Manifold Learning for Regression of Mars Spectra

Thomas Boucher[†], **CJ Carey, Stephen Giguere, Sridhar Mahadevan**
College of Information and Computer Sciences
University of Massachusetts
Amherst, MA 01003
[†]boucher@cs.umass.edu

M. Darby Dyar
Department of Astronomy
Mount Holyoke College
South Hadley, MA 01075

Samuel Clegg, Roger Wiens
Los Alamos National Laboratory
Los Alamos, NM 87545

Abstract

Laser-induced breakdown spectroscopy (LIBS) is currently being used on-board the Mars Science Laboratory rover *Curiosity* to predict elemental abundances in dust, rocks, and soils using a partial least squares regression model developed by the ChemCam team. Accuracy of that model is constrained by the number of samples needed in the calibration, which grows exponentially with the dimensionality of the data, a phenomenon known as *the curse of dimensionality*. LIBS data are very high-dimensional and the number of ground-truth samples (i.e., standards) recorded with the ChemCam before departing for Mars was small compared to the dimensionality, so strategies to optimize prediction accuracy are needed. Manifold learning, a class of dimensionality reduction techniques that preserve geometric characteristics from the feature space in the embedded space, are frequently used to combat the curse of dimensionality. In this study, we introduce a novel extension to manifold learning designed for real-valued regression problems, manifold learning for regression (MLR), which takes into account the known chemical composition of the training data when embedding. We show the effectiveness of our proposed MLR methods and existing manifold learning algorithms for the task of preprocessesing LIBS spectra using three different datasets recorded under Mars-like conditions. MLR methods are shown to outperform traditional manifold methods for predicting real-valued chemical compositions.

Introduction

The laser-induced breakdown spectroscopy (LIBS) instrument ChemCam on the *Curiosity* rover on Mars has played a vital role in supporting geochemical investigations on a daily basis since the rover's landing in August of 2012. To date, over 200,000 laser shots have been fired at Martian materials, generating a rich suite of spectra on a wide range of dust, soils, and rocks. Figure 1 shows a photo the of the weather-beaten rover and ChemCam on Mars. To predict chemical composition from these spectra, Earth-based calibration datasets have been collected under Mars-like conditions using both the flight instrument before launch and ChemCam-like instruments subsequently (Wiens et al. 2013; Clegg et al. 2014). To date, the total number of calibration spectra acquired from ground-truth samples (i.e., standards) remains small compared to the dimensionality of the LIBS spectra (6144 channels), so work is in progress to add thousands of additional standards to the calibration process (Dyar et al. 2015).

However, the accuracy of chemical composition predictions depends not only on the training set data, but also upon the sophistication of the techniques used to interpret data. Interpretation of data from various LIBS applications has benefited from improvements in multivariate analysis techniques in recent years. Sirven et al. (2006; 2007) were among the first to apply multivariate analysis techniques to LIBS spectra of geological samples; they compared partial least squares (PLS) regression with predictions from neural network analyses. Clegg et al. (2009) used multivariate analysis techniques to analyze LIBS spectra of 18 disparate igneous and highly-metamorphosed rock samples. This work was quickly followed by Tucker et al. (2010) and studies of suites of sulfate- and carbonate-rich rocks (Clegg et al. 2007; 2006; Dyar et al. 2011) to use for calibration.

There are some common drawbacks to using multivariate methods. When using the full high-dimensional data, the regression problem will often be underdetermined, i.e., there are many more features than samples. Also, when using thousands of channels, there is often more noise in the data representation. To combat both of these problems, regularized versions of regression have been used, like lasso, ridge regression (Kalivas 2012), and sparse PLS (Chun and Keles 2010). Instead of using a preprocessor to improve the data representation, these methods attempt to solve the problem in the regression algorithm. *Dimensionality reduction* methods like principal component analysis (PCA) (Ivosev, Burton, and Bonner 2008), random projections (Varmuza, Filzmoser, and Liebmann 2010), and kernel PCA (Schölkopf, Smola, and Müller 1997) have been used as preprocessors to improve data representation and problem conditioning. *Manifold embedding* is a form of nonlinear dimensionality reduction that attempts to preserve geometric characteristics of the high-dimensional feature space when embedding the data into a low-dimensional space.

As a nonlinear preprocessing step, Boucher et al. () investigated the application of manifold embedding techniques to LIBS data of geological samples. The regression models PLS and lasso were shown to predict most elements better when the data were preprocessed with locally-linear em-

bedding (LLE), a popular manifold embedding algorithm (Roweis and Saul 2000). Many real-world datasets are assumed to lie on or near a subspace of the overall space called a *manifold*, whose dimensionality is significantly smaller than the original feature space due to physical or statistical constraints (Ma and Fu 2011). Distances on a manifold are measured locally with Euclidean distance but globally with geodesic distance, the distance over the manifold. By maintaining local neighborhood distances and discarding large global distances when constructing its embedding, LLE attempts to recover the geometry and structure underlying the data.

