BASELINE REMOVAL IN RAMAN SPECTROSCOPY: OPTIMIZATION TECHNIQUES. C. Carey[1], M. D. Dyar[2], T. F. Boucher[1], S. Giguere[1], C. M. Hoff[3], L. B. Breitenfeld[2], M. Parente[4], T. J. Tague, Jr.[5], P. Wang[5], and S. Mahadevan[1]. [1]School of Computer Science, Univ. of Massachusetts, Amherst, MA 01003, (ccarey@cs.umass.edu), [2]Mount Holyoke College, Dept. of Astronomy, South Hadley, MA 01075, 3dept. of Earth &Environmental Sci., Rensselaer Polytechnic Inst., Troy, NY 12180, [4]Dept. Electrical & Computer Engineering, Univ. of Massachusetts, Amherst MA 01003, [5]Bruker Optics, Inc., 19 Fortune Dr., Billerica, MA 01821.

Introduction: The imminent use of Raman spectroscopy for planetary exploration on Mars (Exomars, Mars 2020) and other bodies (e.g., Venus Roadmap) requires an infusion of work into development of appropriate databases and software. Successful use of Raman on extraterrestrial surfaces will require expansion of existing mineral databases and a dramatic improvement in accuracy of existing tools for mineral identification [1]. Our research group is addressing these issues on multiple fronts, including development work on optimizing software for mineral fingerprinting with Raman spectroscopy [2]. This report focuses on a a problem common to nearly all types of spectroscopy: baseline or continuum removal, in which some portion of a signal that is not necessary to the features of interest must be removed. Manual baseline removal is unprincipled and may produce variable results.

Many automated techniques have been proposed to subtract baselines from spectra, each with tunable parameters that may be set to refine the quality of the removed baseline. In practice, it is common to empirically find parameters that produce reasonable results on a small set of spectra, then use that setting to process the remaining samples in the spectral library. As no baseline correction algorithm is perfect and users cannot find perfect parameter settings for all data, continuum removal typically introduces systematic error to subsequent processing (**Figure 1**).

In this project, we evaluate the effects of six different baseline removal techniques on Raman spectra of minerals to test their effect on subsequent mineral identification matching routines. We also introduce an optimization method to automate baseline removal for this application.

Background: There are numerous methods for baseline removal in the literature based on polynomial [4], spline [5], LOESS [6], and Whittaker [7] techniques. In this project, we are evaluating three algorithms based on Whittaker smoothing (ALS, airPLS, and FABC), two that include iterative thresholding based on mean and variance (FABC, Dietrich), and two that involve fitting polynomial functions (Kajfosz-Kwiatek, Parente), as follows.

1. Asymmetric Least Squares (ALS) [3] finds a rough baseline with Whittaker smoothing, then lowers the weight for points above the baseline, and iterates until convergence is achieved.

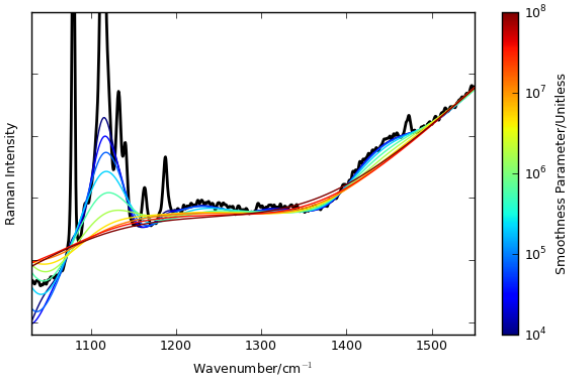

Figure 1. Errors introduced by baseline correction. Ten baselines (colored lines) were fit to a hanksite spectrum (black line) using the Asymmetric Least Squares (ALS) baseline correction algorithm [3], varying the smoothness parameter λ on a log scale between 104 and 108 as suggested by the authors. Choice of the best value for λ is largely subjective, and may require quality trade- offs: the low values of λ in blue clearly overfit the peaks on the left of the figure, but do a better job of modeling the baseline on the right half. Large values of λ (in red) preserve the peaks but often force the baseline above the original spectrum.

2. Adaptive Iteratively Reweighted Penalized Least Squares (airPLS) [8] is similar to ALS except that it uses the sum of differences between signal and baseline to adjust weights intelligently.

3. Fully Automatic Baseline Correction (FABC) [9] applies a continuous wavelet transform (CWT) to approximate the first derivative of the signal. It then finds "peak bands" by selecting all bands with intensities greater than the mean + k × standard deviation (where k is a tuned parameter), a process called *iterative thresholding*. The procedure then performs a single iteration of weighted Whittaker smoothing to produce a baseline followed by recalculation of the mean and standard deviation with any peak bands excluded.

4. The method of Dietrich [10] is related to FABC in that it performs iterative thresholding on the (squared) first derivative of the spectrum to find non-peak points. Then, rather than a smoothing step, it reconstructs the baseline for all bands using simple

linear interpolation between non-peak bands.

