear combination that accounts for the next highest variance after the $k\text{th} - 1$ component. It must be orthogonal to all previous directions. Collectively, all of the vectors of coefficients a_p can be placed as columns in a matrix A, called the loading matrix, such that the transformation of the original variables to the principal components is

$$Y = AX.$$

The rows of A are the eigenvectors of the variance-covariance matrix of the original data. The elements of the eigenvector are the weights a_{ij} and are known as loadings. The elements in the diagonal of the variance-covariance matrix of the principal components are the eigenvalues. The variance explained by each principal component is equal to the eigenvalue (Holland 2008; Varmuza and Filzmoser 2009, 69 – 70).

Each observation in a principal component is called a score and is a linear combination of the original variables, x_{ij}, and the loadings, a_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The score for the r^{th}sample on the k^{th}principal component can be computed as

$$Y_{kr} = a_{k1}x_{k1} + a_{k2}x_{k2} + \ldots + a_{kp}x_{kp}\,.$$

Supervised Classification Techniques

Discriminant Analysis

Discriminant analysis (DA) is a traditional approach to supervised classification. Decision boundaries are constructed that separate the classes from one another.

In linear discriminant analysis (LDA), we create decision boundaries to separate classes that are linear, as shown in Figure 5. Suppose that there are K classes, labeled $1, 2, \ldots, K$, where k represents the kth class. It is assumed that the classes have a normal distribution and a common covariance matrix $\sum_k = \sum, \forall\, k$. Given an input vector x, linear discriminant functions follow the form (Hastie et al. 2009, 109)

$$\delta_k(x) = x^T\sum^{-1}\mu_k - \frac{1}{2}\,\mu_k{}^T\sum^{-1}\mu_k + \log \pi_k.$$

In practice, we do not know the parameters of the distribution, so they are estimated using the training data and maximum likelihood estimation:

$\hat{\pi}_k = \frac{N_k}{N}$, where $N_k$is the number of class-k observations;

$\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$, where x_i represents the ith sample vector;

$$\hat{\sum} = \sum_{k=1}^{K}\sum_{g_i=k}(x_i - \hat{\mu}_k)\,(x_i - \hat{\mu}_k)^T/(N-K).$$

The decision boundary between class k and l is where $\delta_k(x) = \delta_l(x)$. Although there are many different types of discriminant analysis, LDA often produces the best classification results because of its simplicity and low variance (Hastie et al. 2009, 439). Sometimes, however, linear boundaries are not adequate to separate the classes.

There are other types of discriminant analysis that create nonlinear boundaries to separate classes. LDA can be recast as a linear regression problem, which can then be generalized into a more flexible, nonparametric form of regression. This allows for a more flexible form of discriminant analysis, appropriately referred to as flexible discriminant analysis (FDA). The feature vectors can be mapped into a higher dimensional space and LDA is then performed in this enlarged space (Hastie et al. 2009, 439). This is similar to the procedure used with support vector machines, which will be explained shortly.

Penalized discriminant analysis (PDA) is another method that creates nonlinear boundaries. In a similar manner to FDA, the predictors are first expanded. Then an LDA model is fitted, but coefficients are penalized to be smooth (Hastie et al. 2009, 440 – 446).

Stabilized linear discriminant analysis (SLDA) is another method that performs dimension reduction followed by LDA. The data are reduced into linear scores that are left-spherically distributed and are used as predictors for linear discriminant analysis (Peters and Hothorn 2012). Similarly, shrinkage

discriminant analysis (SDA) shrinks the dimension of the data by determining a ranking of predictors by computing correlation-adjusted t-scores (CAT) scores between the group centroids and the pooled mean (Ahdesmaki and Strimmer 2010). Once again, LDA is performed on the shrunken data.

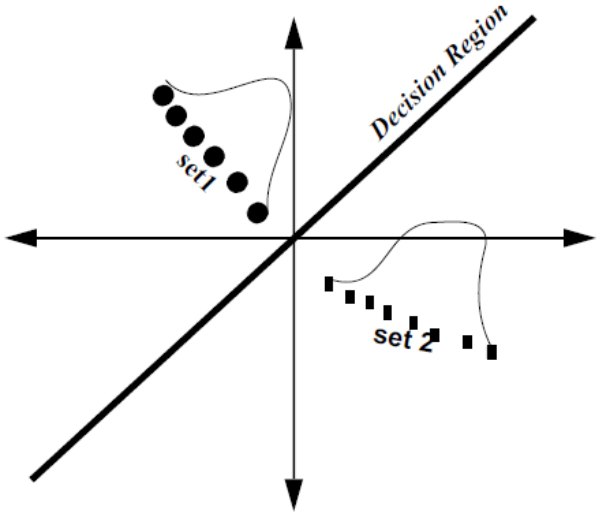

Figure 5: Schematic of discriminant analysis

http://www.music.mcgill.ca/~ich/classes/mumt611_07/classifiers/lda_theory.pdf

The main way to evaluate discriminant analysis models is to compute the accuracy of classification based on *K*-fold cross validation. Calculating the Kappa statistic (Cohen 1960) is another way to evaluate the results from discriminant analysis. The Kappa statistic is a measure of agreement for classification relative to what is expected by chance. A Kappa value of zero indicates a lack of agreement while a value of one indicates perfect agreement. Kappa is a useful statistic when the classes are highly unbalanced (Kuhn 2008).

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o denotes the observed proportion of agreements among all classifications and P_e denotes the expected value of P_o under random agreement (Gross 1986).

Discriminant analysis was performed in R using the `caret` (classification and regression training) package (Kuhn et al. 2012). Using this package, discrimination functions were tuned for the five different variations of discriminant analysis (linear, flexible, penalized shrinkage, and stabilized linear discriminant analysis). A 10-fold cross validation was used for parameter tuning for flexible and penalized discriminant analysis. The parameters selected minimized the expected classification error.

Support Vector Machine

In the similar way to discriminant analysis, the support vector machine (SVM) is used to create hyperplanes to separate different classes. Like FDA, SVM extends to the nonlinear case where the classes overlap. SVMs construct a linear boundary between classes in a large, transformed version of the feature space in such a way to maximize the margin between the groups (Varmuza and Filzmoser 2009, 223).

In the r-dimensional case, a hyperplane separating one class from another is defined by

$$b_0 + \boldsymbol{b}^T \boldsymbol{x} = 0$$

where $\boldsymbol{b}^T = (b_1, b_2, \dots, b_r)$ and $\boldsymbol{x} = (x_1, x_2, \dots, x_r)$ and $\boldsymbol{b}^T \boldsymbol{b} = 1$ (Varmuza and Filzmoser 2009, 224). For a simple example, assume that we only have two classes that we are trying to classify into. Given the feature vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ we know the group membership for each object, which is given by the value y_i. In the two-class case, we can say that $y_i = -1$ for membership in the first group, or $y_i = +1$ for membership in the second group. We hope to find a hyperplane that gives a perfect group separation, such that

$$y_i(b_0 + \boldsymbol{b}^T \boldsymbol{x}_i) > 0 \quad \text{for } i = 1, \dots, n.$$

This is possible when the two groups are linearly separable (Varmuza and Filzmoser 2009, 225). The position of the hyperplane is such that the margin between groups is maximized. The data points that define the division of the hyperplanes are called support vectors (Varmuza and Filzmoser 2009, 225).

In the nonlinearly separable case, we have to allow for points to be on the wrong side of the resulting hyperplane in order to maximize the margin. A slack variable is introduced, ξ_i, for $i = 1, \dots, n$, which are defined by the distance from the hyperplane with margin M. $\xi = 0$ for objects on the correct side of the hyperplane and are positive otherwise (Varmuza and Filzmoser 2009, 225).

These methods are generally performed in a transformed space that is enlarged using basis expansions. Every object vector $\boldsymbol{x}_i$ is replaced with the

vector $\boldsymbol{h}(\boldsymbol{x}_i)$ with r dimensions (Varmuza and Filzmoser 2009, 226). The linear hyperplanes in the transformed space translate to nonlinear boundaries in the original space. For certain basis functions the "kernel trick" can be applied, which allows for implicit mapping into the transformed space with the use of kernel function (Karatzglou et al. 2006). For example, two object vectors, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for $i, j = 1, \dots, n$ that are transformed by basis functions become

$$k(\boldsymbol{x}_i, x_j) = h(\boldsymbol{x}_i)^T h(\boldsymbol{x}_j)$$

where k is a kernel function that computes products in the transformed space (Varmuza and Filzmoser 2009, 226). This allows us to specify the kernel functions without specifying the basis functions. Three popular kernels are the linear kernel, radial basis function (RBF) kernel and the sigmoid kernel. The kernel functions are below:

Linear:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \boldsymbol{x}_i^T \boldsymbol{x}_j$$

Radial basis function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\gamma \left\| \mathbf{x_i} - \mathbf{x_j} \right\|^2\right) \text{ with } \gamma > 0.$$

Sigmoid (also referred to as Neural Network):

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j + r) \text{ with } \gamma > 0 \; and \; r < 0$$

where γ and r are kernel parameters (Varmuza and Filzmoser 2009, 227).

Support vector machine methods are implemented with R's `e1071` package (Dimitriadou et al. 2011). The procedure for support vector machines was very similar to that of discriminant analysis. The three kernel functions explained above are used. 10-fold cross validation is used to tune the parameters of the models, gamma and cost. The parameter values considered were 2^{-15} to 2^{3} for gamma and 2^{-5} to 2^{15} for cost. The optimal parameters were selected in correspondence with the lowest classification error from the cross-validated results.