While manifold embedding proved effective in this preliminary work, traditional LLE is an unsupervised method, and so it is inherently naive to the known chemical composition of the training spectra. Both de Ridder et al. (2003) and Zhang (2009) proposed supervised versions of LLE, but these variants assume the response data are class-based instead of real-valued. Belkin et al. (2006) proposed a manifold regularized least squares variant of laplacian eigenmaps (LE) (Belkin and Niyogi 2003), another common manifold method, for real-valued responses, but their method does not incorporate the response data when calculating its embedding. In this work, we introduce a state-of-the-art supervised extension, manifold learning for regression (MLR), that uses both the feature data and the response data to preprocess the samples. We apply this extension to LLE and LE for preprocessing. Our results show that manifold preprocessing methods can significantly improve the accuracy of element composition prediction for geological samples.

Background and Notation

The LIBS technique as applied to geological samples with diverse and complex chemistries suffers from a fundamental limitation that inhibits use of the simple univariate calibration techniques used in many other types of atomic emission spectroscopy. All of the elements interact with one another and perturb the plasma. These chemical *matrix effects* remove some of the dependence between the intensity of any given emission line and its concentration. In particular, these effects have been shown to impose a non-linear relationship between lines and concentrations (Aguilera et al. 2009), thereby hindering the performance of linear regression methods. Matrix effects arise from the relative abundances of neutral and ionized species within the plasma, collisional interactions within the plasma, laser-to-sample coupling efficiency, and self-absorption. Minor or trace elements in any sample can also be affected by and potentially cause matrix effects on major element emission lines. Local atmospheric composition and pressure also significantly influence LIBS plasma intensity because breakdown products from atmospheric species interact with ablated surface material in the plasma. All these factors combine to create spectra representing complex interactions that result in changes among emission lines. These are likely best addressed with multivariate analysis techniques, many of which were in fact developed to deal with datasets in which the variables interact. However, these multivariate methods are often underdetermined and noisy. A dimensionality reducing preprocessing method, like manifold embedding, is useful to improve the data representation.

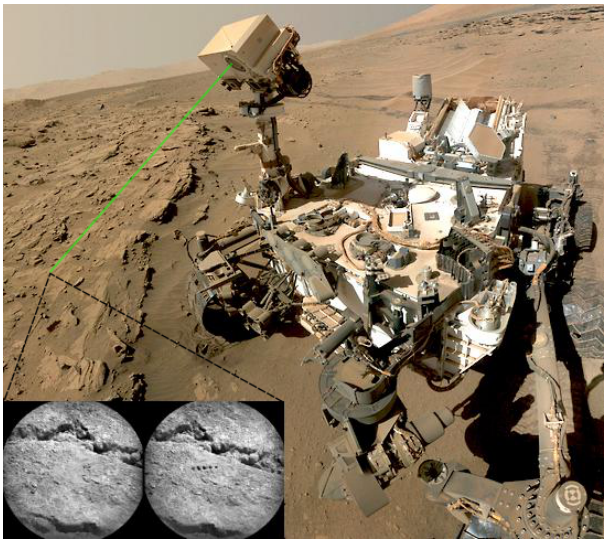


Figure 1: The *Curiosity* rover on Mars with a simulated ChemCam laser pulse. The photos on the left are of a Martian rock surface before and after laser ablation. The rock was lased 50 times in each of the five locations. Photos courtesy of NASA.

To explain the manifold preprocessing methods, the following notation is used. Capital letters (X, Y) represent matrices. Indexed lowercase letters ($x_i = X(i, \cdot)$) represent row vectors and doubly indexed lowercase letters ($x_{ij} = X(i, j)$) represent matrix entries. We use $X \in \mathbb{R}^{N \times p}$ and $Y \in \mathbb{R}^{N \times q}$ to denote the feature (or sample) matrix and response matrix, respectively, and $\tilde{X} \in \mathbb{R}^{N \times d}$ (such that $d \ll p$) to denote the embedded feature matrix. For brevity, we use the phrase *smallest eigenvectors* of a matrix to refer to the eigenvectors associated with the smallest non-zero eigenvalues.

Methods

The goal of this work is to compare and contrast two existing and two manifold preprocessing techniques for calibration model fitting using three different LIBS datasets recorded under Mars-like conditions.

Partial least squares

Partial least squares (PLS) is a linear regression model that uses a latent variable approach to model the covariance between the feature data X and the response data Y. It uses projections similar to principal component regression (PCR), but while PCR maximizes the variance within X, PLS maximizes the covariance between X and Y (Wold et al. 1984). PLS is now a standard tool in chemometrics because it is especially well-suited for problems with many collinear or nearly collinear features, few training samples, and multiple responses (Erdas et al. 2010;

Kalivas 1999). The ChemCam team currently uses a set of PLS models (one per element) to predict the chemical composition of minerals on Mars (Wiens et al. 2013). In this work, we fit PLS models with and without manifold preprocessing to evaluate the effectiveness of the preprocessing methods.

Locally linear embedding

Locally linear embedding (LLE) is an unsupervised manifold learning algorithm that calculates an embedding for each sample from feature space to a low-dimensional manifold space (Roweis and Saul 2000). LLE captures the intrinsic geometry of the feature data by reducing dimensionality while preserving local distances between samples. Other dimensionality reduction methods like PCA and multi-dimensional scaling (MDS) may fail to recover the underlying manifold, where instead of projecting distant points close together or vice versa, creating distortions in both global and local geometry (Saul and Roweis 2003). By reducing the dimensionality of the feature representation to its essential parts, LLE combats the *curse of dimensionality*, the demand for more training samples to well fit a model in higher-dimensional space. LLE is essentially a two-step algorithm. In the first step, each sample is approximated as a linear combination of its nearest neighbors. In the second step, the embedding that best preserves the locally linear approximation is calculated.

