Data Mining the Surface of Mars: Novelty Detection through Least Square Clustering of Remote Sensing Data

Shahin Movafagh Mowzoon
Department of Engineering
University of Arizona
shahinmm@hotmail.com

Abstract

Multiregression has been effectively applied in determining mineral abundances from spectral data captured by probes in planetary orbits around Mars. The analysis of such data can help answer fundamental questions about the mineral distributions on the planet and on topics such as the existence of potential liquid surface water in the past history of the planet and therefore has played an important role in Mars mission science objectives. To facilitate the automated categorization of the large volumes of data and in the interest of locating novel mineral signatures, a new clustering method is presented that combines elements of least square fitting and clustering. This method is then applied to data from the orbiting Thermal Emission Spectrometer (TES) over the Opportunity rover surroundings and the Nili Fossae region.

1. Introduction

The analysis of remote sensing spectral data may often involve signal deconvolution or unmixing from spatial and temporal mixing that occurs within an observation [1]. The orbiting Thermal Emission Spectrometer instrument captures 143 bands with a 3 km by 3 km spatial resolution utilizing a Michelson interferometer. The large number of bands allows for flexibility and compensation with respect to determining unknown potential minerals [2]. The data equation for the analysis of a single observation is therefore a matrix of trial n (typically 10-20) mineral spectra chosen at the time of analysis as the input columns by a reduced set of 73 bands (70 bands are commonly removed when analyzing surface mineralogy) as rows multiplied by a n-dimensional β column vector representing the unknown abundances with the results set equal to an observation response vector of 73 dimensions. The objective is to find what minerals are exposed on the surface from a given observation vector Y and a selected set of lab mineral spectra $\mathbf{x_1}$ to $\mathbf{x_n}$. Here β_1 to β_m represent abundances while β_0 is the constant or intercept term (**1** is a vector of 1's) and **ε** is the observation error:

$$\mathbf{y} = (\beta_0 \mathbf{1} + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \ldots + \beta_n \mathbf{x_n}) + \boldsymbol{\varepsilon} \qquad (1)$$

This is an over-determined system that can be solved using a least square fitting or linear multi-regression. Since much of the surface data exhibits linear mixing, this has been used with great success in the analysis of the data [3,4,5]. The extension of these methods into data mining techniques was the driving background through which this work was formulated.

One of the areas of interest by the TES principal scientist [6] has been the identification of novel surface regions on Mars through an intelligent pattern recognition method. This is particularly useful since the amount of data collected is very large and characterizing a section of the surface can be a tedious task. Such finds are valuable as they can help influence future science objectives.

To accomplish this task, a general method was developed that modifies any of the existing computer based clustering algorithms by converting the clustering distance function into a solution space or least square distance function based on the Grammian matrix of a given set of baseline vectors.

Previous work that has combined clustering with regression include a residual clustering method [7] and an algorithmic approach [8]. The residual clustering method was shown to be very effective in detecting influential observations in the regression while the algorithmic approach relies on a custom algorithm.

The presented method is different in that it is not limited to any single clustering algorithm and it can be applied using any number of existing standard clustering methods. It does not rely on the residuals rather it uses the solution coefficients. It also retains its flexibility on the regression side and can incorporate any modeling technique that provides a solution vector as its output.

Ultimately this approach could be applied to many situations were and grouping of data is needed based on a set of reference data. In general terms, if each data object to be grouped is represented by a vector, then a small set of objects of the same dimension are selected as reference vectors. A distance function is then created within this solution space using the solution parameter vector or its equation. Any clustering algorithm can then be applied so long as this new distance function is used as a replacement for its existing distance function or alternatively the solution vectors are calculated and then clustered.

2. A Geometrical Framework

Since each single observation **y** is actually a vector and **y** does not represent independent rows of observations, a geometrical interpretation of Least Squares is applicable here.

Given a set of input vectors and a response vector, we will derive the multiregression linear equation using a combination of vector and matrix algebra. Pure matrix and algebraic derivations and geometrical interpretations are available in the literature [7,9,10,11,12], but a vector derivation will be useful for building a framework for further analysis methodologies.