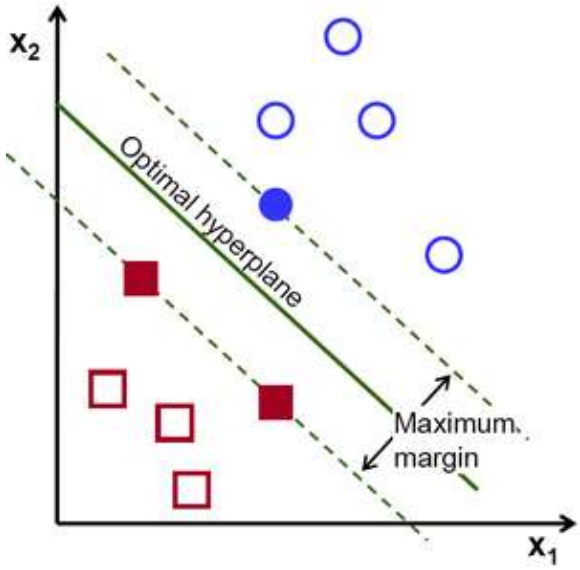

Figure 6: Schematic of support vector machines (SVM). The shaded squares and circles represent the support vectors.

http://opencv.itseez.com/doc/tutorials/ml/introduction_to_svm/introduction_to

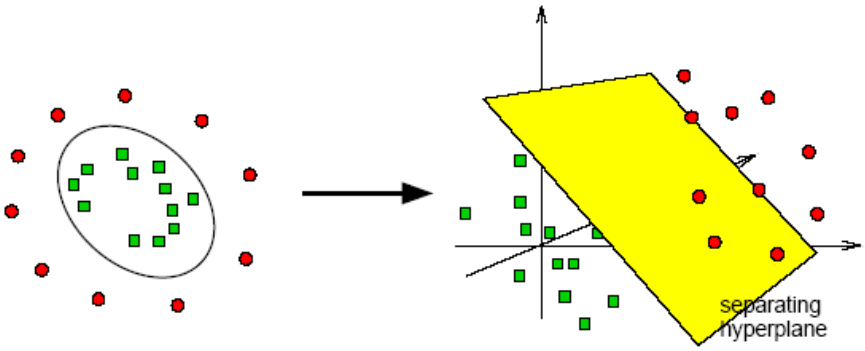

Figure 7: Schematic of the "kernel trick": we can imagine that non-linear separation between vectors in the original space (left image) can be equated to linear separation between vectors in a higher-dimensional space using inner products (right image).

http://www.biostat.pitt.edu/biostat2055/11/110128_W4_Classification2.ppt

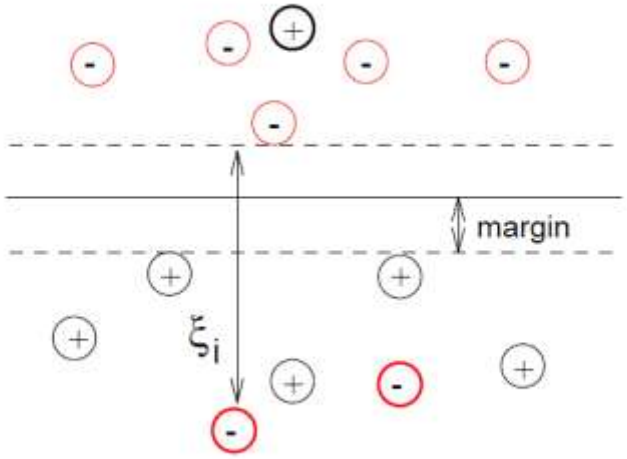

Figure 8: Schematic of support vector machines.

http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

Discriminant Analysis versus Support Vector Machines

Discriminant Analysis and Support Vector Machines may seem similar in their techniques of computing hyperplanes to classifying data, however, they differ in their methodology and assumptions. In LDA, the hyperplane is only optimal if the covariance matrices of the classes are identical, which is commonly violated in practice (Gokcen and Peng 2002). SVMs do not make this assumption.

Unsupervised Learning

k-means Clustering

In the *k*-means algorithm, we start with an unlabeled data set and work top-down to create clusters. We randomly initialize a set of *k* points to be our cluster centroids. Then using our cluster centroids, we examine each point in our data set, determine which cluster centroid is closest to it using typically the Euclidean distance (although other metrics can be used), and assign that point to the corresponding cluster. Once all points have been assigned to a cluster, we then update the cluster centroids so they represent the mean of all of the points that have been assigned to that particular cluster. Then using our new cluster centroids, we repeat the process of assigning points to clusters and updating the cluster centroids until we reach convergence where the clusters no longer change. In this algorithm, we must choose *k*, the number of clusters to create (Hastie et al.

2009). The *k*-means algorithm does not make any assumptions about the data (Varmuzza 2009, 267).

Evaluating Clustering Outcomes

In standard classifications tasks, there is a correct classification against which we can compare the results of classification outcomes based on different algorithms. This is often something like an accuracy rate of classification – the number of samples correctly classified divided by the total number of samples. Labels are not required with clustering algorithms, and therefore several intrinsic metrics exist that compute the quality of the clusters without taking labels into consideration. However, since we do know true class labels for the century set, it is important to compare our clustering outcomes to these true clusterings. This is also useful for being able to compare the results from supervised classification methods with the unsupervised clustering methods.

Purity is a measure that focuses on the frequency of the most common category into each cluster. Purity does not reward the algorithm for grouping items together from the same class, but it does penalize for noise in a cluster (Amigó et al. 2008). To compute purity, each cluster is assigned to the class that is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned objects and dividing by *N*. The purity can be calculated as

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_{k} \max_j |w_k \cap c_j|$$

where $\Omega = \{w_1, w_2, \ldots, w_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes. w_k can be interpreted as the set of samples in w_k and c_j can be interpreted as the set of samples in c_j (Manning et al. 2008). Purity ranges from 0 to 1, where 0 implies a bad clustering and 1 implies a perfect clustering. Purity does not consider the tradeoff between the quality of the clustering against the number of clusters. For example, the purity of a clustering would be 1 if each object were in its own cluster (Manning et al. 2008).

On the other hand, the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) is a different measure to compare clustering results against true clusterings that penalizes both false positives, where two dissimilar objects are assigned to the same cluster, and false negatives, where two similar objects are assigned to different clusters (Manning et al. 2008). The ARI ranges from 0 when the two clusterings have no similarities, to 1 when the clusterings are identical. Let n_{ij} be the number of objects that are in both class u_i and cluster v_j. Let $n_{i.}$ be the number of objects in class u_i and $n_{.j}$ be the number of objects in cluster v_j. The formula for the Adjusted Rand Index is

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} - \sum_j \binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} - \sum_j \binom{n_{.j}}{2}\right]/\binom{n}{2}}$$

The unsupervised *k*-means clustering algorithm was performed using the `stats` package (R Development Core Team 2011) in R. The Adjusted Rand Index was computed using the `mclust` package in R (Fraley and Raftery 2010).

Application of Classification to LIBS Data

As previously stated, an investigation into different techniques for automatic categorization of the century set into groups with similar chemical compositions is a crucial task. Accurate classification can be used for rapid identification of similar unknown samples, allowing us to gain a sense of the distributions of rock types on Mars on a broad scale. Classification can also be used for novelty detection. This involves determining if an unknown sample is different from others previously analyzed. Lastly, classification is critical in the selection of the best possible training set for regression methods used to predict the weight percent oxide of the major elements commonly found in rocks. Using samples that are chemically similar to an unknown sample to train the regression model will produce the most accurate estimates of the unknown sample's composition.

CLASSIFICATION RESULTS

This section explains how the methods explained in the previous section were used with the century set and presents the associated results for evaluating their performance in classifying the century set.

Data Exploration

Data Visualization

The century set was subjected to classical multidimensional scaling and then plotted as a method of data visualization. The samples were color coded by their TAS class as a way to evaluate the similarity of the samples within each TAS class using the Euclidean distance metric. The three-dimensional plot in Figure 10 shows that there is a small amount of natural grouping of samples from the same TAS class, but overall, samples from different classes are generally mixed together. This lack of separation of samples with the same TAS class based on Euclidean distance as the measure of dissimilarity, as shown in the MDS plots, may make classification using Euclidean distance metrics difficult. This will be examined with the use of *k*-means clustering.

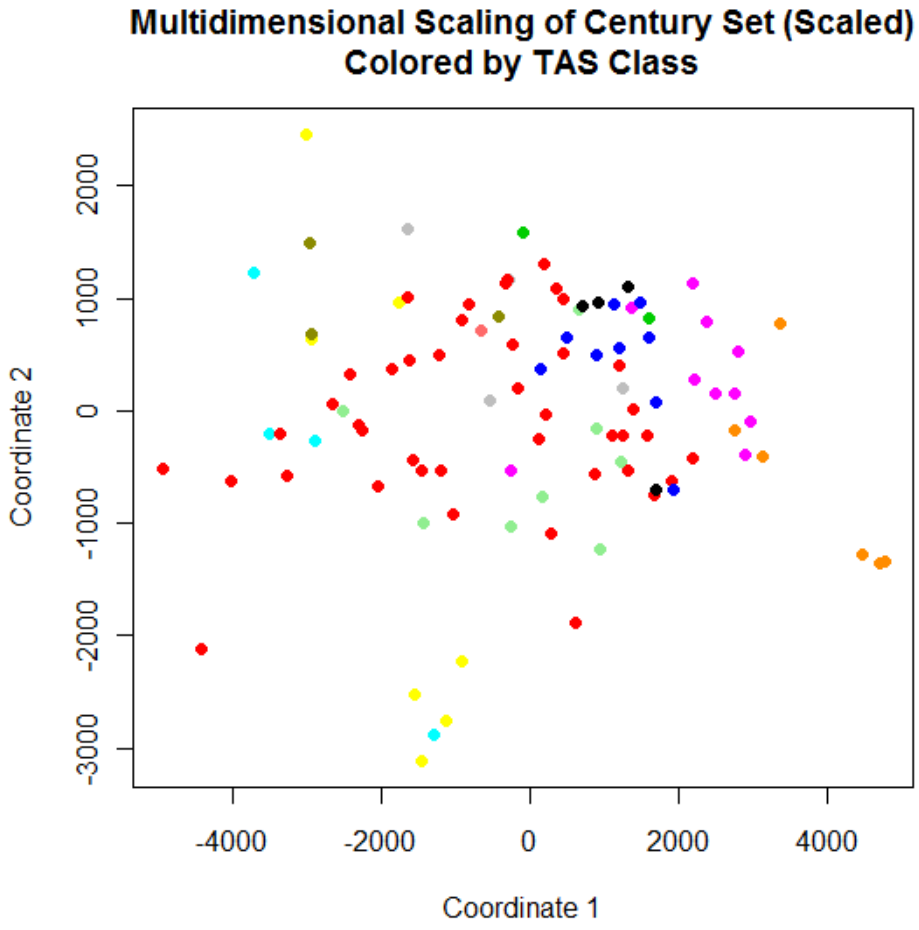

Figure 9: Two-dimensional plot of first two multidimensional scaling coordinates of century set colored by TAS class