To begin the LLE algorithm, the k nearest neighbors are calculated for each sample. Next, each sample is reconstructed as a linear combination of its neighbors. To calculate the reconstruction weight matrix $W \in \mathbb{R}^{N \times N}$, the problem is formulated as the optimization problem below, minimizing the sum of squared reconstruction error,

$$\min_{W} \sum_{i=1}^{N} ||x_i - \sum_{j=1}^{k} w_{ij} x_j||^2, \tag{1}$$

such that $\sum_j w_{ij} = 1$ and $W_{ij} = 0$ if x_i and x_j are not nearest neighbors. The constraint that the weights for each sample must sum to one ensures that the embeddings are invariant to translation. By using linear reconstruction weights, the embeddings are also invariant to scaling and rotation. The 2-norm is used because the manifold space is locally Euclidean in geometry. To solve the optimization problem in equation 1 in practice, it can be posed as series of ordinary least squares problems.

After the reconstruction weights are calculated, an embedding from the original feature space to the manifold space is calculated for each sample. The goal of the embedding is to preserve the local linearity of the feature space. LLE accomplishes this by choosing the sample embedding that best reconstructs the sample from its neighbors in the embedded space using the feature space reconstruction weights. This is formulated as the minimization problem,

$$\min_{\tilde{X}} \sum_{i=1}^{N} ||\tilde{x}_i - \sum_{j=1}^{k} w_{ij} \tilde{x}_j||^2, \tag{2}$$

such that $\sum_i \tilde{x}_i = \mathbf{0}$ and $\tilde{X}^\top \tilde{X} = I$, where I is the $N \times N$ identity matrix. The constraint $\sum_i \tilde{x}_i = \mathbf{0}$ ensures that the embedding is invariant to translation, and the constraint $\tilde{X}^\top \tilde{X} = I$ ensures that the problem is well-posed. This minimization problem can be solved in closed form with the sparse eigenvector decomposition of the matrix $M = (I - W)^\top (I - W)$.

Laplacian eigenmaps

Laplacian eigenmaps (LE) is an unsupervised manifold learning algorithm that uses spectral graph techniques to calculate an embedding for each sample from feature space to a low-dimensional manifold space (Belkin and Niyogi 2003). Unlike LLE, which uses neighborhood reconstruction weights to capture the data's geometric structure, LE uses a weighted graph over the dataset to capture the manifold structure. It is essentially a two-step algorithm. In the first step, a weighted graph is constructed over local neighborhoods. In the second step, the eigenvectors (or eigenmaps) of the *graph Laplacian* are calculated for the low-dimensional embedding, as follows.

To begin the algorithm, the k nearest neighbors are calculated for each sample. A weighted adjacency matrix W is used to describe the neighbors, where $w_{ij} > 0$ if x_i and x_j are neighbors and $w_{ij} = 0$ otherwise. The matrix W defines a weighted graph over the sample neighborhoods. The neighbor weights of W are typically all set to one or are set to the radial basis function (RBF) kernel value (equation 5). An ϵ-ball neighborhood can be used instead of k-nearest neighbors; however, in practice, ϵ can be difficult to tune and the approach can lead to a disconnected graph.

After the weighted adjacency matrix is calculated, an embedding for each sample is calculated. When embedding the data, LE uses the following loss function:

$$\min_{\tilde{X}} \frac{1}{2} \sum_{i,j=1}^{N} ||\tilde{x}_i - \tilde{x}_j||^2 w_{ij}. \tag{3}$$

This can be reformulated using the graph Laplacian matrix L, where $L = D - K$ such that D is the diagonal matrix of row sums $D(i,i) = \sum_j k_{ij}$, as:

$$\min_{\tilde{X}} \frac{1}{2} \tilde{X}^T L \tilde{X} \tag{4}$$

such that $\tilde{X}^T D \tilde{X} = I$. This constraint ensures that the embedding is invariant to rescaling and translation. Equation 4 is minimized by the eigenvectors of the d smallest eigenvalues of the Laplacian L. Use of k-nearest neighbors in the first step of LE may result in a non-symmetric Laplacian matrix. To guarantee a symmetric Laplacian in practice, the *normalized graph Laplacian* $\mathcal{L} = D^{-1/2} L D^{-1/2}$ may be substituted.

Manifold learning for regression

In this section, we describe our novel contribution, manifold learning for regression (MLR). MLR methods are *supervised* variants of LLE and LE that use both the feature

data and the response data when calculating their embeddings. MLR adds an additional penalty to the loss function in the second step of LLE (equation 2) and LE (equation 4) that draws spectra of similar composition closer together in the embedding space. The LLE version (MLR-LLE) is described in Algorithm 1 and the LE version (MLR-LE) is described in Algorithm 2. Our novel contribution is identical in both algorithms, so only the derivation of the MLR-LLE algorithm is described in detail below.

