MULTIVARIATE ANALYSIS IN VIBRATIONAL SPECTROSCOPY

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Chemical Engineering

in the

College of Graduate Studies

University of Idaho

by

Andrew T. Weakley

June 2014

Major Professor: D. Eric Aston, Ph.D.

Authorization to Submit Dissertation

This dissertation of Andrew T. Weakley is submitted for the degree of Doctor of philosophy with a Major in Chemical Engineering and titled "Multivariate Analysis in Vibrational Spectroscopy" has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____Date_____
D. Eric Aston, Ph.D

Committee Members _____Date_____
Matthew Morra, Ph.D

_____Date_____
Vivek Utgikar, Ph.D

_____Date_____
James Moberly, Ph.D

Department Administrator _____Date_____
Wudneh Admassu, Ph.D

Discipline's College Dean _____Date_____
Larry Stauffer, Ph.D

Final Approval and Acceptance

Dean of the College of Graduate Studies:
_____Date_____
Jie Chen, Ph.D.

Abstract

Multivariate analysis in vibrational spectroscopy involves the application of procedures and protocols from multivariate statistics, signal processing, and experimental design to elucidate physico-chemical phenomena studied using high-dimensional data sets acquired from multichannel instrumentation. These so-called chemometric procedures are applicable to a range of questions relevant to the practice of analytical chemistry and engineering. Formal areas such as exploratory data analysis, multivariate classification, multivariate calibration, and curve resolution are common focus areas of chemometricians. This dissertation concerns the exclusive use of latent variable models to assess analytical quantities and chemical systems via vibrational spectroscopy for the purpose of data exploration and calibration. The principal goal of this dissertation was to develop and/or innovate domains pertinent to multivariate calibrations utilizing principal component analysis (PCA), principal component regression (PCR), and/or partial least-squares (PLS) regression. Specific objectives included: developing two novel baseline correction algorithms (chapter 2 and 3) to optimally preprocess vibrational spectra prior to calibration, applying PCA and PCR to probe the specific hydrogen-bonding behavior of thermoplastic polyurethane (TPU) blends (chapter 4), utilizing a PLS regression to determine the quantity of filter-adsorbed silica from metal/non-metal mines (chapter 5), and finally developing rigorous and comprehensive model selection criteria to choose a best PLS regression among viable alternative offered by the novel application of backward Monte Carlo unimportant variable elimination (BMCUVE). Overall, those chapters exclusively focused on multivariate calibration using vibrational spectra (chapter 4-6) demonstrate that statistical and scientific validity converge under the auspices of a well-designed chemometric analysis.

Acknowledgments

First and foremost, it is important to recognize the detailed, thoughtful, and intellectually stimulating contributions to this dissertation project by major professor D. Eric Aston, professor emeritus Peter R. Griffiths, and all feedback from committee members professor Vivek Utgikar, professor James Moberly, and professor Matthew Morra. In many ways, this work represents the cumulative insight of 100+ years of combined experience, without which it would have stood as an unsatisfying fragment of its final form. I'd like to particularly thank Dr. Aston for extensive financial support as well as releasing me from any teaching responsibility so that I could comfortably pursue an independent study of chemometrics and multivariate statistics. Naturally, I'd like to thank all coauthors contributions' to the published works that comprise the majority of this dissertation (Chapters 2-5). Finally, I'd like to thank the department's administrative assistants, in particular Gail Bergman and Margaret Baker, for effectively addressing my often ill-informed questions concerning departmental protocols and procedures. University staff is truly the backbone of this fine institution.

Table of Contents

List of Figures

List of Tables

Chapter 1. Chemometric analysis in vibrational spectroscopy

Chemometrics approaches conflate pertinent aspects of applied statistics, signal processing, and experimental design with computer science to explore, visualize, and interrogate high-throughput chemical data [1, 2]. Contemporary chemometric approaches are routinely applied to assess large-variable data-analytical problems; namely, those employing spectroscopic or related measurements to ill-posed, "large variable (p), small sample (n)" experimental designs. An experimental design, $[X]$, is generally organized as n rows of spectra (or measurements) defined on p variables (*e.g.*, channel-wavelengths). For example, a quantitative structure activity relationship (QSAR) study might employ several hundred molecular descriptors (*e.g.*, polarizability, lipophilicity) as X-variables (columns) to predict the toxicity (Y) of a proposed pharmaceutical drug [3-6]. Alternatively, an analyst might train a neural network to assign an unknown, sample-observation to one of many predefined groups using information inscribed onto thousands of chromatographic features [4-8]. In such instances, chemometric methods provide numerous and capable data reduction, feature selection, importance-weighting, and/or pattern recognition protocols to elucidate analytical quantities or class affiliation(s) in terms of germane physicochemical signatures.

