Automatic Whole-Spectrum Matching

CJ Carey†, Thomas Boucher, Stephen Giguere, Sridhar Mahadevan
College of Information and Computer Science
University of Massachusetts
Amherst, MA 01003
†ccarey@cs.umass.edu

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

Abstract
In many space science applications, there is a need for the ability to match spectra of unknown objects against a reference spectral database. Whole-spectrum matching has been shown to perform this task effectively, but its performance is strongly linked to end-user choices, as well as specific characteristics of the dataset involved. Conventionally, the task of tuning the matching algorithm falls to the user, a time-consuming and often subjective undertaking. In this paper, we generalize these tunable choices within a preprocessing and matching framework, making this tuning process amenable to automated optimization methods. Our novel method dramatically reduces the burden on the user and results in improved matching accuracy.

Introduction
Interpretation of remote-sensed and prospective in-situ spectroscopic data from surfaces of terrestrial bodies (planets, moons, and asteroids) relies on a combination of modeling and empirical fitting routines to produce estimates of mineralogy, rock types, and space weathering. There are many variables that affect the spectra, including the number of different minerals present, their grain size distributions, porosities, the wavelength of energy used as a probe, and the complex interactions of surface materials with incoming and reflected radiation.

The planetary science community typical models such spectra by focusing on small portions of the spectra containing diagnostic peaks that are characteristic of specific minerals. For example, in reflectance spectroscopy, those peaks can then be modeled using linear combinations of end-members for coarser size fractions (Singer and McCord 1979) and non-linear mixtures for microscopic and intimate mixtures (Nash and Conel 1974; Hapke 1981; Singer 1981; Clark and others 1999; Lawrence and Lucey 2007; Shkuratov et al. 1999) or deconvolution of component peaks (Mustard and Pieters 1987; Sunshine and Pieters 1993). Current practice in planetary science relies heavily on these physical mixing models for unmixing spectral signatures of mixtures of minerals. However, these methods tend to be computationally cumbersome, sensitive to noise and other factors, and poorly constrained with confidence limits. Tools that were largely developed more than two decades ago remain the common practice in studies of planetary data. This discipline stands to benefit from new approaches to spectral deconvolution that build upon the availability of improvements in computational power and techniques from the machine learning community.

In this study, we describe a whole-spectrum matching technique that enables large-scale analyses of spectra and avoids costly and error-prone peak-fitting and dimension reduction steps. It exploits all available data to compute similarity scores between pairs of spectra in a framework that is general enough to be applied to a range of different types of spectroscopy. Here, we demonstrate the application of whole spectrum matching to Raman spectroscopy of minerals. These results will be immediately useful to the science teams working on upcoming Mars missions, including Exo-Mars (Lopez-Reyes et al. 2013) and Mars 2020 (Gasnault et al. 2015; Beegle et al. 2014).

The approach described here exhibits state of the art performance on a problem important to the analysis of instrument data from space science missions. Moreover, our techniques are applicable to a broader set of problems in the greater artificial intelligence community. For example, the reformulation of combinatorial optimization problems in terms of bounded optimization is a technique that many AI practitioners may find useful. These insights are applicable to the analysis of many naturally-occurring datasets of interest, which have traditionally relied upon human expert intervention to make optimization tractable (Norris and Ritchie 2008).

Motivation
The task of spectrum matching is commonly presented as a search problem, in which a set of query spectra are compared to a target set, typically comprised of a comprehensive spectral library. If the target set has sufficient coverage of the materials of interest, these matching techniques can be very effective at spectrum identification and analysis. Many methods for spectrum matching have been explored, making use of techniques such as peak analysis, principal components analysis, and signal-processing (Ishikawa and Gulick 2013). These methods are complicated by naturally-occurring variations between spectra in the same class, caused by differences in sample crystal orientation, laser polarization, focus, and other instrumental parameters. These differences
typically manifest as variations in peak intensity, as illustrated in Figure 1, the effects of which are typically mitigated using a sequence of preprocessing steps.

Carey et al. (2015) demonstrated that full-spectrum matching algorithms exhibit excellent performance in classification tasks for Raman spectroscopy without expensive dimensionality reduction or model training. These techniques offer great promise for adaptation to planetary and terrestrial applications, and hint at the potential for expanding the application of full-spectrum algorithms to a wide variety of larger spectral datasets and applications for other kinds of materials, as well as mineral mixtures.