MLR-LLE shares the same first step as LLE, i.e., every sample's nearest neighbors are calculated and then the sample is linearly reconstructed from those neighbors. The second step of MLR-LLE begins with the calculation of the RBF kernel matrix $K \in \mathbb{R}^{N \times N}$ defined as

$$K(i,j) = \exp\left(-\gamma||y_i - y_j||^2\right), \ \forall\, y_i, y_j \in Y, \quad (5)$$

where γ is the *bandwidth* of the kernel. K can be viewed as a similarity score between two samples with a range of $(0, 1]$, where more similar samples have a larger kernel value. The kernel matrix can be constructed pairwise for most LIBS datasets because they contain relatively few samples. There are many methods for quickly and accurately approximating the matrix K. The Nyström approximation method (Williams and Seeger 2001) and random Fourier features (Rahimi and Recht 2007) are two well-practiced methods.

After calculating K, the augmented embedding loss function is minimized:

$$\min_{\tilde{X}} \frac{1}{2}\sum_{i=1}^{N}||\tilde{x}_i - \sum_{j=1}^{k} w_{ij}\tilde{x}_j||^2 + \frac{\mu}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}||\tilde{x}_i - \tilde{x}_j||^2 k_{ij}, \quad (6)$$

where μ is the weight of the response penalty. This loss function inherits the same two constraints as equation 2. The loss function in equation 6 is composed of two parts: the first term ensures that the local geometry from the high-dimensional space is preserved in the low-dimensional embedded space, and the second term pushes samples that are similar in chemical composition closer together in the embedded space. MLR uses K to penalize samples that are similar in composition but are distant in the embedded space. Note that by setting $\mu = 0$, we get back the original LLE loss function in equation 2.

The minimization problem in equation 6 can be reduced as:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{N}||\tilde{x}_i - \sum_{j=1}^{k} w_{ij}\tilde{x}_j||^2 + \frac{\mu}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}||\tilde{x}_i - \tilde{x}_j||^2 k_{ij} \\
&= \frac{1}{2}||\tilde{X} - W\tilde{X}||_F^2 + \frac{\mu}{2}\sum_{i,j=1}^{N}||\tilde{x}_i - \tilde{x}_j||^2 k_{ij} \\
&= \frac{1}{2}tr\left(\tilde{X}^T(I - W^T)(I - W)\tilde{X}\right) + \frac{\mu}{2}tr\left(\tilde{X}^T(D - K)\tilde{X}\right) \\
&= \frac{1}{2}tr\left(\tilde{X}^T M\tilde{X}\right) + \frac{\mu}{2}tr\left(\tilde{X}^T L\tilde{X}\right),
\end{aligned} \quad (7)$$

where $||\cdot||_F$ is the Frobenius norm, $tr(\cdot)$ is the matrix trace, M is the Gram matrix $M = (I - W^T)(I - W)$, and L is the graph Laplacian. Therefore, to minimize the loss function in equation 6, we can minimize the sum of traces in equation 7. To account for the constraint $\tilde{X}^T\tilde{X} = I$, equation 7 must be reformulated using the method of Lagrange multipliers. Define $\mathcal{L}$ to be the loss function:

$$\mathcal{L}(\tilde{X}, \Lambda) = \frac{1}{2}tr\left(\tilde{X}^T M\tilde{X}\right) + \frac{\mu}{2}tr\left(\tilde{X}^T L\tilde{X}\right) - \langle\Lambda, \tilde{X}^T\tilde{X} - I\rangle, \quad (8)$$

where Λ is the diagonal matrix of Lagrange multipliers. To minimize equation 8, we calculate the roots of its partial derivatives:

$$\frac{\partial\mathcal{L}}{\partial\tilde{X}} = M\tilde{X} + \mu L\tilde{X} - \Lambda\tilde{X} = 0, \quad (9)$$

$$\frac{\partial\mathcal{L}}{\partial\Lambda} = \tilde{X}^T\tilde{X} - I = 0. \quad (10)$$

The linear constraint $\sum_i \tilde{x}_i = 0$ becomes irrelevant after calculating the partials of the Lagrangian and can be ignored. Therefore,

$$(M + \mu L)\tilde{X} = \Lambda\tilde{X} \ \text{ such that } \ \tilde{X}^T\tilde{X} = I. \quad (11)$$

The smallest eigenvectors of the matrix $\Omega = M + \mu L$ will minimize this loss function. The covariance constraint on $\tilde{X}$ is satisfied because the eigenvectors are orthogonal. The smallest eigenvalue of Ω is 0, but this is associated with the trivial eigenvector $\mathbf{1}$ and should be discarded. Note that Ω is a positive semidefinite matrix, and all of its eigenvalues will be real-values greater than or equal to zero.

When using nonlinear manifold methods to embed out-of-sample extensions, i.e., samples not in the original training set, a new procedure must be used. Although extension methods have been proposed (Bengio, Paiement, and Vincent 2003; Strange and Zwiggelaar 2011), we found the method described in Saul and Roweis (2003) works well in practice for LLE, LE, and the MLR variants. In general, the algorithm follows the same steps as LLE. First, for each testing sample (an out-of-sample extension not included in the original embedding), nearest neighbors in the *training set* $\tilde{X}_{train}$ are calculated. Next, the reconstruction vector w_i is calculated using the neighbors in the training set. Finally, the embedded test sample $\tilde{x}_i$ is calculated as $w_i\tilde{X}_{train} = \tilde{x}_i$.