Arguably, the *raison d'etre* of chemometrics is the application of multichannel vibrational spectroscopy to the rapid and inexpensive determination of an analytical quantity or property of interest [1, 9, 10]. An example pioneering success involves the total characterization of protein, starch, and water content in fish meal using near infrared (NIR) spectrometry via a principal component regression (PCR) [11, 12]. Since this and related developments, agricultural quality control and process monitoring applications routinely pair NIR spectrometry with multivariate analysis, often employing the partial least-squares (PLS) modeling procedure [13-16]. Advances in nonlinear calibration and improved signal correction approaches have further extended the reach of vibrational spectroscopy to the determination of analytical quantities [17-21].

As mentioned above, it is common to have many more variables than sample-spectra in a calibration problem ($n \ll p$). When variables vastly outnumber sample-observations, the problem is ill-conditioned and insoluble using a least-squares estimator [10]. Phrased differently, spectroscopic variables in $[X]$ are usually highly correlated (collinear). Excessive collinearity results in unstable parameter estimates for a least-squares problem leading to an

ultimately useless predictive model [9]. For this reason a rank-reducing procedure is generally performed to identify the true dimensionality of the calibration problem prior or simultaneously to the prediction of a target analyte [9, 22]. Specifically, the data structure itself (*e.g.*, dominant variations in$[X]$) is mapped onto a handful of latent variables (A) that presumably capture the important, underlying chemical effects linking spectroscopic and analyte response.

A PCR utilizes a rank-reducing procedure via a principal component analysis (PCA) [23, 24]. A PCA captures the dominant sources of variation in $[X]$ by defining a new set variables (A.K.A., scores) that are constrained to be orthogonal (uncorrelated) linear combinations of the original channel-variables (columns in $[X]$). Early algorithms extracted principal components sequentially with the length (2-norm) of the first principal component constrained to be larger than the second, and so on, until the variance in $[X]$ was completely explained by the components [25, 26]. Furthermore, a critical assumption of PCA is that only a handful of principal components (=A) describe the analytically *useful* variation in $[X]$ while the minor components contain purely spectroscopic noise [23, 27]. Once the true rank (=A) of $[X]$ is determined using an approach such as cross-validation [28], the principal component scores ($[T]$) are used as predictors in a least-squares regression to predict an analytical quantity or property ($\boldsymbol{y}$) [29]. By definition, using the uncorrelated principal component scores as the predictors solves the collinearity problem thereby stabilizing the regression. Chapter 4 will demonstrate a supplementary benefit of PCA; namely, the ability to visualize the associations and dissimilarities of sample-spectra on score plots.

To better understand the tangible benefits of PCA, a hypothetical example is in order. Assume that $[X]$ contains n rows of UV-vis absorbance spectra where the concentration of three different alkanes are distinct for each of the n spectra. Furthermore, these n spectra are digitized upon p spectral channels ($n \ll p$). Further assume that the linear Beer-Lambert law applies, the spectroscopic signatures of each alkane are resolved, and any noise is random, normally distributed, and additive [27, 30]. In this ideal case, the significant number of principal components in $[X]$ *precisely corresponds to the number of spectroscopically-active chemical species in the solutions* which in this case are the three alkanes ($A = 3$). As a purely mathematical transformation, PCA captures only the spectroscopic-response of each alkane on the principal components in an abstract/ambiguous way [27]. In other words, the

principal components span the same linear subspace as the three alkanes but the components are not *selective* to any alkane in particular. More advanced methods such as multivariate curve resolution (MCR) or a target factor analysis (TFA) attempt to simultaneously or sequentially *resolve* the pure component spectra as well as the corresponding concentration profiles contained within the principal components [31-36].

Real deviations from Beer's law, or linearity in general, may complicate the estimation of the number of principal components [1]. Deviations might occur due to an inherently nonlinear chemical environment (*e.g.*, intermolecular association among species in solution) or result from instrument and/or measurement biases. Regardless of source, interactions among species in solution, matrix effects, measurement artifacts, scattering, and background can represent large sources of systematic variation in$[X]$. The higher principal components can capture these effects thereby artificially inflating the size of the component space ($A > A_{true}$). This can lead to inaccurate predictions in a PCR and certainly complicate the interpretation the regression's latent structure. Chapter 2-6 will demonstrate that the choice of variable scaling and preprocessing (*e.g.*, baseline correction) can help dampen or completely ameliorate the influence of these effects on the predictive performance and interpretation of latent variable models [11, 37, 38].