For a system with n variables and m observations, lets define the error vector **ε** as:

$$\boldsymbol{\varepsilon} = \mathbf{y} - (\beta_0\,\mathbf{1} + \beta_1\,\mathbf{x_1} + \beta_2\,\mathbf{x_2} + \ldots + \beta_n\,\mathbf{x_n}) \qquad (2)$$

Where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix},\ \mathbf{x_i} = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \cdots \\ x_{mi} \end{bmatrix},\ \mathbf{1} = \begin{bmatrix} 1_1 \\ 1_2 \\ \cdots \\ 1_m \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_m \end{bmatrix} \qquad (3)$$

From a geometrical perspective, we are minimizing the squared length of the difference between the response vector and its projection into the space defined by the regressors:

$$\|\varepsilon\|^2 = \boldsymbol{\varepsilon} \bullet \boldsymbol{\varepsilon} = (\mathbf{y} - (\beta_0\,\mathbf{1} + \beta_1\,\mathbf{x_1} + \beta_2\,\mathbf{x_2} + \ldots + \beta_n\,\mathbf{x_n})) \bullet (\mathbf{y} - (\beta_0\,\mathbf{1} + \beta_1\,\mathbf{x_1} + \beta_2\,\mathbf{x_2} + \ldots + \beta_n\,\mathbf{x_n})) \qquad (4)$$

Setting the derivatives of **ε • ε** with respect to the **β** to zero minimizes the function and results the below dot product relationships:

$$\begin{bmatrix} 1 \bullet \varepsilon \\ x_1 \bullet \varepsilon \\ \ldots\ldots\ldots \\ x_n \bullet \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix} \qquad (5)$$

The above equation indicates an orthogonal relationship will exist between the error vector and each of the regressors $\mathbf{x}_i$. The dot products can be converted to the matrix multiplication below:

$$\begin{bmatrix} 1\ldots\ldots1_m \\ x_{11}\ldots x_{m1} \\ \cdots \\ x_{1n}\ldots x_{mn} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix} \qquad (6)$$

If we now let X represent the matrix of regressors then equation (2) becomes **ε** = **y** – X **β** and substituting this **ε** into (6) we get X'(**y** – X **β**) = 0. This in turn results X'**y** = X'X **β** and solving for **β** will yield the multiregression equations:

$$\boldsymbol{\beta} = (X'X)^{-1}\, X'\mathbf{y} \qquad (7)$$

3. Inner Products and Distances

Most clustering methods rely on a distance function while applying their algorithm. A clustering distance function needs to adhere to four general rules where $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{V}_3$ below are particular objects or vectors and **V**i represents any i^th vector in the dataset in the relations below:

Distance [$\mathbf{V}_1$,$\mathbf{V}_1$] = 0 (8)
Distance [$\mathbf{V}_1$,**V**i] ≥ 0
Distance [$\mathbf{V}_1$,$\mathbf{V}_2$] = Distance [$\mathbf{V}_2$,$\mathbf{V}_1$]
Distance [$\mathbf{V}_1$,$\mathbf{V}_3$] ≤ Distance [$\mathbf{V}_1$,$\mathbf{V}_2$]+Distance [$\mathbf{V}_2$,$\mathbf{V}_3$]

Most algorithms can easily incorporate a general distance function without any major change to their code.

Inner products can be used in measuring a distance.

An inner product is defined as a function on pairs of vectors such that the properties of symmetry ($x'y = y'x$), positivity ($x'x \geq 0$ and $x'x=0$ if and only if $x=0$) and bilinearity (for all real numbers a and b, $(ax+by)'z = ax'z+by'z$ for all vectors x, y and z) are preserved [13]. A new inner product can be defined such that given a regressor vector from the X matrix ($\mathbf{x}_k$) and an observation $\mathbf{y}$ it should return the solution coefficient for that vector β_k. If this was an orthogonal system the Euclidean inner product would suffice. However, an oblique or "best fit" inner product is needed in our solution.