Algorithm 1 MLR-LLE

Calculate the k-nearest neighbors of X.
Calculate the reconstruction weight matrix W.
Calculate the RBF kernel matrix K of Y and its Laplacian L_K.
Calculate the d *smallest* eigenvectors of $(I - W)^T(I - W) + \mu L_K$.

Data and Preprocessing

Three different LIBS datasets were used to evaluate the manifold learning methods: the 329 sample *New LANL* dataset (Clegg et al. 2014), the 61 sample *Cleanroom* dataset (Clegg et al. 2012; Wiens et al. 2013), and the 100 sample *Century* dataset (Tucker et al. 2010; Dyar et al. 2012). All three

Algorithm 2 MLR-LE

Calculate the k-nearest neighbors of X.
Calculate the weighted affinity matrix W and its normalized Laplacian $\mathcal{L}_W$.
Calculate the RBF kernel matrix K of Y and its normalized Laplacian $\mathcal{L}_K$.
Calculate the d *smallest* eigenvectors of $\mathcal{L}_W + \mu\mathcal{L}_K$.

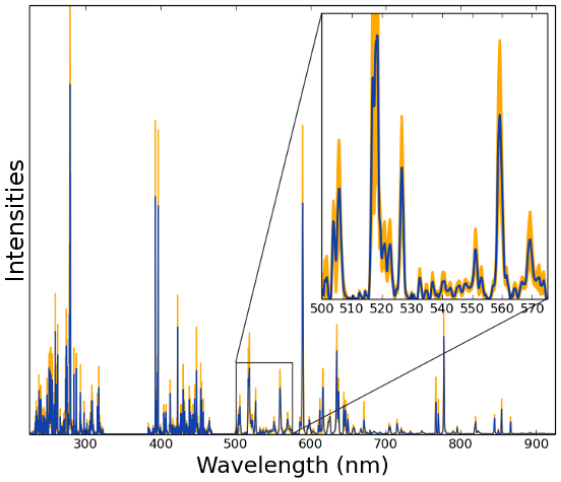


Figure 2: The blue line is the mean LIBS spectrum of 100 igneous rock powders from Tucker et al. (2010), where the horizontal axis displays the three spectrometer wavelength range and the vertical axis displays the number of counts in each channel. Two standard deviations are colored in orange.

datasets were collected in simulated Mars-like conditions at varying distances, where the samples were lased in a sealed, evacuated chamber filled with 7 torr CO_2. These datasets were recorded at Los Alamos National Laboratory in support of the Mars Science Laboratory Team and the ChemCam instrument.

All three datasets were recorded similarly. Each mineral sample was lased in 3-5 locations, and each location was lased 30-50 times. The samples were homogeneous in composition, so the mean spectrum of all the shots was used to represent each sample. The spectra were recorded at 6144 channels corresponding to a wavelength range from 240-906 nm spread between three charge-coupled devices (CCD).

Before conducting our experiments, the spectra were background removed, bad channels were masked (i.e., channels known to have poor or inconsistent instrument response), and the spectra between each of the three CCD's were normalized (Wiens et al. 2013; Tucker et al. 2010). At the start of the experiment, each spectrum was represented as a 5485 dimensional vector.

Table 1: Chemical compositions of the three datasets studied in mean weight % oxide with one standard deviation.

	New LANL		*Cleanroom*		*Century*	
	Mean	1σ	Mean	1σ	Mean	1σ
SiO_2	57.16	13.39	47.33	18.89	52.53	9.64
Al_2O_3	15.00	6.20	11.74	6.48	12.92	3.42
Fe_2O_3T	7.58	6.62	9.05	7.45	10.79	4.21
CaO	4.56	6.67	9.70	9.18	8.12	3.64
MgO	4.45	6.82	4.66	5.51	8.91	7.37
Na_2O	2.09	1.59	2.40	1.64	2.56	1.14
K_2O	2.43	1.86	1.44	1.54	1.35	1.59
TiO_2	0.93	0.80	1.06	1.21	2.04	1.40
MnO	0.12	0.17	0.13	0.11	0.17	0.06
P_2O_5	-	-	-	-	0.43	0.56

Model Tuning and Software

All model hyperparameters were tuned using a cross-validation search over the parameter space. The hyperparameters tuned included the number of PLS components over the range 1-20, the number of LLE and LE neighbors k and embedding dimensionality d over the range 1-75 (where applicable), and the MLR weight μ over the log range (0.1,1,10,100). For all elements and datasets evaluated, ten PLS components were used because the setting performed universally well. The same dimensionality and neighbor settings were used for LLE and MLR-LLE and for LE and MLR-LE. Using a grid search over the Cartesian product space, d and k were tuned. Then using the same dimension and neighbor settings, μ and γ were tuned similarly. To find a range for the kernel bandwidth γ, the pairwise (Euclidean) distances were calculated between the training samples X_{train}, and the range was set to one over the empirical quantiles at $p = 0.1, 0.25, 0.5, 0.75, 0.9$.