Where PCR uses uncorrelated principal components as predictor-variables in a least-squares regression, PLS develops analyte-specific latent variables by maximizing the covariance between both $[X]$ and $\boldsymbol{y}$ [9]. Arguably this improves the relative accuracy of the regression and can lead to a simpler (fewer component) model [1]. In PLS, components are analogously extracted where cross-validation or additional methods are used to estimate the precise number of components to use in the predictive model [28, 39-41]. The details of the non-linear, iterative partial least-squares (NIPALS) algorithm and alterative simultaneous PLS method (SIMPLS) are described elsewhere [26, 42].

Overall, both PCR and PLS are what are known as bilinear models [43]. Specifically, a single-response bilinear model is represented as

$$[X] = \left[\hat{T}\right][\hat{P}]^T + [E]_x \qquad \text{Equation 1.1}$$

$$\hat{\boldsymbol{y}} = \left[\hat{T}\right]\hat{\boldsymbol{q}} + \boldsymbol{e}_y \qquad \text{Equation 1.2}$$

where $[X]$ is an $n\ x\ p$ design matrix of vibrational spectra, $[\hat{T}]$ is an $n\ x\ A$ matrix of estimated component scores, $[\hat{P}]$ are the $p\ x\ A$ component loadings, $\hat{\boldsymbol{y}}$ is a $n\ x\ 1$ estimated vector of analyte response, $\hat{\boldsymbol{q}}$ are the regression coefficients ($A\ x\ 1$), and $[E]_x$ and $\boldsymbol{e}_y$ are X- and y-block errors, respectively. Note, this bilinear form describes many chemometric techniques including methods from multivariate classification and the aforementioned MCR [44-46].

Martens and coworkers [43] provide a concise rubric detailing how to approach a bilinear modeling procedure in analytical chemistry. Inspired by these criteria, this manuscript details numerous innovations and algorithms roughly conforming to the steps outlined by Martens and coworkers [43] when applied to multivariate calibration using vibrational spectra. The algorithms and/or innovations developed by this author address domains related to:

1. Spectral preprocessing including semi-automated baseline correction and variable scaling techniques
2. Influence of preprocessing on the estimation, exploration, and interpretation of PCA applied to a real polymer system
3. Role of feature (wavenumber) selection in the precision/stability improvement of PLS regressions
4. Development of comprehensive model selection criteria to aid in selecting stable, accurate, and precise PLS models given a range of viable alternatives

The principal goal of this dissertation is to develop and improve the design, measurement, preprocessing, interpretation, and long-run stability of multivariate predictive models employing vibrational spectra as design variables. Prior to executing the primary objective, priority was assigned to two major technical areas: *numerical analysis* as well as *experimental design, analysis, and validation*. Given the critical role of latent variable estimation in multivariate models, the first technical area focused on directly developing novel computational routines to remove background and baseline (Domain 1; Chapter 2 and 3) as well as improve the precision of PLS models (Domain 3; Chapter 5 and 6). The second technical area concerned applying these numerical routines and preexisting approaches to

interrogate complex chemical mixtures (Domain 2; Chapter 4) and develop comprehensive model selection criteria to identify an optimal calibration given the choice of viable alternatives (Domain 4; Chapter 6). The specific objectives conform to each chapter as follows:

Chapter 2 presents a novel method of data preprocessing by improving the computational baseline correction of Raman, infrared, surfaced-enhanced infrared absorption, vibrational circular dichroism spectra using a locally-weighted, adaptive least-squares thresholding procedure. This method of baseline correction is particularly unique in that the estimation procedure was model-free (i.e., assumed no continuous function captured the low-frequency baseline components), accommodated bidirectional (both positive- and negative-going) spectral lineshapes, and remains completely generalizable to any spectrum that requires baseline estimation and removal. Chapter 3 modifies and extends model-free baseline correction to the correction of solvent baselines in vibrational circular dichroism (VCD) spectra. Chapter 3 elucidates an absolutely novel, first attempt to remove solvent background from VCD spectra using purely computational estimation-correction.

Chapter 4 successfully explores the specific hydrogen-bonding behavior of segmented, thermoplastic polyurethane (TPU) blends using PCA applied to TPU Raman spectra. Note, the algorithm developed in Chapter 2 was applied to suppress fluorescence baseline perturbations observed in the Raman spectra of TPU blends. Following the interpretation of PCA, a PCR is performed to predict the hard segment fraction of each polymer blend using the PCA scores derived from relevant regions of the Raman spectra. The composition of these blends ($\boldsymbol{y}$) were estimated from nuclear magnetic resonance (NMR) spectra. Applying chemometric analyses to study the hydrogen-bonding behavior of TPU blends is a proof-of-principle representing the first major effort to elucidate the complex morphological behavior of segmented polyurethanes using Raman spectroscopic signatures, PCA, and multivariate calibration.