The Inner Product for β_k can be obtained as follows:

Where $\mathbf{x}_k$ is the k^th predictor or regressor vector and $\hat{u}_k$ is the unit vector for the k^th dimension (example: k=3 would give $\hat{\mathbf{u}}_3 = <0,0,1,0,0,\ldots>$) and X is the matrix of all predictors, then the k^th regressor can be obtained as:

$$\mathbf{x}_k = \hat{\mathbf{u}}_k'X'$$

It follows that:

$$\mathbf{x}_k'X = \hat{\mathbf{u}}_k'X'X$$

Swapping the sides of the equation and multiplying by $(X'X)^{-1}$ yields:

$$\hat{\mathbf{u}}_k'(X'X)(X'X)^{-1} = \mathbf{x}_k'X(X'X)^{-1}$$

$$\hat{\mathbf{u}}_k' = \mathbf{x}_k'X(X'X)^{-1}$$

Since from (7) we have:

$$\boldsymbol{\beta} = (X'X)^{-1}X'\mathbf{y}$$

We then get:

$$\beta_k = \hat{\mathbf{u}}_k'\,\boldsymbol{\beta} = \mathbf{x}_k'X(X'X)^{-1}(X'X)^{-1}X'\mathbf{y} \qquad (9)$$

This creates an inner product in the solution space and if $\mathbf{x}_k$ and $\mathbf{y}$ are replaced with two observations $\mathbf{y}_1$ and $\mathbf{y}_2$ it is equivalent to $\boldsymbol{\beta}_1'\boldsymbol{\beta}_2$.

Note that since $[\mathbf{y}_1'\ (X(X'X)^{-1})] = \boldsymbol{\beta}_1'$, where $\boldsymbol{\beta}_1$ is the beta vector for observation vector $\mathbf{y}_1$ and $[((X'X)^{-1}X')\ \mathbf{y}_2] = \boldsymbol{\beta}_2$, where $\boldsymbol{\beta}_2$ is the beta vector for observation vector $\mathbf{y}_2$ then substituting into equation $\boldsymbol{\beta}_1'\boldsymbol{\beta}_2$ above also yields:

$$[\mathbf{y}_1'(X(X'X)^{-1})][((X'X)^{-1}X')\mathbf{y}_2] \qquad (10)$$

This is a similarity measure and we need a distance measure. To accomplish this we simply change the function to:

$$[(\mathbf{y}_{1-}\,\mathbf{y}_2)'(X(X'X)^{-1})][((X'X)^{-1}X')(\mathbf{y}_{1-}\,\mathbf{y}_2)] \qquad (11)$$

This can be also thought of as $(\mathbf{y}_{1-}\,\mathbf{y}_2)'Q\,(\mathbf{y}_{1-}\,\mathbf{y}_2)$ with Q as a matrix formed from all the X terms. Although this approach would also result the same distance calculation, it is more computationally resource intensive than (10). But all this demonstrates the fact that a least square distance function can be defined based purely on the X matrix and two observations $\mathbf{y}_1$ and $\mathbf{y}_2$. If the solution vectors are already computed it is also equivalent to $(\boldsymbol{\beta}_{1-}\,\boldsymbol{\beta}_2)'(\boldsymbol{\beta}_{1-}\,\boldsymbol{\beta}_2)$ or the clustering of the $\boldsymbol{\beta}$ solution vectors. If the solution vectors are obtained first for each y then the complexity order of the solution is simply equivalent to that of the basic clustering technique used.

4. Least-Squared Clustering

Clustering is often used as an exploratory technique where variables are considered inputs and the data is usually grouped by distance or dissimilarity functions using various algorithms and methods. It is the main method for unsupervised learning. Its great strength lies in its ability to group data into a set of groups with no requirements for training or an output variable.

The hierarchical method of "agglomerative clustering" and the partitioning method k-means are the most common clustering techniques [9] and are found in most software packages.

Additional types of clustering would include density-based methods, grid-based methods, model-based methods, high-dimensional data clustering methods and constraint-based methods [14].

In the k-means partitioning method for example, given k number of partitions, it randomly selects k of the objects, each one will then represent a cluster mean or center, then each of the remaining objects is assigned to the cluster that has the most similarity to the object based on the distance between each given object and the cluster mean. It then recalculates the mean for each cluster and iterates again relocating the data based on the new mean values, it continues iterations until the mean values for the partitions stop changing [15]. Different initial partitions may result different results.