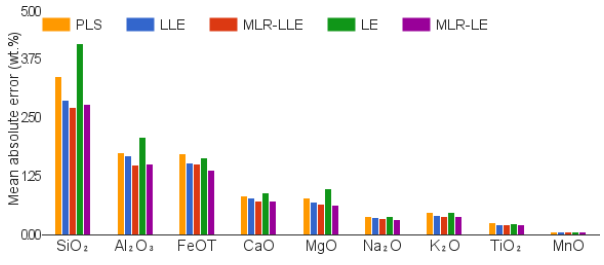
The open-source machine learning Python library Scikit-learn (Pedregosa et al. 2011) was used for the implementations of PLS, LE and LLE and to train and evaluate all models. Implementations of MLR-LLE and MLR-LE in Python are available for download from the author's website[1].

Experimental Setup and Results

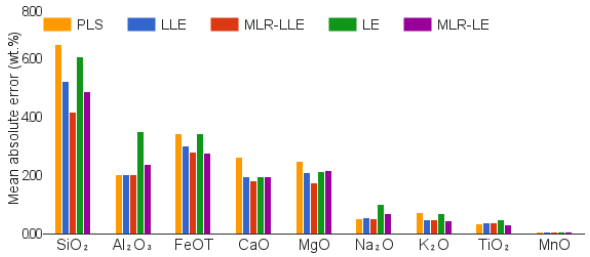
In this study, leave-one-out cross-validation was used to compare the predictive performance of a PLS regression model before and after applying manifold preprocessing to the data. For each element in each dataset, a PLS model was trained and tested: (1) on the original spectra, (2) on spectra preprocessed with LLE or LE, and (3) on spectra preprocessed with MLR-LLE or MLR-LE.Mean absolute error (MAE) was used to evaluate the performance of the models throughout. MAE was chosen for three reason: it is easily interpretable because it is in the original units (weight % oxide), it makes no parametric assumptions about the data, and it is more robust to outliers than a quadratic loss.

The results for the *New LANL* dataset, the *Cleanroom* dataset, and the *Century* dataset are listed in Figures 3a, 3b, and 3c, respectively. In general, using LLE and MLR-LLE

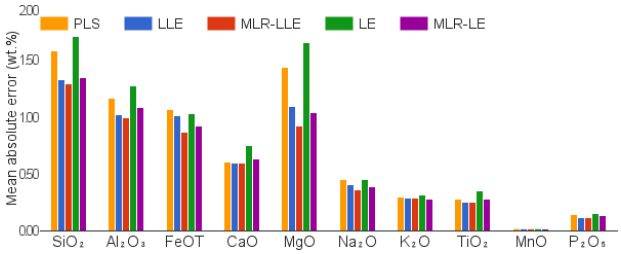
[1]https://github.com/all-umass/mlr



(a) Results of the 329 sample *New LANL* dataset. Here, LE preprocessing yields the worst performing model. We suspect this is due to the algorithm's graph-based representation of the manifold.



(b) Results of the 61 sample *Cleanroom* dataset. Manifold embedding yields no improvement for predicting Al_2O_3, because of the uniform distribution of the element across samples.



(c) Results of the 100 samples *Century* dataset. The large error in predicting MgO is due to its highly irregular distribution across samples.

Figure 3: The MAE using leave-one-out cross-validation for the three datasets. The evaluated methods were partial least-squares (PLS), locally linear embedding (LLE), manifold learning for regression LLE (MLR-LLE), laplacian eigenmaps (LE), and manifold learning for regression LE (MLR-LE).

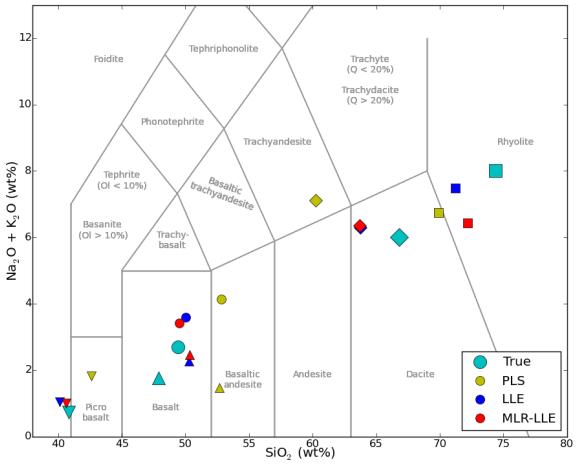


Figure 4: A TAS volcanic rock classification diagram showing a selection of samples from the *Century* dataset that were misclassified using PLS alone but properly classified when used in conjunction with manifold methods. The different symbol shapes indicate the same sample predicted using each of the three methods and the true sample value.

to preprocess the data resulted in significantly lower prediction errors than using using PLS alone. Moreover, MLR-LLE outperformed all other evaluated methods on all three datasets.

Our proposed MLR variants consistently matched or outperformed the traditional manifold methods. The only predictions worsened using LLE preprocessing were the trace elements in the *Cleanroom* dataset, however these differences were insignificant. In contrast, using unsupervised LE preprocessing was shown to frequently degrade the predictive performance of PLS. The reconstruction weight approach of LLE proved superior for real-valued prediction than the graph-based approach of LE. Our proposed supervised variant MLR-LE significantly improved the predictive performance over traditional LE, achieving results comparable to MLR-LLE.