That study recommended that optimal mineral spectrum matching performance could be achieved using a weighted-neighbors classifier based on a cosine similarity metric, with spectra that have been preprocessed via square-root squashing, then maximum intensity normalization, then sigmoid squashing. The primary difficulties with this spectrum matching pipeline are the choice of preprocessing steps, their order of application, and the choice of distance metric. These choices are critically important to the performance of the spectral matching algorithm, but the only way to optimize them is a time-consuming guess-and-check process. For example, the recommended sequence of preprocessing steps in Carey et al. (2015) was determined empirically, and is not guaranteed to be optimal for all spectrum matching tasks. This shortcoming presented challenges for users, as the results for Raman matching were not easily generalizable to other types of spectroscopy.

Accordingly, we now address this issue by introducing a mathematical generalization of these preprocessing and distance metric choices. Our generalized preprocessing and matching framework can be parameterized using three real-valued numbers, each constrained to a finite range. This framework dramatically simplifies the spectrum matching pipeline and allows for automated parameter tuning using existing methods from the field of stochastic optimization.

**Methods**

Our generalized framework includes both preprocessing and matching steps. In this section, we discuss preprocessing and distance metric choices separately, and then show how they can be combined.

**Generalized Preprocessing**

The conventional preprocessing pipeline consists of a series of transformation functions, each applied to the result of the previous step, starting with the raw intensity values of a spectrum, $y_{raw}$. If a particular choice of preprocessing consists of the step ordering $\{f_1, f_2, f_3\}$, the resulting processed intensities are computed:

$$y_{pp} = f_3 (f_2 (f_1 (y_{raw}))) .$$

We can generalize this idea by introducing a parameterized function $p(y)$ that can express all reasonable series of traditional preprocessing steps.

Without loss of generality, we restrict raw and preprocessed intensity values to lie between 0 and 1. This is achieved by first normalizing each spectrum by its maximum intensity. With this restriction in place, we can tie the ends of the preprocessing function, such that $p(0) = 0$ and $p(1) = 1$. These constraints eliminate many redundant preprocessing functions, reducing the search space significantly.

Our search space is further reduced by enforcing a constraint introduced in Carey et al. (2015), which states that each individual preprocessing step must be a monotonic transformation. This guarantees that the relative ordering of intensity values remains fixed, preserving the inherent shape of the spectrum. In our generalized preprocessing framework, this constraint requires that the first derivative of the preprocessing function be non-negative for all valid intensity values.

We also restrict our preprocessing functions to be concave over the range of valid intensity values. This constraint diminishes the influence of uninformative variations in peak amplitude, which is appropriate for spectrum matching applications. A concave preprocessing function will increase the relevance of smaller peaks, which Carey et al. (2015) demonstrated to be a useful property.

Collecting these observations about spectrum preprocessing, we define the following constraints:

$$p(0) = 0, \quad p(1) = 1, \quad p'(y) \geq 0, \quad p''(y) \leq 0,$$

$$0 \leq y \leq 1.$$ 

This set of constraints can be satisfied reasonably with many possible functions. In this paper, we present and explore two such functions.

1. **Polynomial Preprocessing**

The first is a cubic polynomial, which in the general case can be written as

$$p_{poly}(y) = ay^3 + by^2 + cy + d.$$ 

After applying the endpoint constraints, this can be simplified to a function with two free parameters:

$$p_{poly}^{(a,b)}(y) = ay^3 + by^2 + (1-a-b)y.$$ 

The monotonicity and concavity constraints bound the feasible values for the parameters $a$ and $b$, resulting in the ranges $-\frac{1}{2} \leq a \leq 1$ and $-2a - 1 \leq b \leq \min(0, -3a)$. These ranges define a closed triangle, in which any $(a, b)$ point defines a unique and valid preprocessing function which fulfills each of the constraints listed above.