Hierarchical methods can be either agglomerative or divisive. The agglomerative method is a bottom-up method as it starts with each single data point in its own cluster and joins the closest points or groups iteratively based on the distance functions until a single cluster is formed. There are variations on how the algorithm decides to merge the clusters, these "linkage" variations determine the distances between the clusters and in doing so they may use the nearest members of the groups (single linkage), farthest members of the groups (complete linkage), average distances (average linkage) or other criterion such as Ward's measures of variance to determine the next grouping [9]. The divisive method is a top-down approach where it starts from all the data in a single cluster and continues dividing until each data point is in its own cluster.

Clustering methods suffer from the curse of dimensionality where too many dimensions can often make the data sparse and the distance measures less meaningful.

The number of clusters is commonly one of the inputs in most algorithms and it is often difficult to determine how many clusters are the right number of clusters to generate without some problem domain knowledge.

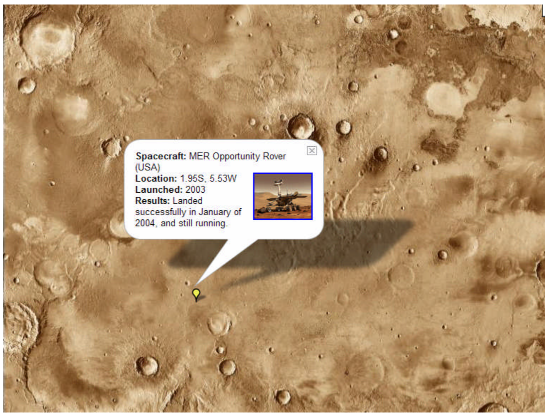
Due to the fact that clustering lacks guidance from users or classifiers it may not generate highly desirable clusters [14]. There is often a need for clustering but in comparison to a set of known objects. Where clustering is currently applied in settings where there is no set of predicted **y** responses being predicted by a number of predictor X variables, the approach here creates a perspective wherein a set of predictor X variables create a solution space where clustering distances are measured. This approach combines notions of **x** and **y** variables in the model making them interchangeable.

Since the data being clustered is being represented in the solution space with reduced dimensions now equal to the number of X predictor variables, an additional benefit is that distances become more meaningful and the clustering results become more useful and indicative of valid groupings. Also any number of existing clustering algorithms can be used depending on the problem without having to change their algorithms. Finally since its foundation is multiregression, variations of the solution could be developed to adapt to different needs.

This method allows the user to apply domain knowledge and guide the clustering based on a known set of criteria.

5. Implementation and Simulation

The Thermal Emission Spectrometer data is stored in a well documented format and available for public use [16]. An extraction tool is also provided that has a runtime for multiple platforms. Using this tool the data in a 1 degree by 2 degree region around the Opportunity lander area was extracted. The landing site is located at 1.95 S latitude and 354.47 east longitude as shown in Figure 1.



Spacecraft: MER Opportunity Rover (USA)
Location: 1.95S, 5.53W
Launched: 2003
Results: Landed successfully in January of 2004, and still running.

Fig. 1. The opportunity rover landing site. (Courtesy of NASA / JPL / JSFC / Arizona State University and Google Mars)

Infrared spectra for various minerals was also downloaded into a file [16]. A band sampling of the lab files was performed to reduce the bands from the lab spectra making it match the probe data in its dimensionality. A subset of the lab minerals were selected based on current science results [17] as regressors or mineral "end-members".

The Mathematica software package was used for all the data manipulations as well as the analysis. This proved very useful since Mathematica has a fully documented clustering capability and all the functions and logic are exposed and programmable.

Agglomerative clustering methods were used with a change in the distance function. Both Single Linkage and Complete Linkage options were used in the analysis and various number of clusters were tried out.

A spectra simulator was first created to test the system against known mixtures. The simulator generated both random or static abundances as needed within configurable ranges.