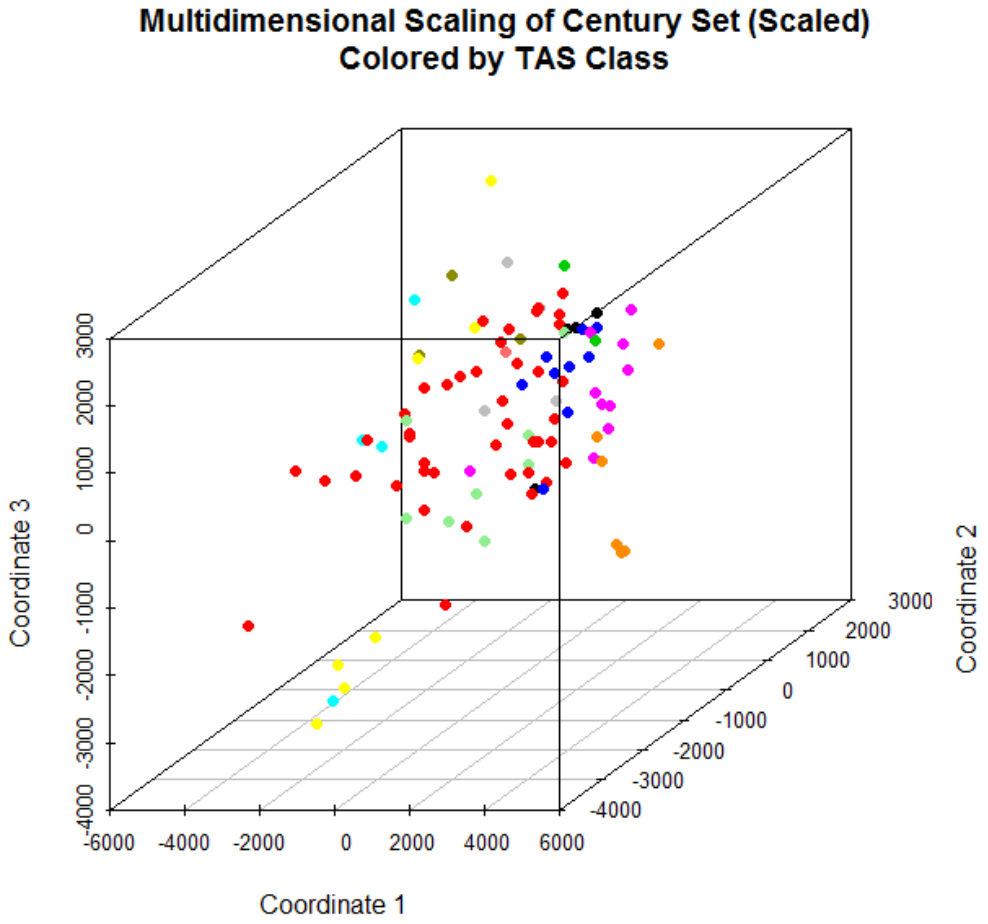

Figure 10: Three-dimensional plot of the first three coordinates based on multidimensional scaling of century set, colored by TAS class

Similar plots can be made using principal components. Figure 11 is a three-dimensional plot of the first three principal components with the samples colored based on their TAS class. Here we see better separation of the TAS classes in this plot as compared to the plot from MDS.

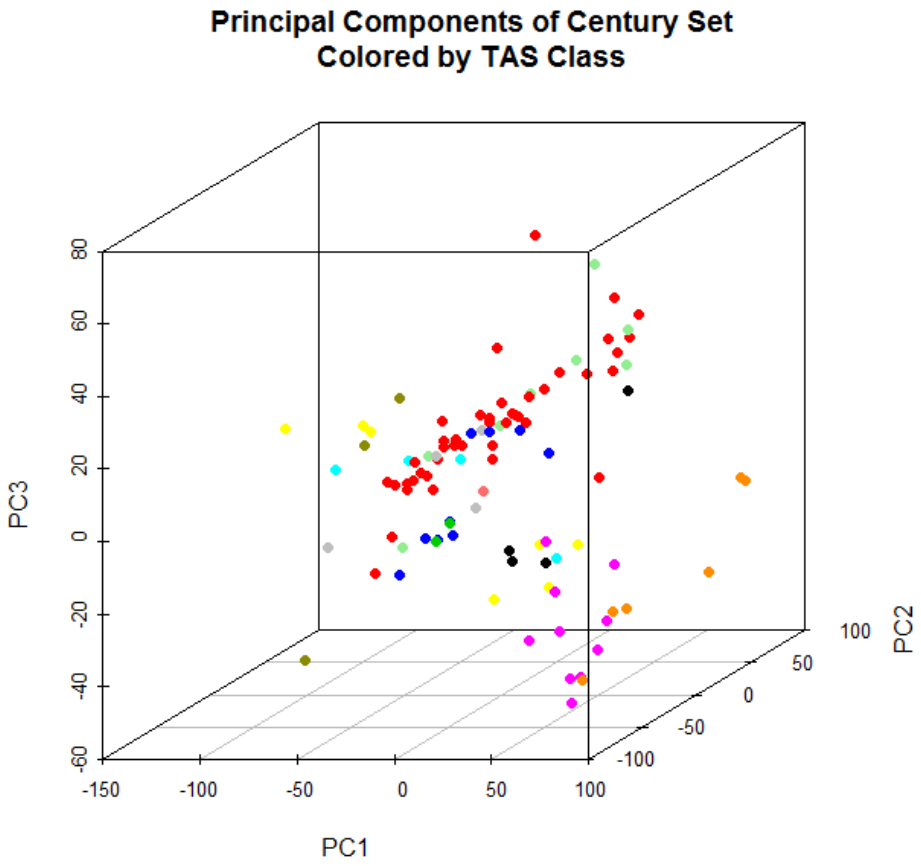

Figure 11: Three-dimensional plot of first three principal components of century set colored by TAS class

Dimension Reduction for Classification

Features (variables) in the data that had the same value across all samples were removed, reducing the number of features to 6,451. This allowed the data to then be centered and scaled in order to account for the possible variations in the data due to differences across samples and shots. Then, principal component analysis was used as a dimension reduction technique. An important aspect in using principal components for training classification algorithms is choosing the appropriate number of components. Generally, this is determined based on an examination of the variance explained by each component. A scree plot showing the amount of variance in the original data that is explained by the first 20 principal components can be seen in Figure 12. Table 2 also displays the standard deviation, proportion of variance, and cumulative proportion of variance explained by several of the principal components. Based on this information, two different numbers of components were selected for use in classification. Six principal components were chosen using the "hinge heuristic" because there is a leveling off or hinge in the total variance explained at around the sixth component, shown in Figure 12. After that, all subsequent components explain less than 2% of the total variance. The principal components beyond the sixth may only reflect noise in the data (Varmuza and Filzmoser 2009, 64). However, it is important to consider that cumulatively, the first six components only explain 65.6% of the variance in the data. Hence, 23 components, which explain a larger proportion of the total variance, 80.2%, will also be used as training data for the

classification algorithms. Using a large number of components such as 23 for classification model training may cause concern for overfitting, but it is interesting to see if the predictive power of the algorithms is very different from the use of six components.

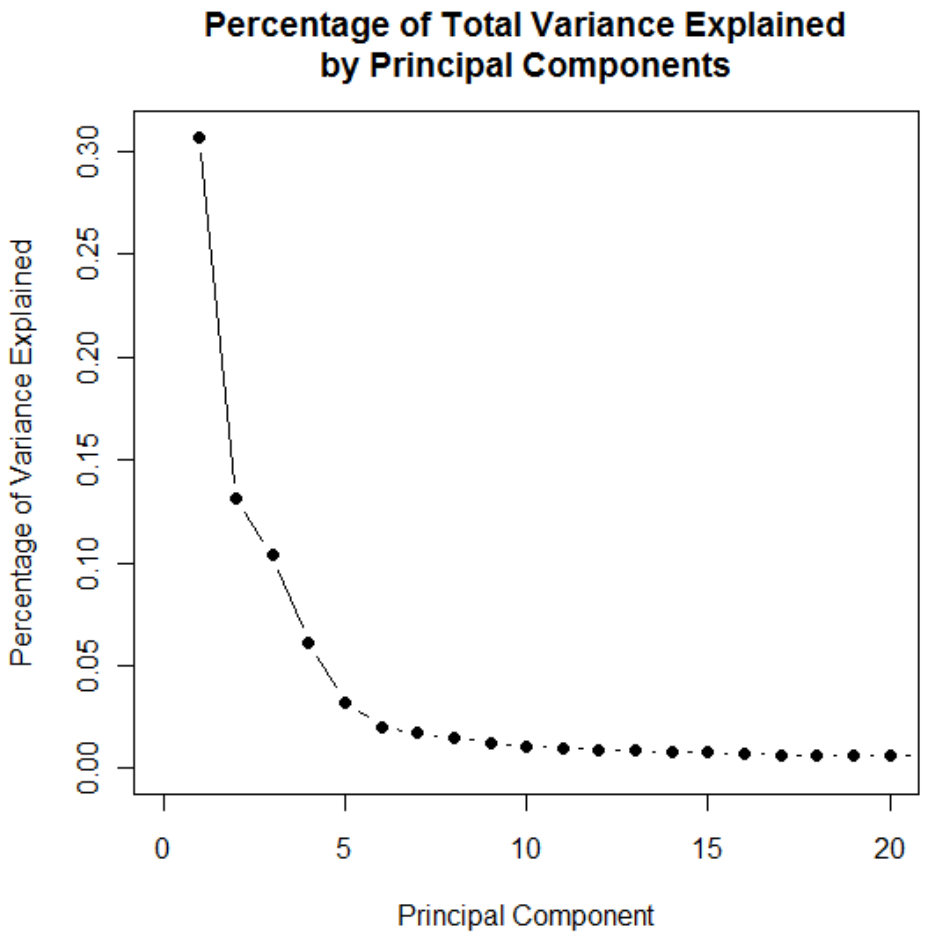

Figure 12: Scree plot: variance explained by principal components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC23	PC48
Standard Deviation	44.52	29.06	25.87	19.85	14.43	11.43	10.46	5.89	4.46
Proportion of Variance	0.307	0.131	0.104	0.061	0.032	0.020	0.017	0.005	0.003
Cumulative Proportion of Variance	0.307	0.438	0.542	0.603	0.635	0.656	0.673	0.802	0.900

Table 2: Importance of the century set principal components

Outlier Removal and Evaluation of Normality

Next the normality of the data was examined and outliers were removed. A Q-Q plot of the first principal component, shown in Figure 13 was constructed to compare the distribution of the century set factors to a theoretical standard normal distribution, which is represented by the straight line. The heavy tails and deviation from the straight line in the Q-Q plot reveal that the first principal component does not appear to be normally distributed. After removing the potential outliers apparent in the plot and creating a new Q-Q plot, the data did not become any more normal. Table 2 shows that the first principal component only explains 44.5% of the variability in the data. Therefore, in an effort to evaluate the normality of a better representation of the data as a whole rather than just the first principal component, we produce Q-Q plots based on the sum of multiple principal components. Assuming that each principal component is normally distributed, the sum of multiple components would also be normally distributed. Q-Q plots are constructed for the sum of six and 23 principal

components separately, as displayed in Figure 14 and Figure 15, for reasons previously explained. Samples that appeared to be outliers in the Q-Q plot were identified and removed from the data. After outlier removal in both cases, we see in Figure 15 and Figure 17 that the normality of the data has improved. The six principal components had six outliers removed and the 23 components had four.