2. **Bezier Preprocessing**

The second preprocessing function derives from the requirement of a smooth, monotonic function tied at two endpoints. This can be elegantly represented with a Bezier curve with control points at $(0, 0)$, $(1, 1)$, and a third point $(a, b)$ inside the unit square. The traditional
The formulation of a quadratic Bezier curve with these control points is
\[
\begin{align*}
  p(t) &= 2at(1-t) + t^2 \\
  y(t) &= 2bt(1-t) + t^2 \\
  t &\in [0,1].
\end{align*}
\]

Using this system of equations, we can eliminate the dependence on \( t \) and rewrite this as a preprocessing function:
\[
\begin{align*}
  p_{\text{bezier}}(y) &= 2(b-a)\left(\sqrt{a^2 - 2ay + y - a}\right) + y(4ab - 2a - 2b + 1) \\
  &\quad (2a-1)^2.
\end{align*}
\]

Applying the monotonicity and concavity constraints, we obtain parameter ranges \( 0 \leq a \leq b \leq 1 \). As with the polynomial function, this Bezier preprocessing function satisfies all constraints at all valid parameter settings.

The result of this generalization is that the task of optimizing the preprocessing procedure, which previously required a search over all permutations of discrete preprocessing steps, is now a simple search over a bounded two-dimensional parameter space. Figure 2 illustrates both types of generalized preprocessing functions, showing their effects on the normalized intensity values from a hypothetical input spectrum. Figure 3 demonstrates the expressivity of the proposed generalized preprocessing methods by comparing the empirical method of Carey et al. (2015) with the polynomial preprocessing function described above.

**Generalized Spectrum Similarity**

Carey et al. (2015) introduced the idea of “MatchScore” functions to more accurately capture the semantics of whole-

---

**Figure 1**: Raman spectra acquired using a 532 nm laser of trolleite samples 30565 (blue) and 32267 (green) from the RRUFF database. Variations in peak intensities occur even in samples of the same mineral species and laser energy.

**Figure 2**: Both types of generalized preprocessing functions are shown. The polynomial preprocessing function (shown as a line) and the Bezier preprocessing function (shown as diamonds) are compared.

**Figure 3**: Comparison of preprocessing procedure used in Carey et al. (shown as diamonds) vs. the proposed polynomial preprocessing function (shown as a line). The \( x \) axis is the normalized input intensity for each of five RRUFF spectra, and the \( y \) axis is their corresponding preprocessed intensity. This figure shows that the polynomial preprocessing function is able to accurately reproduce the empirically determined best sequence of discrete preprocessing steps, using the parameters \( a = 0.49 \) and \( b = -1.95 \).
spectrum matching. These are simple functions that compare two scalar intensity values at one particular wavelength from a pair of spectra. We can then define a simple similarity metric based on this class of functions, for a pair of spectra $A$, $B$:

$$\text{sim}(A, B) = \sum_{i=0}^{n} \text{MatchScore}(A_i, B_i),$$

where $n$ is the number of channels in each spectrum, and $A_i$ and $B_i$ are the intensity values associated with the $i$th channels of $A$ and $B$, respectively.

This formulation encompasses each of the popular spectrum-matching metrics. Cosine similarity can be recovered with

$$\text{MatchScore}(y_A, y_B) = y_A y_B,$$

assuming without loss of generality that $\|A\| = \|B\| = 1$. Squared Euclidean ($l^2$) similarity uses

$$\text{MatchScore}(y_A, y_B) = -(y_A - y_B)^2,$$

and Manhattan ($l^1$) similarity uses

$$\text{MatchScore}(y_A, y_B) = -|y_A - y_B|.$$

The primary advantage of the MatchScore framework for similarity metrics is the ease with which it can be generalized. For example, we can define a parameterized MatchScore based on a linear combination of cosine and $l^1$ metrics as follows:

$$\text{MatchScore}^{(w)}(y_A, y_B) = (1 - w) y_A y_B - w |y_A - y_B|.$$

The single parameter $w \in [0, 1]$ controls the extent to which this new similarity metric emulates cosine similarity or $l^1$ similarity. The cosine component is useful for giving high similarity to pairs of intensities with large values, while the $l^1$ component penalizes pairs of intensities with differing values. The effect of this parameter is illustrated in Figure 4.

**Algorithm 1** Generalized Whole-Spectrum Matching

**Input:** query spectra $Q$, target spectra $T$, parameters $w, a, b$

**Output:** $|Q| \times |T|$ matrix of pairwise similarities, $S$

1. For each spectrum in $Q$ and $T$ individually, normalize intensities to the $[0, 1]$ range.
2. Preprocess all spectral intensities using the preprocessing function $p^{(a,b)}$.
3. For each spectrum in $Q$ and $T$ individually, normalize intensities such that the sum of squared intensities is one.
4. For each pair of spectra $(Q_i, T_j)$, compute $S_{ij} \leftarrow \sum \text{MatchScore}^{(w)}(y_q, y_t)$.