Chapter 5 describes the feasibility of applying portable Fourier transform-infrared (FT-IR) instrumentation at non-coal mines to rapidly predict the occupational exposure of mine personnel to airborne silica (α-quartz). Typically, regulatory enforcement methods for airborne silica exposure in US non-coal mines mandates the quantification of silica using a

very specific absorbance range in the IR spectrum of mine dust [47]. This range, which captures the so-called "α-quartz doublet," is susceptible to mineral confounds (e.g., kaolin clays) and requires expertise on the part of the analyst to accurately (and consistently) predict the mass of silica. Presently, such procedures are time consuming where an assessment of personnel exposure lags well behind the pace of mining operations (+2 weeks).

A PLS regression was developed in lieu of the existing protocol (which included tedious sample preparation and rudimentary predictive modeling) to determine the mass of α-quartz present in mine-dust adsorbed onto personal polymeric sampling filters using FT-IR spectra. Major improvements over the existing FT-IR enforcement method [47] included: leaving the polymeric sampling filter (and any organic contaminants) in place during the FT-IR measurement, improving accuracy using α-quartz vibrations outside the doublet range, and applying a numerical wavelength selection algorithm, designated backward Monte Carlo unimportant variable elimination (BMCUVE), to objectively select the most stable and pertinent spectroscopic feature in $[X]$ to predict α-quartz. The precision and stability of the regression coefficient estimates were evaluated and compared to an FT-IR enforcement analog to demonstrate performance improvement.

Chapter 6 elucidates of the fundamental operations of BMCUVE as well as the ability for comprehensive model selection criteria to inform the choice of an optimum PLS regression comparing common approaches from the literature to give context to the relative effectiveness of BMCUVE. Feature selection using BMCUVE was applied in tandem with PLS regression to predict hard segment fraction in TPU blends (Raman), moisture content in corn (NIR), and airborne silica content adsorbed onto polymeric sampling filters (FT-IR). Seven comprehensive model selection criteria that prioritized measures of long-run predictive ability and validation sample-statistics were used to select the best PLS models given dozens of viable alternatives offered by BMCUVE. Overall, these seven comprehensive criteria substantially aided the selection of the most stable, precise, and accurate PLS regressions given a range of viable candidates provided by BMCUVE. More importantly, this chapter demonstrates that any future model selection problem should assign importance to multiple criteria, which is not yet common practice, in the interest of identifying predictive models free from spectroscopic artifacts, superfluous baseline, and excessive redundancy.

A final chapter summarizes the strengths and limitations of the foregoing work in Chapters 2-6. Practical recommendations concerning the exploration, prediction, and optimization of multivariate calibrations using vibrational spectra are provided according to the four domains (1-4) presented above.

References

[1] P. Gemperline, *Practical Guide to Chemometrics*: CRC/Taylor & Francis, 2006.

[2] R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*: Wiley, 2003.

[3] M. Olah, C. Bologa, and T. Oprea, "An automated PLS search for biologically relevant QSAR descriptors," *Journal of computer-aided molecular design,* vol. 18, pp. 437 - 449, 2004.

[4] M. A. Lill, "Multi-dimensional QSAR in drug discovery," *Drug Discovery Today,* vol. 12, pp. 1013-1017, 12// 2007.

[5] L. Eriksson and E. Johansson, "Multivariate design and modeling in QSAR," *Chemometrics and Intelligent Laboratory Systems,* vol. 34, pp. 1-19, 8// 1996.

[6] W. J. Dunn Iii, "Quantitative structure—activity relationships (QSAR)," *Chemometrics and Intelligent Laboratory Systems,* vol. 6, pp. 181-190, 9// 1989.

[7] X.-M. Sun, X.-P. Yu, Y. Liu, L. Xu, and D.-L. Di, "Combining bootstrap and uninformative variable elimination: Chemometric identification of metabonomic biomarkers by nonparametric analysis of discriminant partial least squares," *Chemometrics and Intelligent Laboratory Systems,* vol. 115, pp. 37-43, 6/15/ 2012.

[8] G. J. Patti, O. Yanes, and G. Siuzdak, "Innovation: Metabolomics: the apogee of the omics trilogy," *Nat Rev Mol Cell Biol,* vol. 13, pp. 263-269, 04//print 2012.