The TAS classes associated with the outliers were examined and are displayed in Table 3 and Table 5. We can see that the outliers removed generally come from separate TAS classes, which gives us more confidence that these are true outliers. If the outliers had come from the same TAS class, it could have been indicative of important differences in the distribution of that particular class.

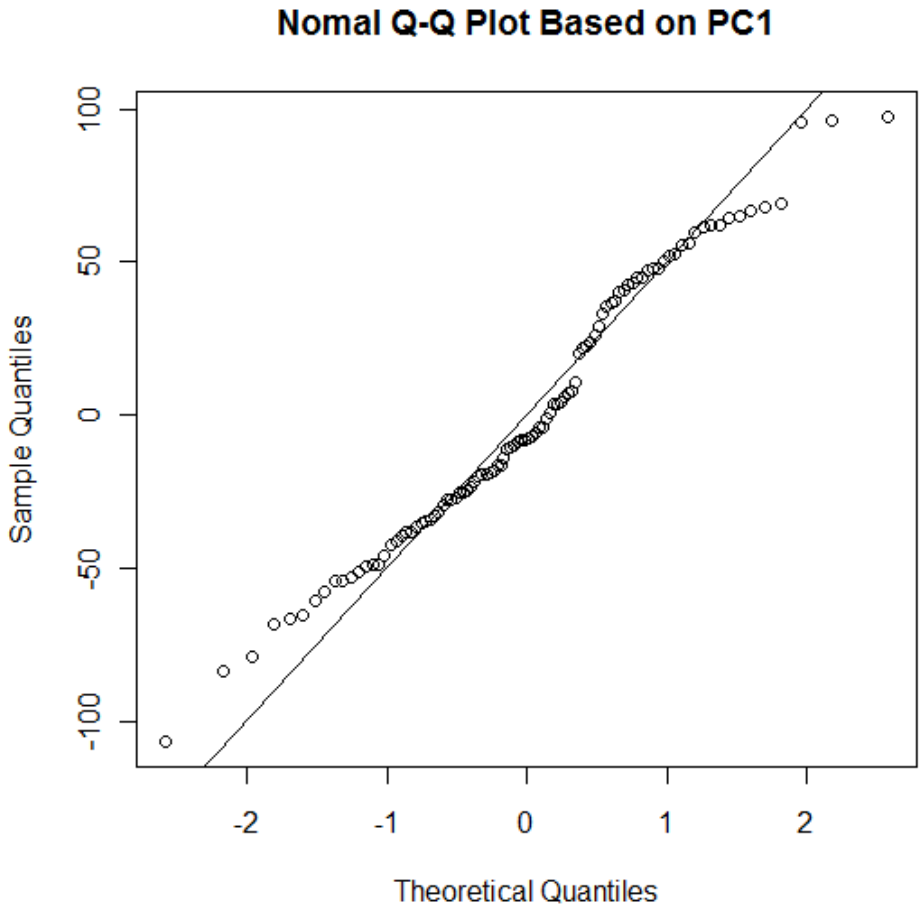

Figure 13: Normal Q-Q Plot based on the first principal component of the century set

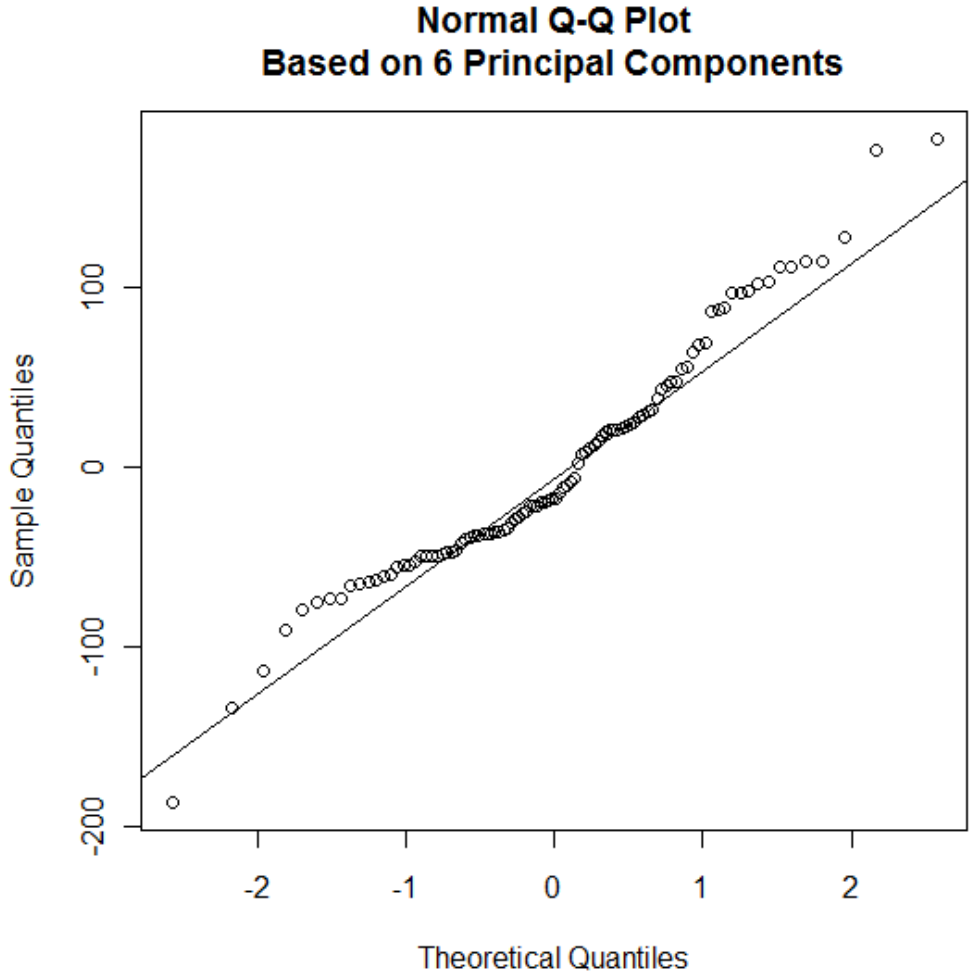

Figure 14: Normal Q-Q Plot of the sum of the first six principal components of the century set

Outliers Removed from 6PCs	
TAS Class	**Frequency**
Andesite	2
Basalt	1
Foidite	1
Picrobasalt	1
Trachybasalt	1

Table 3: TAS class membership of outliers removed from the first six principal components

6PC Century Set (n = 94)	
TAS Class	**Frequency**
Andesite	2
Basalt	41
Basaltic andesite	2
Basaltic trachyandesite	9
Basanite tephrite	4
Dacite	10
Foidite	6
Phonotephrite	4
Picrobasalt	2
Rhyolite	6
Trachyandesite	1
Trachybasalt	7

Table 4: TAS class membership of the samples in data set with first six principal components after outlier removal

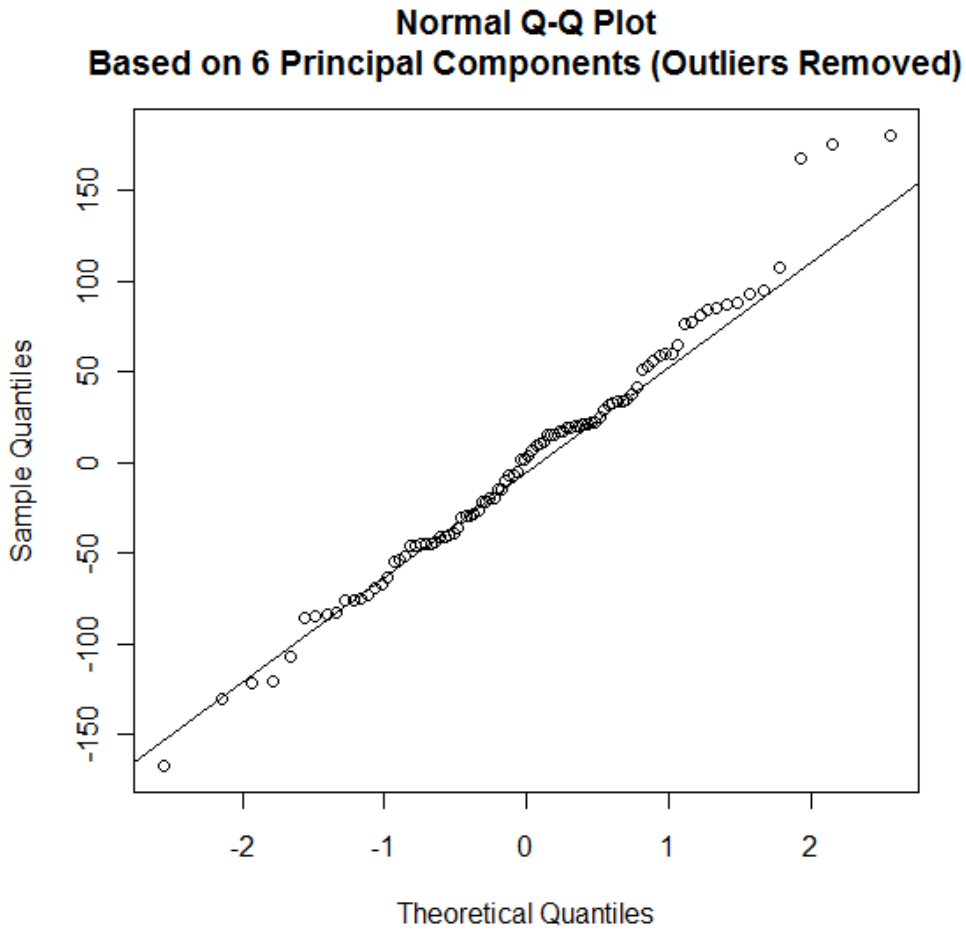

Figure 15: Normal Q-Q Plot of the sum of the first six principal components of the century set after outlier removal