MLR preprocessing had the greatest improvement on the *Cleanroom* dataset, the dataset with the fewest samples, and the smallest improvement (but least variant between elements) on the *New LANL* dataset, the dataset with the most samples. We suspect this is because with many more samples the unsupervised manifold learning algorithm can properly learn the embedding space without the need of the response data, while with fewer samples manifold embedding benefits from supervision. The PLS model using the original data performed best on the *New LANL* dataset also, likely due to the datasets larger size. In general, manifold preprocessing had the greatest effect on the elements in greatest abundance.

To quantify the geological significance of the improvement manifold methods provide, Figure 4 shows a total alkali-silica (TAS) volcanic rock classification diagram (Le Bas et al. 1986) and predictions from the evaluated methods. The five samples shown are selected from the *Cen-*

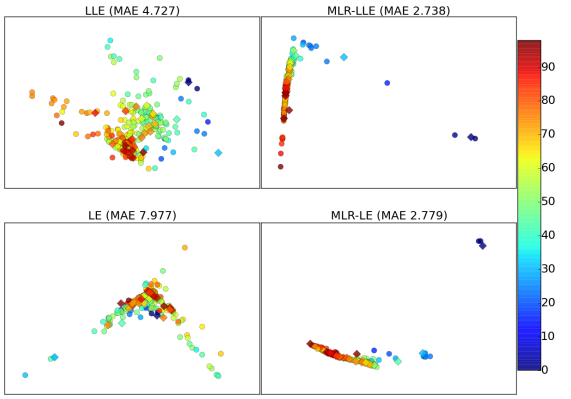


Figure 5: A comparison of 2-D embeddings from LLE, LE, MLR-LLE, and MLR-LE using the *New LANL* dataset. The circles are the training set and the diamonds are the test set. The shapes are colored according to their SiO_2 content. While the 2-D constraint hinders the performance of traditional manifold methods, our supervised extensions are not inhibited by this constraint.

tury dataset and are all misclassified when using PLS alone and correctly classified when used in conjunction with LLE and MLR-LLE. This demonstrates that the improved accuracy of manifold methods enables deeper scientific analyses, like TAS mineralogical classification.

To visually compare the four manifold algorithms, Figure 5 shows their 2-D embeddings of the *New LANL* dataset. Ten-fold cross validation was used so both the training and the (out-of-sample extension) testing examples can be seen. The figure shows the results from a single fold. Of the unsupervised methods, LLE yielded a superior embedding to LE while both methods were hindered by the 2-D embedding constraint. However, the proposed MLR methods greatly improved the performance of both manifold methods. Both training and testing sample embeddings were improved. Even when reducing the dimensionality down to just 2-D, the MLR methods were able to greatly improve the performance of the PLS models.

Conclusion and future work

In this study, we introduced a novel extension of manifold learning and used it as a preprocessing step before fitting a PLS regression model for predicting the abundance of major elements in geological samples on Mars. The proposed MLR methods consistently improved PLS model performance, and MLR-LLE preprocessed models were the best-performing models on all three datasets. These results suggest that the LIBS data inherently lie in a lower dimensional space, and that the manifold algorithms preserve the geometry of that space well. It was shown that using traditional manifold methods resulted in mixed performance, where LLE often decreased prediction error while LE often increased error. By preprocessing the original data with MLR methods into a more compact representation, the linear PLS models were able to better fit the LIBS data.

In the future, we plan to explore variants of manifold preprocessing that maximize the cross covariance of the the feature matrix X and the response matrix Y. We also plan to investigate fully embedded versions of manifold regression, in which both X and Y are coembedded in a shared manifold space and the regression is performed completely in the shared space. Methods of this type would be well suited for problems involving complex response surfaces.

Acknowledgments

We are grateful for support from NSF grants CHE-1306133 and CHE-1307179. We thank both Michael Vollinger and Michael Rhodes for contributing analyzed samples.

References

Aguilera, J.; Aragn, C.; Madurga, V.; and Manrique, J. 2009. Study of matrix effects in laser induced breakdown spectroscopy on metallic samples using plasma characterization by emission spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy* 64(10):993 – 998.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15(6):1373–1396.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* 7:2399–2434.

Bengio, Y.; Paiement, J.-F.; and Vincent, P. 2003. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *In Advances in Neural Information Processing Systems*, 177–184. MIT Press.

Boucher, T.; Carey, C.; Dyar, M.; Mahadevan, S.; Clegg, S.; and Wiens, R. Manifold preprocessing for laser-induced breakdown spectroscopy under Mars conditions. *J. Chemometrics*. In press.

Chun, H., and Keles, S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(1):3–25.

Clegg, S.; Wiens, R.; Dyar, M.; Vaniman, D.; Thompson, J.; Sklute, E.; Barefield, J.; and Maurice, S. 2006. Laser induced breakdown spectroscopy (LIBS) remote detection of sulfates on Mars science laboratory rover. *Workshop on Martian Sulfates as Recorders of Atmospheric-Fluid-Rock Interactions*.

Clegg, S.; Weins, R.; Dyar, M.; Vaniman, D.; Thompson, J.; Sklute, E.; Barefield, J.; Sallé, B.; Sirven, J.-B.; Mauchien, P.; Lacour, J.-L.; and Maurice, S. 2007. Sulfur geochemical analysis with remote laser-induced breakdown spectroscopy on the 2009 Mars science laboratory rover. *Lunar and Planetary Science Conference*. Abstract no. 1960.