[9] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems,* vol. 58, pp. 109-130, 2001.

[10] T. Næs, Isaksson, T., Fearn, T. Davies, T., *An User-friendly Guide to Multivariate Calibration and Classification*. Chichester, West Sussex: Nir Publications, 2002.

[11] T. Isaksson and T. Næs, "The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy," *Applied Spectroscopy,* vol. 42, pp. 1273-1284, 1988/09/01 1988.

[12] T. Næs and T. Isaksson, "Selection of Samples for Calibration in Near-Infrared Spectroscopy. Part I: General Principles Illustrated by Example," *Applied Spectroscopy,* vol. 43, pp. 328-335, // 1989.

[13] P. Williams and K. H. Norris, *Near-infrared technology in the agricultural and food industries*: American Association of Cereal Chemists, 1987.

[14] B. G. Osborne and T. Fearn, *Near infrared spectroscopy in food analysis*: Longman Scientific & Technical, 1986.

[15] J. S. Shenk and M. O. Westerhaus, "Population Definition, Sample Selection, and Calibration Procedures for Near Infrared Reflectance Spectroscopy," *Crop Sci.,* pp. 469-474, 1991.

[16] Y. Wu, Y. Jin, Y. Li, D. Sun, X. Liu, and Y. Chen, "NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of an extraction process," *Vibrational Spectroscopy,* vol. 58, pp. 109-118, 1// 2012.

[17] P. J. Gemperline, J. R. Long, and V. G. Gregoriou, "Nonlinear multivariate calibration using principal components regression and artificial neural networks," *Analytical Chemistry,* vol. 63, pp. 2313-2323, 1991/10/01 1991.

[18] S. Wold, H. Antti, F. Lindgren, and J. Öhman, "Orthogonal signal correction of near-infrared spectra," *Chemometrics and Intelligent Laboratory Systems,* vol. 44, pp. 175-185, 12/14/ 1998.

[19] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics,* vol. 16, pp. 119-128, 2002.

[20] F. Despagne and D. Luc Massart, "Neural networks in multivariate calibration," *Analyst,* vol. 123, pp. 157R-178R, 1998.

[21] H. Martens, J. P. Nielsen, and S. B. Engelsen, "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures," *Analytical Chemistry,* vol. 75, pp. 394-404, 2003/02/01 2003.

[22] L. Zhang and S. Garcia-Munoz, "A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective," *Chemometrics and Intelligent Laboratory Systems,* vol. 97, pp. 152-158, 7/15/ 2009.

[23] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, pp. 433-459, 2010.

[24] I. T. Jolliffe, *Principal Component Analysis*: Springer, 2002.

[25] S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics,* vol. 20, pp. 397-405, 1978.

[26] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta,* vol. 185, pp. 1-17, // 1986.

[27] E. R. Malinowski, *Factor Analysis in Chemistry*: Wiley, 2002.

[28] P. Eshghi, "Dimensionality choice in principal components analysis via cross-validatory methods," *Chemometrics and Intelligent Laboratory Systems,* vol. 130, pp. 6-13, 1/15/ 2014.

[29] I. T. Jolliffe, "Note on the use of principal components in regression," *Applied Statistics,* vol. 31, pp. 300-3, 1982.

[30] M. Otto and W. Wegscheider, "Spectrophotometric multicomponent analysis applied to trace metal determinations," *Analytical Chemistry,* vol. 57, pp. 63-69, 1985.

[31] H. Abdollahi and R. Tauler, "Uniqueness and rotation ambiguities in Multivariate Curve Resolution methods," *Chemometrics and Intelligent Laboratory Systems,* vol. 108, pp. 100-111, 10/15/ 2011.

[32] J. Jaumot and R. Tauler, "MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution," *Chemometrics and Intelligent Laboratory Systems,* vol. 103, pp. 96-107, 10/15/ 2010.

[33] J. Jaumot, R. Gargallo, A. de Juan, and R. Tauler, "A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB," *Chemometrics and Intelligent Laboratory Systems,* vol. 76, pp. 101-110, 3/28/ 2005.

[34] E. R. Malinowski and M. McCue, "Qualitative and quantitative determination of suspected components in mixtures by target transformation factor analysis of their mass spectra," *Analytical Chemistry,* vol. 49, pp. 284-287, 1977/02/01 1977.

[35] L. Shao and P. R. Griffiths, "Obtaining Qualitative Information on Trace Species in Continuous Open-Path Fourier Transform Spectroscopic Measurements Using Target Factor Analysis and Related Techniques," *Analytical Chemistry,* vol. 79, pp. 2118-2124, 2007/03/01 2007.