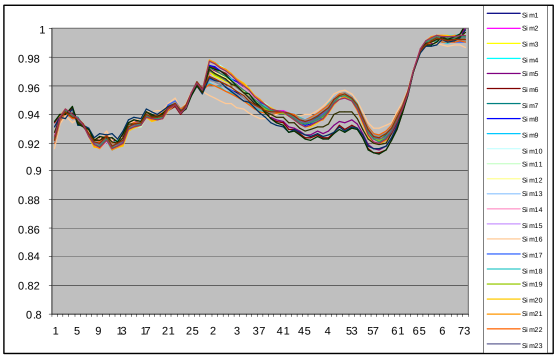


Fig. 2. A simulated set of spectra with exactly the same "end member" abundances but varying atmospheric components.

Figure 2 shows simulated output spectra with static mineral abundances but varying atmospheric components. Each line in the graph is a different signal or observation vector but the surface mixture is consistently composed of 14% Acidalia-type surface, 58% Andesine, 16% Hematite, 11% Sulfate and 1-2% random noise. The atmosphere components (dust and cloud) are based on documented spectral shapes and vary between 0-20% dust and 0-10% cloud between observations in this simulation.

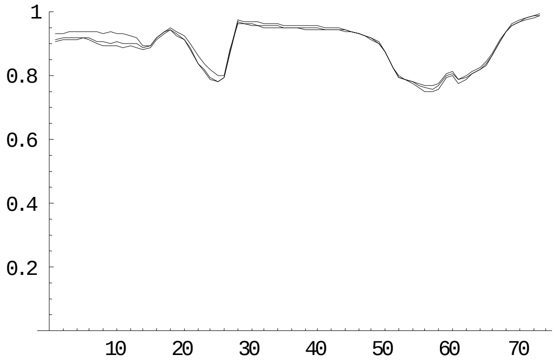


Fig. 3. A cluster of simulated spectra containing quartz without atmosphere removal.

Additional random spectra were added to the first data with random ranges of occurrence and abundance percentages. As shown in Figure 3, the clustering algorithm successfully separated spectra that were very dominated by quartz and gypsum.

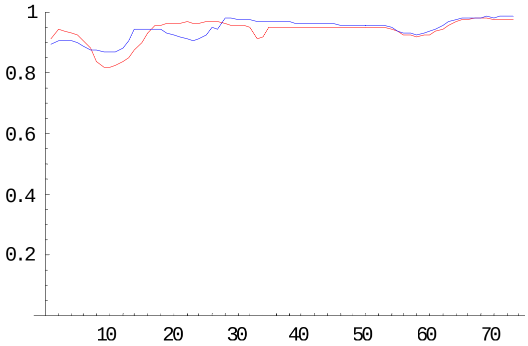


Fig. 4. Average cluster values for two distinct groups from simulation with the removal of atmospheric components.

The Least Square inner product described earlier was then used to remove the atmosphere and dust components and the algorithm was run again resulting in better separation, distinguishing between end-member minerals Hematite and Calcite. The resulting cluster comparison is shown in Figure 4. Both observations also had a Sulfate component. This is a good example of potential novelty detection since each cluster was very small with only 4 data points present in the simulated data out of all the entire dataset.

6. Preliminary Results from Spectrometer Data

The TES data includes a quality flag that indicates whether a glitch may exist in the data. Such glitches could for example occur from simultaneous antenna transmission during an observation. These data were excluded from the analysis. Interestingly some such outlier data still existed within the dataset without the flag being set and the clustering algorithm grouped these separately into additional clusters as shown in Figures 5 and 6. This is significant because it is a good example of the algorithm's ability to identify outlier or novelty type signals. In this case signals that were not in the X regressor matrix.

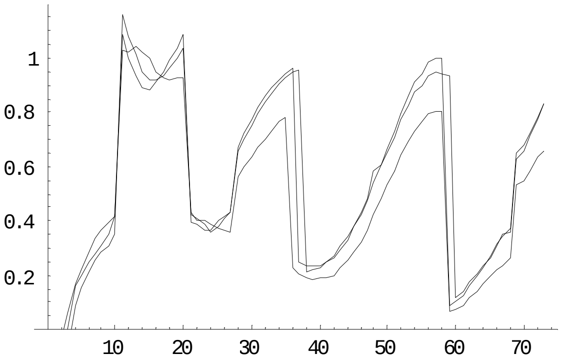


Fig. 5. Several spectra signals successfully grouped into one of the clusters by the algorithm. Note this does not look like spectra rather some instrument glitch was captured by the device.