5. The Kajfosz-Kwiatek method, which uses a non-polynomial approximation of background [11], is commonly used in processing x-ray near-edge absorption spectra. It fits a set of polynomials to the signal, both concave up and down, and then takes the maximum value over all polynomials as the baseline.

6. Parente [12,13] developed a baseline removal technique for remote-sensed hyperspectral data from the visible and near-IR regions of the electromagnetic spectrum. His novel algorithm casts baseline removal as a linear program that solves for the coefficients of a Hermite polynomial, maximizing the size of the baseline while constraining the baseline to lie below the input signal across all bands.

In any of these approaches, the difficulty is in choosing which of the adjustable parameters is appropriate for any given spectrum; this is usually done manually. We treat this choice as an optimization problem in which baseline parameters are adjusted until the removed baseline has zero predictive value and is not correlated with any compositional variable.

Data: For this project, we tested six baseline correction algorithms by applying them to a set of 184 Raman spectra of pure mineral samples collected at Bruker Optics, Inc., in Billerica, MA. Once corrected, each spectrum was matched against a 3950-sample subset of the RRUFF mineral library. The matching process used a cosine distance metric to perform *k*-nearest neighbors classification. Granularity in match accuracy was provided by considering matches at each level of the Dana classification hierarchy.

The parameter for success of baseline removal was accuracy of correctly matching each unknown spectrum against its equivalent in the RRUFF database. Each mineral name was matched with its four-part (separated by periods) Dana classification number in which the first number is mineral class, the second is mineral type, the third is the mineral group, and the fourth is the specific mineral species.

Spectra in the RRUFF library are provided with baseline correction already applied [14], so the target set remained unchanged across all classification trials. This experimental setup was chosen to mimic the common scenario of a preparing small set of unknown spectra to match against a well-curated library.

Results: Classification accuracy was predicted for each of the six described baseline correction algorithms, as well as for a control setting where no baselines were removed before matching. Each algorithm was invoked with the suggested default parameters, and no parameter tuning was performed. **Table 1** shows classification accuracy at multiple Dana levels. Scores are reported as the percent of query spectra with correct classification labels. For example, without baseline removal, only 8% of mineral species were correctly identified. Best results were obtained when using the airPLS method of Zhang et al. [6].

Table 1. Classification Accuracy at Different Dana Levels

	Class	Type	Group	Species
Control	17.39	9.24	9.24	8.15
ALS	66.30	61.41	57.07	46.74
airPLS	**66.85**	**61.41**	**57.61**	**48.37**
Dietrich	52.72	47.28	46.2	39.67
FABC	60.87	56.52	54.89	45.65
Kajfosz-Kwiatek	65.22	60.33	55.43	45.11
Parente	52.17	47.83	46.74	37.5

Discussion: These preliminary results already demonstrate important conclusions. Although Raman matching performed with the CrystalSleuth matching software provided on the RRUFF site does not use any baseline correction, our state-of-the-art automatic baseline algorithms yield much better performance. Of the models tested so far, those based on Whittaker smoothing (ALS, airPLS, FABC) tend to work well even without tuning. Finally, there is room for improvement in mineral identification software from Raman spectroscopy. Table 1 shows that correct mineral species identifications are obtained < 50% of the time, even with pure mineral spectra (confirmed by x-ray diffraction) and high-quality baseline removal. Work on database development and software improvement will be needed before Raman spectroscopy can be usefully employed on remote planetary surfaces.

Our group is working not only to acquire new data to supplement RRUFF but to introduce new algorithms for both baseline removal and mineral identification. Work is in progress to develop better methods for automatic parameter tuning given limited data.

Acknowledgments: This research was supported by NSF grants CHE-1306133 and CHE-1307179 as well as by the RIS^{4}E node of NASA SSERVI.

References: [1] Carey C. et al. (2015) *J. Raman Spectrosc.*, in review. [2] Carey C. (2014) Geo-Raman 8, Abstract #5053. [3] Eilers P. H. C. and Boelens H. F. M. (2005) Baseline correction with asymmetric least squares smoothing. Leiden Univ. Medical Centre Report. http://www.science.uva.nl/~hboelens/. [4] Gan F. et al. (2006) *Chemom. Intell. Lab. Syst.* 82, 59–65. [5] de Boor C. (1978) *A Practical Guide to Splines*. Springer, New York. [6] Ruckstuhl A. F. et al. (2001) *J. Quant. Spectrosc. Radiat. Transf.*, 68, 179–193 [7] Eilers P. H. C. (2004) *Anal. Chem.*, 76, 404–411. [8] Zhang Z.-M. et al. (2010) *Analyst*, 135, 1138-1146. [9] Cobas J. C. et al. (2006) *J. Mag. Reson.*, 183, 145-151. [10] Dietrich W. et al. (1991) *J. Mag. Reson.*, 91, 1-11. [11] Kajfosz J. and Kwiatek W. M. (1987) *Nucl. Instrum. Methods Phys. Res.*, 22, 78-81. [12] Parente M. (2008) *LPSC XXXIX*, Abstract #2528. [13] Parente M. (2010) Ph.D. thesis, Stanford Univ. [14] http://RRUFF.info/.