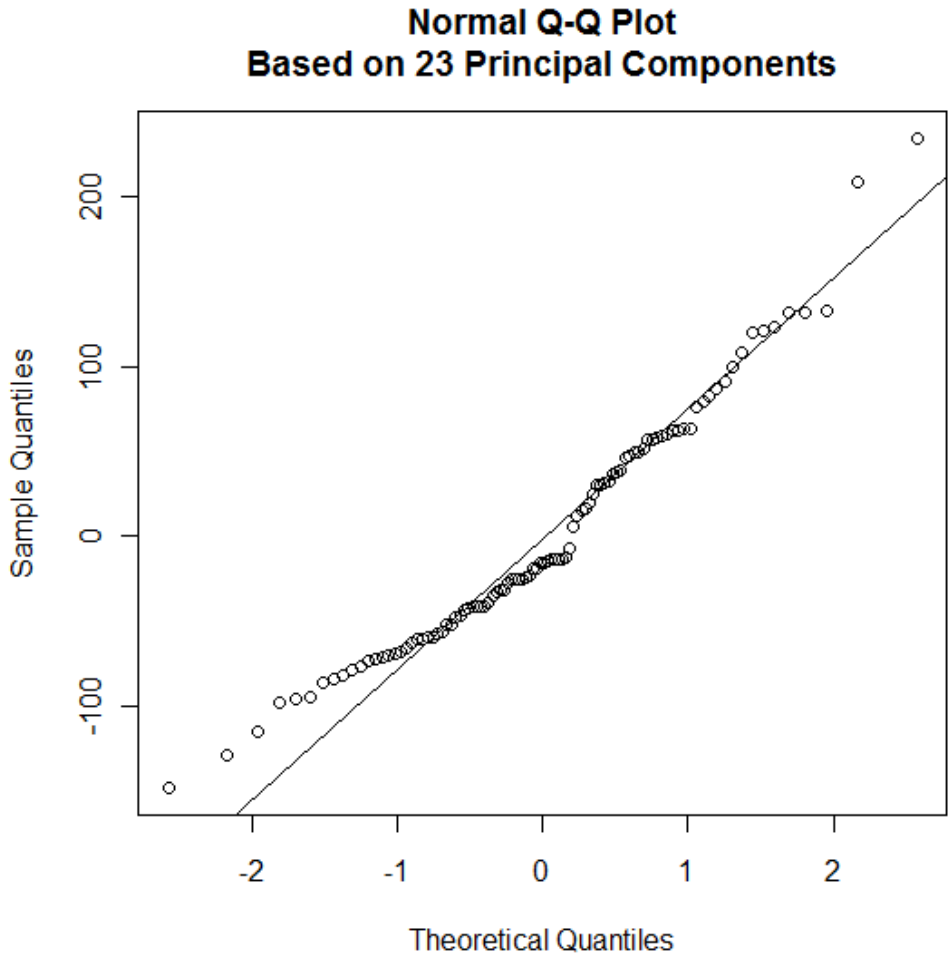

Figure 16: Normal Q-Q Plot of the sum of the first 23 principal components of the century set

Outliers Removed from 23PCs	
TAS Class	**Frequency**
Phonotephrite	1
Picrobasalt	1
Rhyolite	2

Table 5: TAS class membership of outliers removed from the first 23 principal components

23PC Century Set (n = 96)	
TAS Class	**Frequency**
Andesite	4
Basalt	42
Basaltic andesite	2
Basaltic trachyandesite	9
Basanite tephrite	4
Dacite	10
Foidite	7
Phonotephrite	3
Picrobasalt	2
Rhyolite	4
Trachyandesite	1
Trachybasalt	8

Table 6: TAS class membership of the samples in the data set with the first 23 principal components after outlier removal

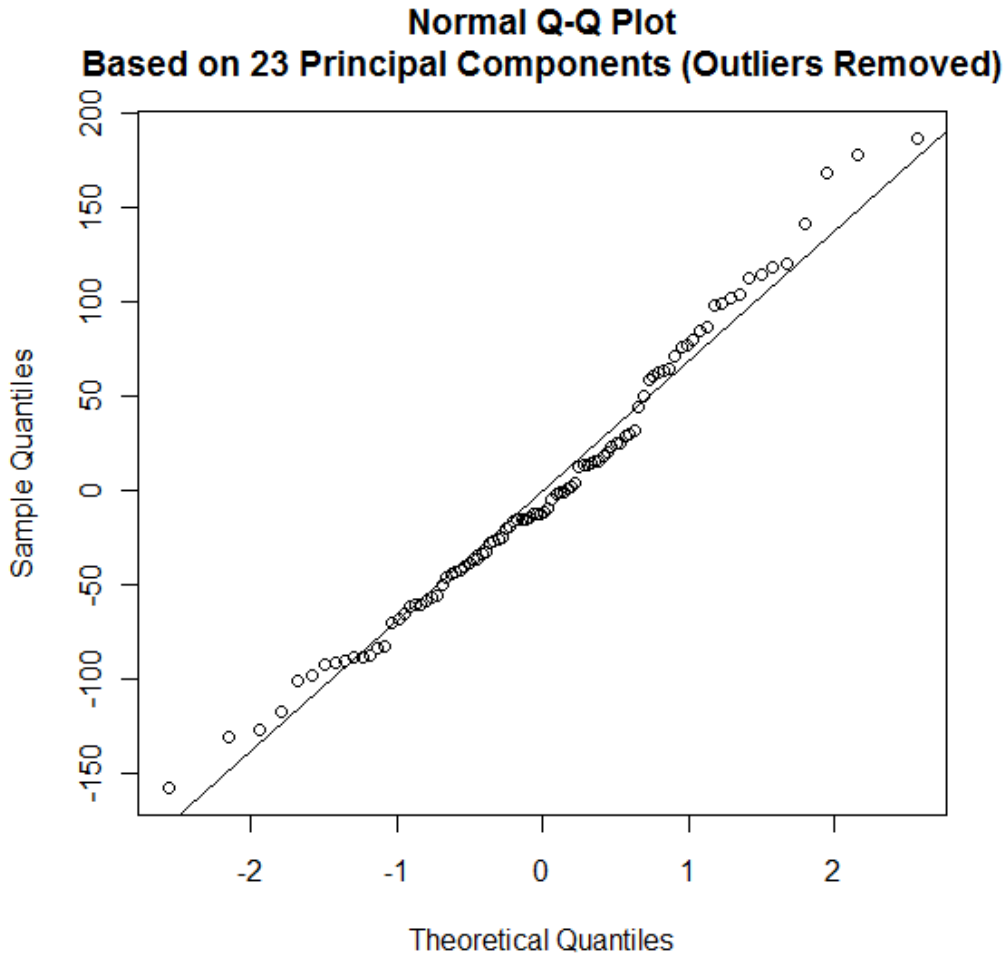

Figure 17: Normal Q-Q Plot of the sum of the first 23 principal components of the century set after outlier removal

Classifiers for separation into 12 TAS classes

k-means Clustering

Table 7 displays the ARI for various k –means clustering outcomes using different subsets of data. The full data set was used (all 6,415 channels in the LIBS spectra), as well as six and 23 principal components. The value of k was set to be 12 and 13, as 12 is the number of true TAS classes and 13 allows for an extra or "garbage" cluster. The ARIs are all extremely small, less than 0.18, indicating that the *k*-means clustering outcomes do not match well with the true TAS classes. The ARI does not change much depending on whether k equals 12 or 13. The ARI is slightly higher when clustering on six principal components.

Data Used	k	Adjusted Rand Index
6 Principal Components	12	0.1607
23 Principal Components	12	0.1558
Full Spectrum (6,415 features)	12	0.1403
6 Principal Components	13	0.1537
23 Principal Components	13	0.1707
Full Spectrum (6,415 features)	13	0.1503

Table 7: Adjusted Rand Indices for *k*-means clustering outliers for k = 12, 13

Table 8 shows a closer examination of one of the clustering outcomes with k = 12 using the full LIBS spectrum (6,415 features) for each sample. The samples assigned to each cluster were examined in order to find the most common TAS class. Then the purity of each cluster was calculated by finding the

proportion of samples that belong to the most common TAS class for each cluster. This can be considered a measure of accuracy because we hope that each cluster should hold samples from the same TAS class. The overall purity for the clustering algorithm was then calculated to be 0.6. This number is deceivingly high after closer examination of the makeup of the samples in each cluster because purity does not penalize for grouping items from the same class in different clusters. We can see that eight of the 12 clusters are comprised of mostly basalts, only two of which are clusters of only basalt samples. This suggests that the *k*-means algorithm is highly affected by the disproportionate number of basalt samples in the century set. Ideally, we would hope to see one cluster containing all of the basalt samples.

Cluster Size	Most Common TAS class	Cluster Purity
13	Dacite	0.6923
5	Foidite	0.8000
16	Basalt	0.5000
11	Basalt / Foidite	0.2727
3	Rhyolite	1.000
10	Basalt	0.7000
10	Basalt	0.8000
5	Basalt	0.6000
7	Basalt	0.7143
2	Basalt	1.000
11	Basaltic Trachyandesite	0.4545
7	Basalt	0.4286
Overall Purity = 0.6		
Adjusted Rand Index = 0.1403		

Table 8: *k*-means clustering on full spectra (*k* = 12). Most common TAS class and cluster purity, the proportion of samples in the cluster that are contained in the most common TAS class, is displayed for each cluster.

Discriminant Analysis

We can see that in all cases other than for SLDA, accuracy rates and Kappa statistics were the highest when using the first six principal components to train the classifier. SLDA is essentially performing dimension reduction on the 23 components, which is most likely why it performed better with a larger number of components. The best classifier based on accuracy, SDA with six principal components, had an accuracy rate of 60.8% and a Kappa value of 0.448. Accuracy rates ranged from approximately 50 to 60% with Kappa statistics all

less than 0.5 with values as low as 0.25, indicating fairly poor classification results.