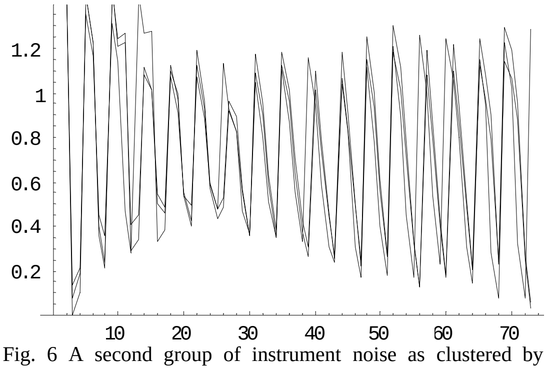


Fig. 6 A second group of instrument noise as clustered by the algorithm.

Valid observations with similar signatures were also usually nicely grouped together as shown in Figure 7. In one or two rare instances an outlier signal was also grouped with otherwise nicely grouped data. This is because it is possible to have matching abundances as well as a large noise component mixed into the signal. If the main purpose of the analysis is purely removing all glitches, a residual clustering method could also be used [7] as a separate pass.

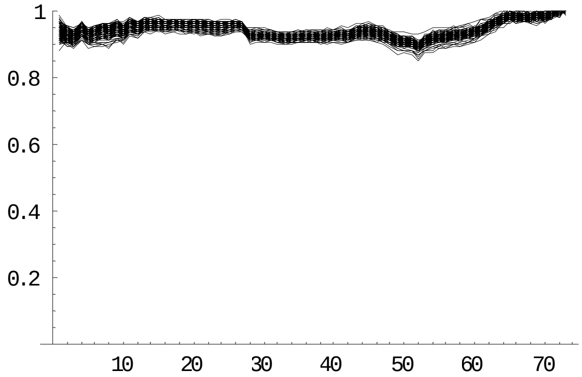


Fig. 7. One of the groupings in the opportunity lander site with many spectra grouped together.

The Nili Fossae region on Mars is a good candidate source for novelty detection [6]. Subspace clustering was performed in this region—after running the clustering algorithm, one of the resulting clustered groups was then clustered by itself breaking it up to sub-clusters. In this second pass, several small sets of clusters were separated from the great majority of the extracted data (< 1%) as shown in Figures 8 and 9 where blue is the observed and red is the modeled spectra. This spectra could not be fitted well with the standard end-members and spectral shapes commonly used suggesting further investigation is merited. It was verified with the TES research staff that Nili Fossae can have minerals that may be unusual and the spaceship clock number (unique identifier) of this data is being provided to the team for further analysis.

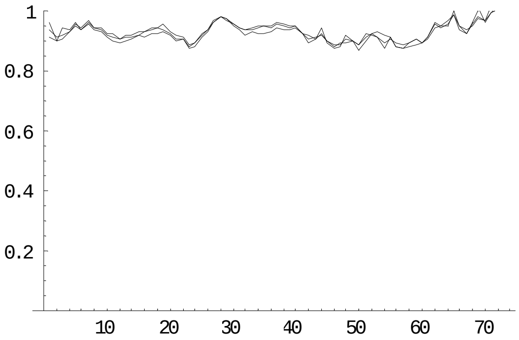


Fig. 8. Nili Fossae spectra.

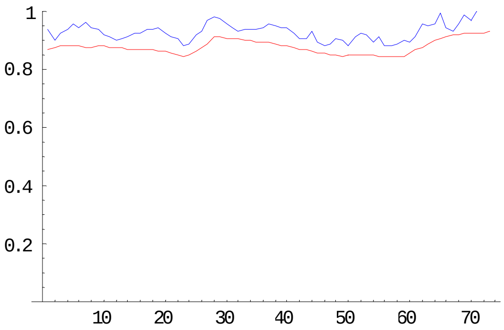


Fig. 9. Nili Fossae blue is the observed vector and red is the modeled spectra.