[36] P. J. Gemperline, "A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis," *Journal of Chemical Information and Computer Sciences,* vol. 24, pp. 206-212, 1984.

[37] T. Bocklitz, A. Walter, K. Hartmann, P. Rosch, and J. Popp, "How to pre-process Raman spectra for reliable and stable models?," *Anal Chim Acta,* vol. 704, pp. 47-56, Oct 17 2011.

[38] L. Xu, Y. P. Zhou, L. J. Tang, H. L. Wu, J. H. Jiang, G. L. Shen, *et al.*, "Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration," *Anal Chim Acta,* vol. 616, pp. 138-43, Jun 2 2008.

[39] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," vol. 56, pp. 1-11, 2001.

[40] J. Shao, "Linear Model Selection by Cross-validation," *Journal of the American Statistical Association,* vol. 88, pp. 486-494, 2013/08/16 1993.

[41] N. M. Faber and R. Rajkó, "How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative," *Analytica Chimica Acta,* vol. 595, pp. 98-106, 7/9/ 2007.

[42] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems,* vol. 18, pp. 251-263, 3// 1993.

[43] H. Martens, M. Høy, F. Westad, D. Folkenberg, and M. Martens, "Analysis of designed experiments by stabilised PLS Regression and jack-knifing," *Chemometrics and Intelligent Laboratory Systems,* vol. 58, pp. 151-170, 10/28/ 2001.

[44] Y. Tominaga, "Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN," *Chemometrics and Intelligent Laboratory Systems,* vol. 49, pp. 105-115, 9/6/ 1999.

[45] S. J. Dixon and R. G. Brereton, "Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure," *Chemometrics and Intelligent Laboratory Systems,* vol. 95, pp. 1-17, 1/15/ 2009.

[46] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg, "OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification," *Journal of Chemometrics,* vol. 20, pp. 341-351, 2006.

[47] Mine Safety and Health Adminstration, “Infrared Determination of Quartz in Respirable Coal Mine Dust - Method No. MSHA P7,” Pittsburgh Safety and Health Technology Center, Pittsburgh, 2013.

Chapter 2. Automatic baseline subtraction of vibrational spectra using minima identification and discrimination via adaptive, least-squares thresholding

Reproduced with permission (Appendix A): Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, Applied Spectroscopy, 2012, **66**(5): 519-529.

Abstract:

A method of automated baseline correction has been developed and applied to Raman spectra with a low signal-to-noise ratio and surface-enhanced infrared absorption (SEIRA) spectra with bipolar bands. Baseline correction is initiated by dividing the raw spectrum into equally spaced segments in which regional minima are located. Following identification, the minima are used to generate an intermediate second-derivative spectrum where points are assigned as baseline if they reside within a locally defined, threshold region. The threshold region is similar to a confidence interval encountered in statistics. To restrain baseline and band point discrimination to the local level, the calculation of the confidence region employs only a predefined number of already-accepted baseline minima as part of the sample set. Statistically-based threshold criteria allow the procedure to make an unbiased assessment of baseline points regardless of the behavior of vibrational bands. Furthermore, the threshold region is adaptive in that it is further modified to consider abrupt changes in baseline. The present procedure is model-free insofar as it makes no assumption about the precise nature of the perturbing baseline nor requires treatment of spectra prior to execution.

Key Words: Baseline correction; automatic; Raman spectra; bipolar bands; adaptive least-squares thresholding

Introduction:

The characteristically narrow bands observed in Raman and infrared spectra are typically superimposed on broad, low-frequency components referred to as baseline or background.[1] Such baselines distort the bands of the components of the sample and hence diminish the accuracy, robustness, and interpretation of chemical effects unearthed from univariate or multivariate analysis, such as partial least squares regression or target factor

analysis.[2-4] The removal of baselines with minimal operator bias is frequently an important step in certain spectroscopic protocols, since during manual baseline correction procedures, a bias is sometimes inappropriately introduced into a spectral measurement that may degrade reproducibility. An automated procedure for baseline correction becomes essential in the case of vibrational hyperspectral imaging, where thousands of spectra must be pre-processed accurately prior to image generation.