Full Spectra, All 12 TAS Classes (n = 100)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.496	0.252	0.0968	0.139	n prune = 3
PDA	0.539	0.373	0.0894	0.137	lambda = 1 : 20
SDA	0.551	0.427	0.261	0.284	

6 Principal Components, All 12 TAS Classes (n = 94)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.55	0.273	0.0915	0.171	n prune = 4 : 14
LDA	0.606	0.452	0.142	0.164	
PDA	0.6	0.452	0.127	0.167	lambda = 1 : 20
SDA	0.608	0.448	0.166	0.226	
SLDA	0.53	0.259	0.0938	0.142	

23 Principal Components, All 12 TAS Classes (n = 96)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.52	0.214	0.0786	0.0912	n prune = 2
LDA	0.573	0.422	0.127	0.167	
PDA	0.53	0.395	0.16	0.188	lambda = 1 : 20
SDA	0.526	0.323	0.102	0.156	
SLDA	0.579	0.377	0.119	0.168	

Table 9: Discriminant analysis results from 10-fold cross validation for classification into all 12 TAS classes

Support Vector Machines

Once again, using the first six principal components produces the best prediction results for algorithms using SVM. The best performance by far is seen in the use of the linear kernel with six principal components with an accuracy rate of approximate 65%. However, the high number of support vectors, 76, may suggest overfitting or could be because of the large number of classes and small sample size.

Full Spectra, All 12 TAS Classes (n = 100)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.42	100	0.167	1
Sigmoid	0.42	68	0.167	1

6 Principal Components, All 12 TAS Classes (n = 94)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.585	88	0.125	0.5
Sigmoid	0.436	77	4	0.03125
Linear	0.649	76	3.05e-05	1

23 Principal Components, All 12 TAS Classes (n = 96)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.438	96	0.5	2
Sigmoid	0.438	84	0.25	0.125
Linear	0.438	93	3.05e-05	1

Table 10: Support vector machine results from 10-fold cross validation for classification into all 12 TAS classes

Summary

All variations of the three main methods produced mediocre results when attempting to classify into 12 TAS classes. Using six principal components seems to produce the best results. Support vector machines generally perform the best; however, we must take into consideration the high number of support vectors that may suggest overfitting.

Removing TAS Classes with Small Sample Size

The poor results from our attempt to classify into groups based on the 12 TAS classes prompted a further investigation into the effects of having classes that contain a very small number of samples on classification accuracy. TAS classes that contained less than four samples were removed from each of the data sets. The new distributions of TAS classes after removing small groups are displayed in Table 11 and Table 12 for each of the data sets containing different numbers of principal components separately.

6PC Century Set Small TAS Classes Removed (where $n < 4$) ($n = 87$, # TAS classes = 8)	
TAS Class	**Frequency**
Basalt	41
Basaltic trachyandesite	9
Basanite tephrite	4
Dacite	10
Foidite	6
Phonotephrite	4
Rhyolite	6
Trachybasalt	7

Table 11: Frequency of TAS classes for six principal components after removal of small classes where $n < 4$. The TAS classes that were removed include andesite, basaltic andesite, picrobasalt and trachyandesite.

23PC Century Set Small TAS Classes Removed ($n < 4$) ($n = 88$, # TAS classes = 8)	
TAS Class	**Frequency**
Andesite	4
Basalt	42
Basaltic trachyandesite	9
Basanite tephrite	4
Dacite	10
Foidite	7
Rhyolite	6
Trachybasalt	8

Table 12: Frequency of TAS classes for 23 principal components after removal of small classes where $n < 4$. The TAS classes that were removed include basaltic andesite, phonotephrite, picrobasalt, and trachyandesite.

k-means Clustering

Purity values and ARIs increased slightly when the small TAS classes were removed; however, since the ARI is still low at around 0.2, the *k*-means clustering outcome is very different from the true TAS class assignment.

Data Used	*k*	Adjusted Rand Index	Purity
6 Principal Components	8	0.2173	0.6897
23 Principal Components	8	0.1942	0.6364

Table 13: Adjusted Rand Indices for *k*-means clustering where $k = 8$ after small TAS classes were removed

Discriminant Analysis

For both six principal components and 23 principal components, the best results based on both accuracy and Kappa were from linear discriminant analysis. Slightly better results are obtained from the use of 23 components, with an accuracy rate for classification of approximately 70% and Kappa value of about 0.56.

6 Principal Components, Small TAS Classes Removed (n = 89, # TAS Classes = 9)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.629	0.448	0.157	0.238	n prune = 5
LDA	0.656	0.512	0.112	0.124	
PDA	0.64	0.49	0.158	0.204	lambda = 1 : 20
SDA	0.61	0.412	0.129	0.212	
SLDA	0.621	0.387	0.103	0.182	

23 Principal Components, Small TAS Classes Removed (n = 88, # TAS Classes = 8)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.635	0.452	0.15	0.225	n prune = 6 : 29
LDA	0.698	0.563	0.141	0.184	
PDA	0.624	0.485	0.202	0.28	lambda = 1 : 20
SDA	0.581	0.375	0.134	0.223	
SLDA	0.642	0.422	0.0833	0.121	

Table 14: Discriminant analysis results from 10-fold cross validation for classification into TAS classes after removal of small classes

Support Vector Machines

For both six principal components and 23 principal components, the best results occurred with the use of the linear kernel. The accuracy rates for classification with the linear kernel using six and 23 components were almost identical at around 70%, but the model trained on six components had seven fewer support vectors.

6 Principal Components, Small TAS Classes Removed (n = 89, # TAS Classes = 9)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.621	80	0.25	1
Sigmoid	0.471	67	8	0.03125
Linear	0.701	63	3.05e-05	1

23 Principal Components, Small TAS Classes Removed (n = 88, # TAS Classes = 8)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.477	72	3.05e-05	0.3125
Sigmoid	0.477	71	3.05e-05	0.3125
Linear	0.697	70	3.05e-05	1

Table 15: Support vector machine results from 10-fold cross validation for classification into TAS classes after removal of small classes

Summary

Classification accuracy rates did improve slightly after the removal of small TAS classes. In the *k*-means clustering and support vector machines, training the model on six principal components once again produced better results. Accuracy rates for discriminant analysis were slightly higher with 23 components. Support vector machines produce the highest single accuracy rate, 70.1%, but the number of support vectors is still high at 63.

Binary Classifiers

In order to evaluate the performance of these methods for a simpler classification task with the century set data, we can consider binary classification. Since our data is comprised of mostly basalts, the data can be divided fairly evenly into two groups by considering if a sample belongs to a TAS class in the basalt family (basalt, basaltic andesite, basaltic trachyandesite, and trachybasalt) or if it falls into one of the other eight TAS classes.

First, the data is replotted in three dimensions with the use of principal components and colored based on this binary division. Figure 18 displays the plot and shows that there is good separation of samples from these two groups.

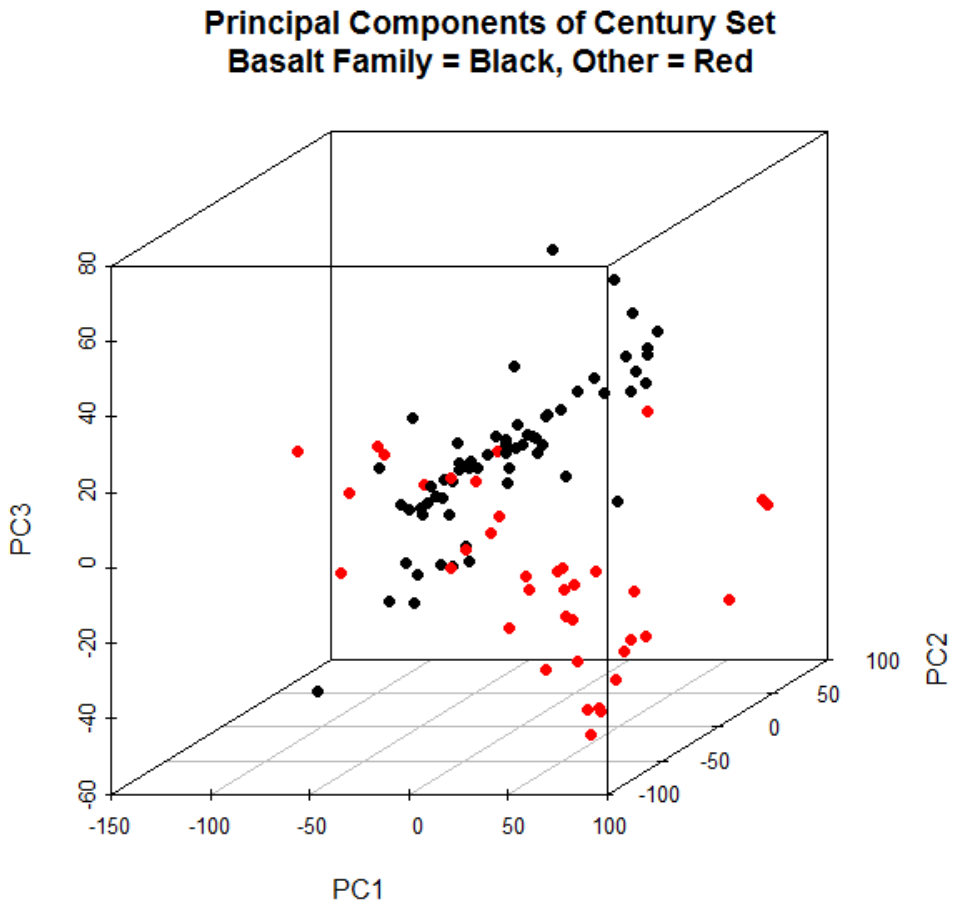

Figure 18: Three-dimensional plot of principal components colored by binary classification. Black points are samples from the basalt family of TAS classes and the red points are samples from all other TAS classes.

k-means Clustering

Once again, as shown in Table 16, purity values are moderate, but the ARIs are extremely low at around 0.1. The *k*-means algorithm does not perform well when dividing the samples into groups with similar TAS classes, even in this simpler classification task.

Data Used	*k*	Adjusted Rand Index	Purity
6 Principal Components	2	0.1176	0.6809
23 Principal Components	2	0.0981	0.6667

Table 16: Adjusted Rand Index and Purity for k-means clustering when $k = 2$

Discriminant Analysis

We see much higher accuracy rates for this binary classification as compared to the multi-label classification. Flexible discriminant analysis had the lowest errors for six components but all methods perform equally well when using 23 components.