The subspace clustering approach was very useful. This is in part because multicolinearity can sometimes exist between the regressor spectra in the X matrix and so it is sometimes useful to cluster the subspace by removing end-members or regressors that do not exist in the subspace. The number of clusters were specified generally as twice the number of end-members in these experimental runs.

7. Conclusion

The method shown here has proven effective in the grouping of spectral data. It demonstrates that doing a best fit to reduce data dimensionality combined with clustering can be useful with certain datasets. It shows that it is possible to cluster data using a few guiding signals while capturing outlier signals into distinct groups.

Another strength of this approach is clearly in its versatility. For example, ridge regression was used with the initial passes where multicolinearity due to larger number of end-members was a factor [7]. In fact various optimization methods could be used simply by adding a clustering of the solution vectors ($\boldsymbol{\beta}_i$) as one of the last steps in the process.

Although dimension reduction and high dimension techniques have sometimes been incorporated with clustering [18], using this method, the user has control over the data reduction through the selection of the end-members.

Finally it is possible to use a set of observed **y** spectra and place them into the X end-member matrix instead of the lab spectra, this approach it worthy of experimentation. Such an approach should be compared with the FCM (Fuzzy C-Means) clustering as a possible alternative for grouping unknown surface features [19].

8. Acknowledgment

This work was made possible through support and encouragement from the ASU TES science team. Special thanks goes to Joshua L. Bandfield, Philip R. Christensen, Tom Taylor and Mary Poulton for their guidance and encouragement in this project.

9. References

[1] Twomey, S., Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurement, Dover Publications Inc., Mineola N.Y., 1977.

[2] Christensen, Philip R., Lecture notes, 2004.

[3] Ramsey, Michael S., and Christensen, Philip R, "Mineral abundance determination: Quantitative deconvolution of thermal emission spectra", Journal of Geophysical Research, American Geophysical Union, January 10 1998, pp. 577-595.

[4] Christensen et. al., "Results from the Mars Global Surveyor Thermal Emission Spectrometer", Science, March 13, 1998, pp.1692-1698.

[5] Christensen et. al., "Evidence for magmatic evolution and diversity on Mars from infrared observations", Nature, July 28, 2005, pp.504-509.

[6] Christensen, Philip R., Personal Communication, 2005.

[7] Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G. Geoffrey, Introduction to Linear Regression Analysis, John Wiley & Sons Inc., New York, 2001.

[8] Zhang, B., "Regression Clustering", ICDM 2003, 2003, pp.451-458.

[9] Johnson, Richard A., and Wichern, Dean W., Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey, 2002.

[10] Weisstein Eric W., CRC Concise Encyclopedia of Mathematics, Chapman & Hall/CRC, NewYork,1999.

[11] Ravishanker, Nalini, and Dey, Dipak K., A First Course In Linear Model Theory, Chapman & Hall/CRC, NewYork, 2002.

[12] Montgomery, Dougles C., Runger, George C, and Hubele, Norma Faris, Engineering Statistics, John Wiley & Sons Inc., New York, 2004.

[13] Ramsay, J.O., and Silverman, B.W., Functional Data Analysis, Springer, New York, 2005.

[14] Han, Jiawei, and Kamber, Micheline, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, New York, 2006.

[15] Witten, Ian H., and Eibe, Frank, Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, New York, 2005.

[16] TES public web site: http://tes.asu.edu

[17] Bandfield, Joshua L., Hamilton, Victoria E. and Christensen, Philip R., "A Global View of Martian Surface Compositions from MGS-TES", Science, March 3, 2000, pp.1626-1629.

[18] Han, Eui-Hong, Karipis, George, Kumar, Vipin, and Mobasher, Bamshad, "Hypergraph Based Clustering in High-Dimensional Data Sets: A Summary of Results", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 1997.

[19] Keshava, Nirmal, and Mustard, John F., "Spectral Unmixing", IEEE Signal Processing Magazine, January 2002, pp. 44-56.