Methods of semi-automatic and automatic baseline correction have been reported and vary in complexity and computational rigor. Methods derived from the field of digital signal processing employ such techniques as Fourier or wavelet transformations to remove low frequency components.[3, 5] Other methods include applying first and/or higher order derivatives to identify and distinguish between baseline and spectral bands,[2, 6] the use of statistical and probabilitistic approaches such as principal components analysis or maximum entropy algorithms,[7-9] or assuming that the baseline has a continuous functional form, usually a low-order polynomial, in which case a routine is employed that gradually "strips" the baseline from the raw spectrum.[10-12] In general, methods tend to employ a combination of the aforementioned techniques depending on application and need. [13-15] For example, just about every algorithm allows a slowly varying baseline to be removed when all the bands in the spectrum are narrow (see later in Figure 2.3, for example.) On the other hand when the frequency components of one or more of the bands in the spectrum are of the same order as the baseline, as may be the case for the O-H stretching mode of hydrogen-bonded hydroxyl groups, most baseline correction algorithms remove the broad spectral features along with the baseline. Even lower success is encountered when the spectrum contains bipolar bands, as the minima of the negative-going region of bipolar bands are often chosen to represent the baseline.

The approach that we report in this paper continues the tradition of drawing on multiple approaches to automate baseline correction. It proceeds by locating minima in equally spaced regions of an infrared or Raman spectrum, taking the first-derivative of the intermediate minima-only spectrum, generating a second-order difference spectrum, and then developing a statistically based, adaptive thresholding scheme to reject or accept minima points as part of the baseline. Once baseline points are located, simple linear interpolation is employed to remove the baseline from the raw spectrum. Unlike most derivative methods,

the present technique makes few assumptions about the positivity, smoothness, band shape, or noise level in the raw spectrum. In fact this method is "model free" insofar as it assumes that the perturbing baseline has no global functional form but contains relatively smooth local variations.[16] Adaptive least-squares thresholding adds an additional layer of complexity in that it relaxes the assumption that the baseline minima are always varying gradually.

Experimental Methods:

The present procedure assumes that the slope differences between adjacent baseline points are much smaller than differences between adjacent baseline and Raman or IR band points. Figure 2.1 illustrates the rationale behind the method. The algorithm is initialized by dividing the spectrum into sixty-four equally spaced segments where spectral subdivision is similar to that employed by Bruker's "rubber-band" method.[17, 18] The choice of spectral subdivision is entirely arbitrary but we have found that sixty-four points per segment is generally appropriate.

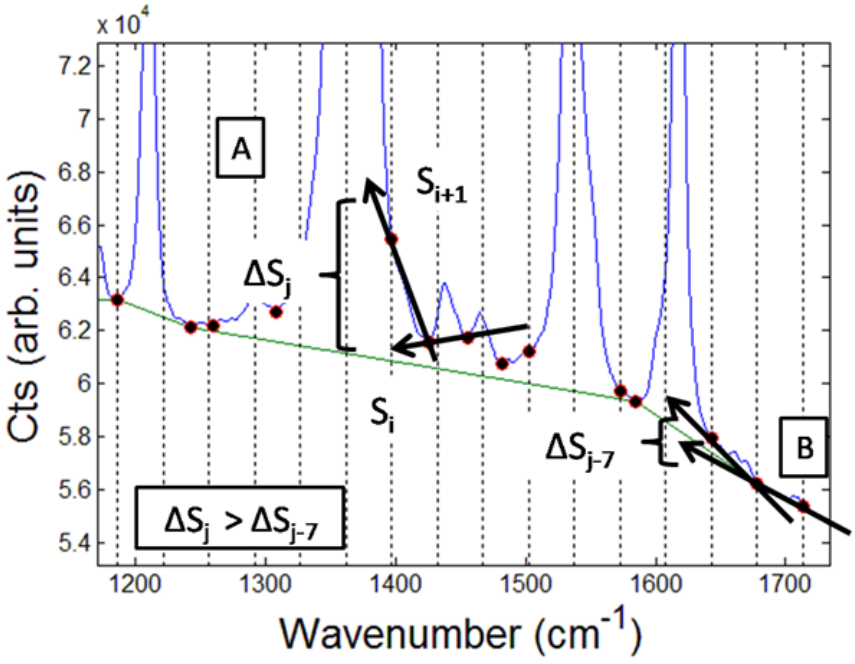


Figure 2.1: Reasoning underlying the proposed baseline subtraction method. First, the spectrum is divided into 64 equally spaced regions and the minima of each region are located (black circles). From lowest to highest wavenumber, the first derivative (or local slope) is taken (S_i). If the difference in slope (ΔS_i) containing the current point (i+2) is much larger than the differences in slope for previously accepted baseline points (B), then the i+2 point is considered a vibrational band (A) and rejected as a baseline value ($\Delta S_i > \Delta S_{i-7}$).