6 Principal Components, Binary Classification (Basalt Family/Other) (n = 94)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.849	0.656	0.14	0.312	n prune = 7
LDA	0.799	0.516	0.123	0.307	
PDA	0.828	0.593	0.106	0.259	lambda = 1 : 20
SDA	0.807	0.548	0.0984	0.227	
SLDA	0.705	0.244	0.0979	0.281	

23 Principal Components, Binary Classification (Basalt Family/Other) (n = 96)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Parameter**
FDA	0.867	0.691	0.0922	0.204	n prune = 6
LDA	0.854	0.658	0.0523	0.137	
PDA	0.867	0.691	0.135	0.31	lambda = 1 : 20
SDA	0.853	0.649	0.0872	0.214	
SLDA	0.886	0.73	0.0562	0.128	

Table 17: Discriminant analysis results from 10-fold cross validation for binary classification

Support Vector Machines

For SVM, the radial kernel function performed the best for six components and 23 components and the sigmoid kernel performed equally as well with 23 components. Accuracy rates are around 85%, which is much higher than in mutli-label classification. The number of support vectors is much lower for this binary classification as compared to multi-label classification.

6 Principal Components, Binary Classification (Basalt Family/Other) (n = 94)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.862	54	0.125	0.5
Sigmoid	0.830	40	0.002	128
Linear	0.830	38	3.05e-05	1

23 Principal Components, Binary Classification (Basalt Family/Other) (n = 96)				
SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Gamma**	**Cost**
Radial	0.875	50	3.05e-05	1024
Sigmoid	0.875	44	0.0039	32
Linear	0.844	37	3.05e-05	1

Table 18: Support vector machine results from 10-fold cross validation for binary classification

Summary

Results from this binary classification task are much improved over the results from the two attempts at multi-label classification. SLDA with 23 components produces the highest accuracy rate, 88.6%, which is very good. Support vector machines also perform well overall, especially with the use of the radial kernel. The number of support vectors is much lower in the binary classification task as compared to the multi-label classification tasks. This implies that the high number of support vectors seen in the multi-label classification is most likely due to a combination of the many boundaries modeled between classes and the small sample size. This reduces the concern for overfitting in the SVM algorithms.

MODEL TO SIMULATE SPECTRA

Methods

The results in the previous section showed an increase in accuracy rates for classification when small TAS classes were removed and an even bigger increase for a simpler classification task where the sample size in each class was larger. This suggests that the small sample size and unequal balance of samples from different TAS classes may be having a significant impact on the performance of attempted algorithms. With the scarcity of LIBS data but the abundance of data providing the chemical composition of rock samples, this led to the idea for a second investigation. What if we could accurately simulate the LIBS spectrum of a rock sample based on its chemical composition? If we could come up with a valid model that could produce realistic spectra, we could increase the sample size of the data we have to work with and even out the distribution of different sample types. This could be extremely useful for evaluating the effects of sample size on the performance of the classification algorithms.

Available to work with are the LIBS spectra from the binary compounds commonly found in rocks. If we make the assumption that the LIBS spectrum for each sample is the combination of LIBS spectra from the binary compounds that make up the rock sample weighted by the proportion of that compound contained

in the rock, these data are extremely useful. Following this idea, the formula to simulate the LIBS spectrum for one sample is

$$\overrightarrow{\text{predicted spectrum}} = \sum_{e \in \text{Major Elms}} \frac{\text{weight \% oxide}_e}{100} \times \overrightarrow{\text{binary spectrum}_e}$$

In other words, we produce a weighted sum of LIBS spectrum for each of the 10 major elements found in a rock sample, weighted by the percent oxide of each element found in the sample. It must be noted that this model does not account for the chemical matrix effect that is associated with variations in the peak intensities.

To evaluate the model, we can calculate a mean squared error for a single sample:

$$\text{Sample MSE} = \frac{1}{k}\sum_{i=1}^{k}(\text{predicted Intensity}_i - \text{true Intensity}_i)^2$$

where i represents the i^{th} wavelength channel. Using these sample errors, the formula used to calculate the overall error across all samples in the century set is

$$\text{Overall MSE} = \frac{1}{n}\sum_{j=1}^{n}\text{Sample MSE}_j$$

where n is the number of samples (100). One MSE is calculated to compare the true century set spectra with the simulated spectra based on our model. Also, we can permute the samples to create a new data set of simulated spectra. A second

MSE can be calculated that compares the true spectra with the permuted spectra. Since spectral data have many values at zero with occasional peaks, the MSEs will be small by nature. Therefore, this second MSE gives us a baseline value that can help us evaluate the relative significance of the first MSE calculated with the model.

Results

The MSEs are displayed in Table 19. The MSE for the model predictions is small, as expected, but it is actually larger than the baseline MSE calculated with the permuted samples as the predicted spectra. To further investigate the appropriateness of this model, several plots were created comparing the true and simulated spectra for individual samples. An example of this is shown in Figure 19. The simulated spectra are quite different than the true spectra for the majority of these different plots. Overall, this shows us that because of the interaction due to the chemical matrix effect, this is not an accurate model for simulating spectra from chemical compositions. Interaction terms should be included in the model in order to make improvements. Unfortunately, time did not permit for further exploration in the scope of this thesis.

Source	MSE
Model	2.17e-07
Permutation	5.25e-08

Table 19: Calculated mean squared error for results based on proposed model and MSE calculated from a permutation of samples

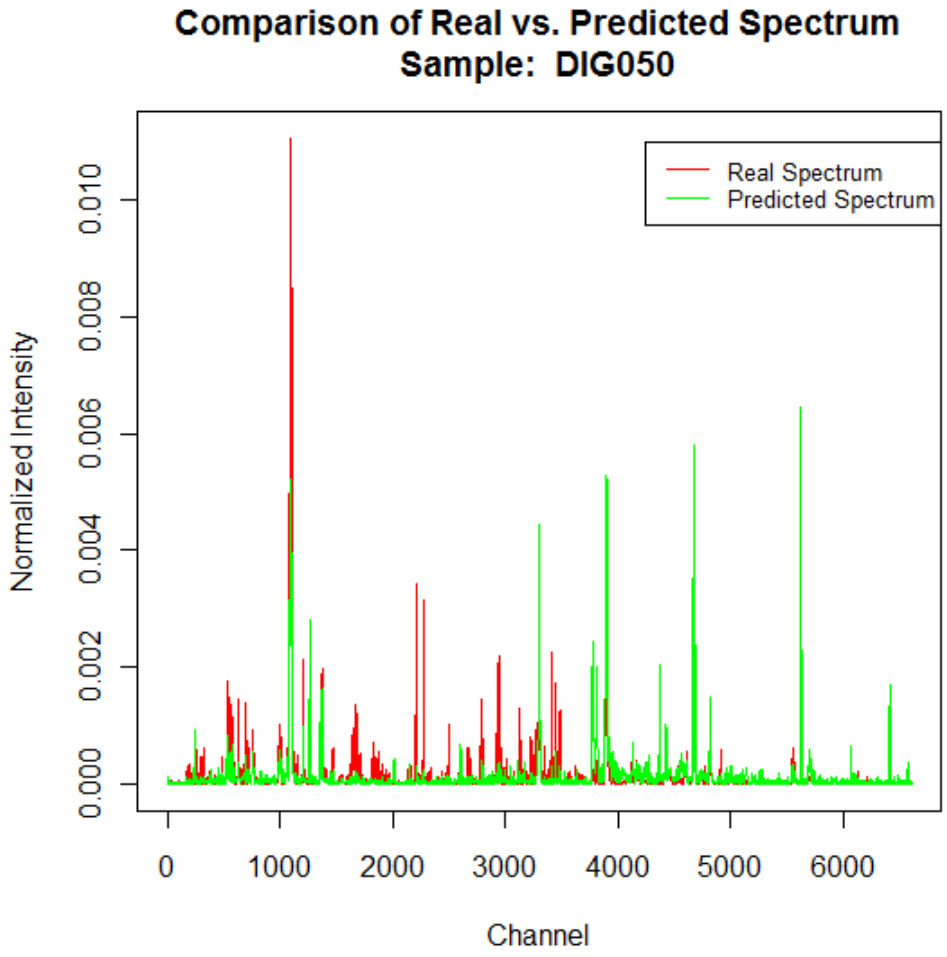

Figure 19: An example of true versus simulated spectrum for one sample

CONCLUSION

Summary of Findings

The methods explored in this thesis were not able to classify the samples of the century set into groups based on TAS classes accurately enough for practical use. Yet, we see that by removing a few TAS classes that contained a small number of samples, we can improve the classification error rates. Additionally, when we perform a binary classification task where the sample size of both classes is roughly equal, we see much more reasonable results. While being able to determine if a sample is a basalt or not is useful in some ways, being able to distinguish a larger number of rock types (based on multi-label classification) is more advantageous.

Overall, support vector machines produce classifiers with the lowest error rates for classification for each of the three categorization tasks based on 10-fold cross validation, but the number of support vectors used in the model is always high. While this sometimes suggests overfitting, it may not in this situation because of the small sample size. Discriminant analysis constructs classifiers that perform almost as well as support vector machines without the risk of overfitting. On the other hand, the *k*-means algorithm performs poorly in all situations, even when the classification task was simple. This does show that the *k*-means clustering algorithm is not very effective at separating samples from

different TAS classes and seems to be affected by the unequal distribution of samples from different classes. The natural structure of the data based on Euclidean distance does not seem to provide good separation of samples based on TAS class.

A comparison of the results based on the number of principal components used shows different outcomes based on the classification task and the data used. The use of 23 principal components in the training set as compared to six does not impact the results for multi-label classification. However, we see that in binary classification, training with 23 components always produced higher accuracy rates as compared to the rates from using six components in the training set.

Discussion

There are many possible factors playing into the inability to classify the samples of the century set into groups based on TAS classes using the methods explored in this thesis. These are mainly normal based methods, so the potential violation of the normality assumptions is a likely possibility. The improved performance of the supervised classifiers for simpler tasks, however, may suggest otherwise. This leads to the possibility that the small sample sizes and uneven spread of samples in different classes is the root of the problem. It is also possible that the results are better strictly because the binary classification task is much simpler as compared to a multi-label classification task. This of course warrants

further investigation. The best way to explore this further is to obtain more data and reevaluate.

The *k*-means algorithm did not produce clusterings that were similar to the true groupings we created using TAS classes, even in the simple binary task. As was expected based on the MDS plots, this suggests that when using the Euclidean distance as a similarity metric, the samples do not form natural clusters based on their TAS class. In other words, the true class indices we selected, TAS classes, don't correspond well with the natural structure of the data. It is possible that the use of a different distance metric could produce better results. It should also be considered that creating true groups of similar samples with a method other than TAS classes could yield different results.