Optimizing the Spectrum Matching Pipeline

Combining these ideas, we present a novel spectrum matching pipeline in Algorithm 1, which depends on three bounded, real-valued parameters. This pipeline produces a pairwise similarity matrix, which may then be used in a classifier of the user’s choice, such as the “weighted-neighbors” classifier used in Carey et al. (2015).

The accuracy score for this classifier forms a natural loss
Experimental Results

To demonstrate the efficacy of this approach, we consider the problem of identifying minerals based on their Raman spectra. This is the same identification task studied by Carey et al. (2015); it is directly applicable to the challenges presented by mineral analysis in remote locations, including Mars and the Moon.

We evaluate the accuracy of mineral species identification using cross validation with semi-randomized splits of Raman spectra from the RRUFF mineral database (Downs 2006) into query and reference sets. To ensure that each query spectrum has a true match in the reference set, the reference set for each trial is constructed by selecting three spectra per mineral species at random. The remaining spectra are assigned to the trial’s query set. This means that species with three or fewer total spectra are not present in the query set for any trial and always appear in the reference set. Under this testing regime, each trial selects 2643 reference spectra and 1307 query spectra. Each reference set covers all 1,215 mineral species in the RRUFF database, while each query set contains only the 349 mineral species with more than three samples.

Accuracy results for a weighted-neighbors classifier are computed by taking the ratio of the number of correct matches to the total number of spectra in the query set. We replace the preprocessing and similarity metrics from (Carey et al. 2015) with the pipeline described in Algorithm 1. Many optimization techniques will be applicable to this problem, but here we test only a random parameter search strategy as an example. To account for the randomness in the choice of spectra for the query set, the match accuracies used for each optimizer were computed as the mean accuracy over ten independent sets of query spectra. The results of the random parameter search are displayed in Figure 6.

Figure 7 further shows that this problem is easily optimized. The combinatorial preprocessing approach requires many trials to achieve its highest observed accuracy. The proposed generalized preprocessing method achieves improved accuracy initially, then converges to the highest observed accuracy in a small number of trials. This illustrates the significant time savings available with the generalized

function for an optimization problem that searches for the setting of $w$, $a$, and $b$ that maximizes classification accuracy. Using this loss function, we can apply well-known techniques from stochastic optimization to solve for an optimal pipeline for our specific task. This process is illustrated by the flow chart in Figure 5.

This optimization problem is actually quite feasible, in spite of the lack of an analytic gradient. The three-dimensional parameter space is fully bounded, forming a triangular prism-shaped search space. More importantly, both the MatchScore and preprocessing parameters change their components of the matching pipeline continuously and smoothly, which translates into smooth changes in match accuracy.
Combinatorial Preprocessing
Generalized Preprocessing
Match accuracy (%)
Number of random preprocessing configurations tested

Figure 7: Comparison of learning curves. The horizontal axis shows the number of iterations of random preprocessing configurations tested, and the vertical axis shows mineral species matching accuracy. Each filled region indicates the minimum and maximum performance over 100 random orderings of preprocessing configurations, and the black lines indicate empirical estimates of the expected matching accuracy. The red region represents the combinatorial preprocessing steps from Carey et al. (2015), while the blue region represents the novel technique for generalized preprocessing described in this paper.

Acknowledgments
This work was supported by NSF grant DUE-1140312 and NASA grant NNA14AB04A to the RIS4E node of the Remote, In Situ, and Synchrotron Studies for Science and Exploration (SSERVI). This is paper SSERVI-2015-158. We thank Bob Downs for making the RRUFF data available to us in a convenient format.

References
Norris, K. H., and Ritchie, G. E. 2008. Assuring specificity for a multivariate near-infrared (nir) calibration: The exam...
Figure 6: Results of a randomized parameter search for the automated whole-spectrum matching pipeline. The left side of each pair of figures shows the influence of the MatchScore parameter $w$ on overall accuracy. The right side shows the influence of preprocessing parameters $a$ and $b$ on overall accuracy, with higher accuracy in red and lower accuracy in blue. The white regions of the figures at right represent parameters outside the valid ranges for their respective preprocessing functions. These results demonstrate that each parameter has some influence on the resulting accuracy, and that incremental changes in a given parameter result in smooth changes in accuracy. The polynomial preprocessing function produces more consistent matching accuracy.