Clegg, S.; Wiens, R.; Barefield, J.; Sklute, E.; and Dyar, M. 2009. Quantitative remote laser-induced breakdown spectroscopy by multivariate analysis. *Spectrochim. Acta, Part B* 64:79–88.

Clegg, S.; Lasue, J.; Forni, O.; Bender, S.; Wiens, R.; Maurice, S.; Barraclough, B.; Blaney, D.; Cousin, A.; deFlores, L.; Delapp, D.; Dyar, M.; Fabre, C.; Gasnault, O.; Lanza, N.; Morris, R.; Nelson, H.; Newsom, H.; Ollila, A.; Perez, R.; Sautter, V.; and Vaniman, D. 2012. ChemCam flight model calibration. *Lunar and Planetary Science Conference*. Abstract no. 2076.

Clegg, S.; Anderson, R.; Forni, O.; Lasue, J.; Dyar, M.; Morris, R.; and B., E. 2014. Expansion of the ChemCam calibration database. *Lunar and Planetary Science Conference*. Abstract no. 2378.

de Ridder, D.; Kouropteva, O.; Okun, O.; Pietikainen, M.; and Duin, R. 2003. Supervised locally linear embedding. In *Artificial Neural Networks and Neural Information Processing*, volume 2714. Springer Berlin Heidelberg. 333–341.

Dyar, M.; Tucker, J.; Humphries, S.; Clegg, S.; Wiens, R.; and Lane, M. 2011. Strategies for Mars remote laser-induced breakdown spectroscopy analysis of sulfur in geological samples. *Spectrochim. Acta, Part B* 66:39–56.

Dyar, M.; Carmosino, M.; Breves, E.; Ozanne, M.; Clegg, S.; and Wiens, R. 2012. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochim. Acta, Part B* 70:51–67.

Dyar, M.; Breves, E.; Lepore, K.; Boucher, T.; Bender, S.; Tokar, R.; Berlanga, G.; Clegg, S.; and Wiens, R. 2015. Calibration suite for Mars-analog laser-induced spectroscopy. *Lunar and Planetary Science Conference*. Abstract no. 1510.

Erdas, O.; Buyukbingol, E.; Alpaslan, F.; and Adejare, A. 2010. Modeling and predicting binding affinity of phencyclidine-like compounds using machine learning methods. *J. Chemometrics* 24:1–13.

Ivosev, G.; Burton, L.; and Bonner, R. 2008. Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry* 80(13):4933–4944.

Kalivas, J. 1999. Interrelationships of multivariate regression methods using eigenvector basis sets. *J. Chemometrics* 13:111–132.

Kalivas, J. 2012. Overview of two-norm (l2) and one-norm (l1) regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *J. Chemometrics* 218–230.

Le Bas, M. J.; Le Maitre, R.; Streckeisen, A.; Zanettin, B.; et al. 1986. A chemical classification of volcanic rocks based on the total alkali-silica diagram. *J.Petrology* 27(3):745–750.

Ma, Y., and Fu, Y. 2011. *Manifold Learning Theory and App.* CRC Press.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *J. Machine Learning Research* 12:2825–2830.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE* 290:2323–2326.

Saul, L., and Roweis, S. 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Machine Learning Research* 4:119–155.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1997. Kernel principal component analysis. In Gerstner, W.; Germond, A.; Hasler, M.; and Nicoud, J.-D., eds., *Artificial Neural Networks at ICANN*, volume 1327 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 583–588.

Sirven, J.-B.; Bousquet, B.; Canioni, L.; and Sarger, L. 2006. Laser-induced breakdown spectroscopy of composite samples: Comparison of advanced chemometric methods. *Anal. Chem.* 78:1462–1469.

Sirven, J.-B.; Sallé, B.; Mauchien, P.; Lacour, J.-L.; Maurice, S.; and Manhes, G. 2007. Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. *J. Anal. At. Spectrom.* 22:1471–1480.

Strange, H., and Zwiggelaar, R. 2011. A generalised solution to the out-of-sample extension problem in manifold learning. *Proc. of AAAI*.

Tucker, J.; Dyar, M.; Schaefer, M.; Clegg, S.; and Wiens, R. 2010. Optimization of laser-induced breakdown spectroscopy for rapid geochemical analysis. *Chem. Geol.* 277:137–148.

Varmuza, K.; Filzmoser, P.; and Liebmann, B. 2010. Random projection experiments with chemometric data. *J. Chemometrics* 24(3-4):209–217.

Wiens, R.; Maurice, S.; Lasue, J.; Forni, O.; Anderson, R.; Clegg, S.; Bender, S.; Blaney, D.; Barraclough, B.; Cousin, A.; Deflores, L.; Delapp, D.; Dyar, M.; Fabre, C.; Gasnault, O.; Lanza, N.; Mazoyer, J.; Melikechi, N.; Meslin, P.-Y.; Newsom, H.; Ollila, A.; Perez, R.; Tokar, R.; and Vaniman, D. 2013. Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars science laboratory rover. *Spectrochim. Acta, Part B* 82:1 – 27.

Williams, C., and Seeger, M. 2001. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*.

Wold, S.; Ruhe, A.; Wold, H.; and Dunn, III, W. 1984. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* 5(3):735–743.

Zhang, S. 2009. Enhanced supervised locally linear embedding. *Pattern Recognition Letters* 30:1208–1218.