It appears that discriminant analysis and support vector machines have the potential to accurately classify LIBS spectra given a proper training set containing a larger number of samples and an even spread of samples in the different true classes. However, unsupervised methods should not be ruled out completely either.

Future Work

More data is necessary to evaluate the effects of sample size and sample imbalance on the error rates of classification. Obtaining more real data would of course be ideal, but if it is not possible to do so in a timely manner, it may be worth exploring models for simulating spectra further. As was seen with the

attempt to create a model to simulate spectra, more complicated models that take the elemental interaction due to the chemical matrix effect into account would likely be better. A good model that uses partial least squares regression is already in place to estimate elemental compositions from LIBS spectra. Taking this model and running it backwards could be a viable option, although it would be complicated. Using simulated data for classification of course adds more possible variation in the results.

There are many possibilities of other classification and unsupervised learning algorithms that could be examined. Once more data is obtained, it is certainly worth further exploration into more robust methods that are more complicated, but may produce better results. For example, mixture modeling using the expectation maximization (EM) algorithm might be worth trying because of its flexibility in choosing component distributions. While the *k*-means algorithm makes a *hard* assignment of data points to clusters where each cluster is associated with only one cluster, the EM algorithm makes a *soft* assignment based on a probabilistic model (Bishop 2006, 443). Also, different distance metrics could be examined in *k*-means clustering. There are also other kernel functions for SVM that could be useful, such as the Fourier kernel.

Final Remarks

This thesis shows that it is critical to use rigorous methods for error analysis in the classification of LIBS data. This work has provided evidence against the optimistic results of many of the LIBS classification studies that did not use statistical best practices. There is an essential need for more data before reevaluating these and other methods for the classification of LIBS data.

APPENDIX

Note: These are the same results displayed in earlier sections, however, here they are displayed in a different format that allows for easier comparison of results from different methods.

Key of abbreviations:

FDA = Flexible Discriminant Analysis (with tuning parameter n prune)

LDA = Linear Discriminant Analysis

PDA = Penalized Discriminant Analysis (with tuning parameter lambda)

SDA = Shrinkage Discriminant Analysis

SLDA = Stabilized Linear Discriminant Analysis

Full Spectra, All 12 TAS Classes (n = 100)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.496	0.252	0.0968	0.139	n prune = 3
PDA	0.539	0.373	0.0894	0.137	lambda = 1:20
SDA	0.551	0.427	0.261	0.284	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.42	100	0.167	1
Sigmoid	0.42	68	0.167	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
12	0.1403
13	0.1503

6 Principal Components, All 12 TAS Classes (n = 94)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.55	0.273	0.0915	0.171	n prune = 4 : 14
LDA	0.606	0.452	0.142	0.164	
PDA	0.6	0.452	0.127	0.167	lambda = 1 : 20
SDA	0.608	0.448	0.166	0.226	
SLDA	0.53	0.259	0.0938	0.142	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.585	88	0.125	0.5
Sigmoid	0.436	77	4	0.03125
Linear	0.649	76	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
12	0.1607
13	0.1537

23 Principal Components, All 12 TAS Classes (n = 96)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.52	0.214	0.0786	0.0912	n prune = 2
LDA	0.573	0.422	0.127	0.167	
PDA	0.53	0.395	0.16	0.188	lambda = 1 : 20
SDA	0.526	0.323	0.102	0.156	
SLDA	0.579	0.377	0.119	0.168	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.438	96	0.5	2
Sigmoid	0.438	84	0.25	0.125
Linear	0.438	93	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
12	0.1558
13	0.1707

6 Principal Components, Small TAS Classes Removed (n = 89, # TAS Classes = 9)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.629	0.448	0.157	0.238	n prune = 5
LDA	0.656	0.512	0.112	0.124	
PDA	0.64	0.49	0.158	0.204	lambda = 1 : 20
SDA	0.61	0.412	0.129	0.212	
SLDA	0.621	0.387	0.103	0.182	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.621	80	0.25	1
Sigmoid	0.471	67	8	0.03125
Linear	0.701	63	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
8	0.1939

23 Principal Components, Small TAS Classes Removed (n = 88, # TAS Classes = 8)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.635	0.452	0.15	0.225	n prune = 6 : 29
LDA	0.698	0.563	0.141	0.184	
PDA	0.624	0.485	0.202	0.28	lambda = 1 : 20
SDA	0.581	0.375	0.134	0.223	
SLDA	0.642	0.422	0.0833	0.121	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.477	72	3.05e-05	0.3125
Sigmoid	0.477	71	3.05e-05	0.3125
Linear	0.697	70	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
8	0.1942

6 Principal Components, Binary Classification (Basalt Family/Other) (n = 94)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.849	0.656	0.14	0.312	n prune = 7
LDA	0.799	0.516	0.123	0.307	
PDA	0.828	0.593	0.106	0.259	lambda = 1 : 20
SDA	0.807	0.548	0.0984	0.227	
SLDA	0.705	0.244	0.0979	0.281	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.862	54	0.125	0.5
Sigmoid	0.830	40	0.002	128
Linear	0.830	38	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
2	0.1176

23 Principal Components, Binary Classification (Basalt Family/Other) (n = 96)					
DISCRIMINANT ANALYSIS					
Method	**Accuracy**	**Kappa**	**Accuracy SD**	**Kappa SD**	**Optimal Parameter**
FDA	0.867	0.691	0.0922	0.204	n prune = 6
LDA	0.854	0.658	0.0523	0.137	
PDA	0.867	0.691	0.135	0.31	lambda = 1 : 20
SDA	0.853	0.649	0.0872	0.214	
SLDA	0.886	0.73	0.0562	0.128	

SUPPORT VECTOR MACHINES				
Kernel	**Accuracy**	**# Support Vectors**	**Optimal Gamma**	**Optimal Cost**
Radial	0.875	50	3.05e-05	1024
Sigmoid	0.875	44	0.0039	32
Linear	0.844	37	3.05e-05	1

K-MEANS CLUSTERING	
k	**Adjusted Rand Index**
2	0.1942

BIBLIOGRAPHY

Ahdesmaki, Miika, and Korbinian Strimmer. 2010. Feature selection in omics prediction problems using CAT scores and false nondiscovery rate control. *The Annals of Applied Statistics* 4 (1): 503-19.

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12 (4) (AUG): 461-86.

Anzano, Jesus, Beatriz Bonilla, Beatriz Montull-Ibor, Roberto-Jesus Lasheras, and Justiniano Casas-Gonzalez. 2010. Classifications of plastic polymers based on spectral data analysis with laser induced breakdown spectroscopy. *Journal of Polymer Engineering* 30 (3-4) (MAY-JUN): 177-87.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37-46.

Cremers, David A., and Leon J. Radziemski. 2006. *Handbook of laser-induced breakdown spectroscopy.* Hoboken: John Wiley & Sons, Ltd.

Dimitriadou, Evgenia, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. 2011. e1071: Misc functions of the department of statistics (`e1071`). *R Package Version 1.6,* TU Wien, http://CRAN.R-project.org/ package=e1071.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. *Pattern classification*. New York: Wiley.

Fraley, C., and Raftery, A. E. 2010. `MCLUST` version 3 for R: Normal mixture modeling and model-based clustering, Technical Report No. 504. Department of Statistics, University of Washington.

Gokcen, Ibrahim, and Jing Peng. 2002. Comparing linear discriminant analysis and support vector machines. *Advances in Information Systems* 2457: 104-13.

Gross, Shulamith T. 1986. The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 42 (4) (DEC): 883-93.

Hastie, Trevor, Andreas Buja, and Robert Tibshirani. 1995. Penalized discriminant analysis. *The Annals of Statistics* 23 (1): 73-102.

Hastie, Trevor, Robert Tibshirani, and Andreas Buja. 1994. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* 89 (428): 1255-70.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. Springer series in statistics. New York: Springer.

Holland, Steven M. Principal components analysis (PCA). 2008. Available from http://strata.uga.edu/software/pdf/pcaTutorial.pdf.

Hubert, Lawrence, and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2 (1): 193-218.

Karatzoglou, Alexandros, David Meyer, and Kurt Hornik. 2006. Support vector machines in R. *Journal of Statistical Software* 15 (9) (APRIL).

Kuhn, Max. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28 (5) (NOV).

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, and Allan Engelhardt. 2012. Caret: Classification and regression training. *R Package Version 5.13-20*, http://CRAN.R-project.org/package=caret.

Lasue, J., R. C. Wiens, T. F. Stepinski, O. Forni, S. M. Clegg, S. Maurice, and ChemCam Team. 2011. Nonlinear mapping technique for data visualization and clustering assessment of LIBS data: Application to ChemCam data. *Analytical and Bioanalytical Chemistry* 400 (10) (JUL): 3247-60.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. New York: Cambridge University Press.

Miziolek, Andrzej W., V. Palleschi, and Israel Schechter. 2006. *Laser-induced breakdown spectroscopy (LIBS): Fundamentals and applications*. Cambridge, UK ; New York: Cambridge University Press.

NASA. Schematic of laser-induced breakdown spectroscopy. 2011. Available from http://www.nasa.gov/mission_pages/msl/multimedia/pia15103.html.

NASA Jet Propulsion Laboratory. MSL Science Corner: Chemistry & Camera (ChemCam). 2010. Available from http://msl-scicorner.jpl.nasa.gov/ Instruments/ ChemCam/.

Peters, Andrea, and Torsten Hothorn. 2012. Package 'ipred'. Available from http://cran.r-project.org/web/packages/ipred/ipred.pdf.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raudys, SJ, and AK Jain. 1991. Small sample-size effects in statistical pattern-recognition - recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (3) (MAR): 252-64.

Ukwatta, Eranga, Jagath Samarabandu, and Mike Hall. 2012. Machine vision system for automated spectroscopy. *Machine Vision and Applications* 23 (1): 111-21.

Varmuza, Kurt, and Peter Filzmoser. 2009. *Introduction to multivariate statistical analysis in chemometrics*. Boca Raton: CRC Press.

Yueh, Fang-Yu, Hongbo Zheng, Jagdish P. Singh, and Shane Burgess. 2009. Preliminary evaluation of laser-induced breakdown spectroscopy for tissue classification. *Spectrochimica Acta Part B-Atomic Spectroscopy* 64 (10) (OCT): 1059-67.