Once regions are specified, minima are located. These minima act to define a new, intermediate spectrum on which all further calculations are performed. (Note, if a spectrum is plotted from highest to lowest intensity, e.g., a spectrum that is linear in transmittance, the spectrum must be inverted prior to identification of the minima. After baseline correction, the spectrum may then be reinverted.) After minima identification, the first derivative of the intermediate minima-only spectrum is calculated using a simple forward-difference approach. Restricting derivative calculations to regional minima (which are assumed as widely distributed across spectral channels) eliminates the need for smoothing and places fewer constraints on the requirement for an adequate signal-to-noise ratio (SNR).[19] For the present purposes, it is convenient to maintain the elementary view that the first derivative is the local slope between two adjacent minima. Additionally, in the event that minima are closely spaced, i.e., within three channels, the minimum at lower wavenumber is retained while the other is removed from the calculations. Next, the absolute value of the difference between each slope (ΔS_j) is calculated. Retaining the absolute value of this intermediate second-order difference spectrum adheres strictly to positivity; the benefits of which are discussed elsewhere.[15]

Depending on the window size (n) specified, a simple moving average[20] using the n^{th} previous ΔS_j values is calculated. The most appropriate windows size, as well as additional model parameters, was selected based on the results of baseline correction applied to simulated Raman and SEIRA spectra, the methodology and results of which are discussed in detail below. This local-moving average, $\overline{\Delta S_k}$, is used in conjunction with the local standard deviation, σ_k, of the past n points to determine whether the i+2 minimum value is accepted as baseline or rejected as belonging to a vibrational band. When minima values are rejected they are removed from successive calculations of the moving average and local standard deviation. This ensures that only minima values corresponding to baseline points govern the rejection criteria of each region. Figure 2.2 illustrates the reasoning behind this approach.

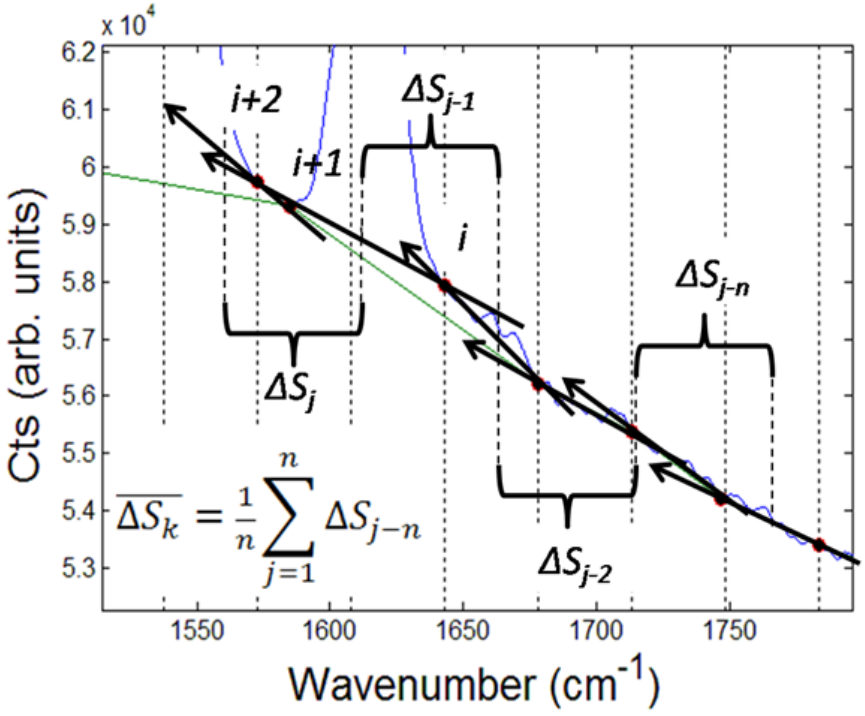


Figure 2.2: Detailed illustration of baseline subtraction method. If the absolute difference in slope between the current value (ΔS_j) is larger than $\overline{\Delta S_k} \pm n_{tol}\sigma_k$ then the *i*+2 point is assigned as a vibrational band point and is thus rejected as being a baseline point.

Additionally, a tolerance scalar, referred to herein as n_{tol}, scales each value of σ_k and determines how tolerant the rejection criterion is for the i+2 point. Equation 1 shows the acceptance region when a constant n_{tol} used. Formally,

$$X_{i+2} \leq \overline{\Delta S}_k + n_{tol}\sigma_k \qquad \text{Equation 2.1}$$

where X_{i+2} is the measured absorbance, transmittance, or intensity of the i+2 minimum. When X_{i+2} is within this threshold region it is considered baseline.

At first appearance, Equation 1 seems to display a positively-bound confidence interval for the local population mean estimate ($\Delta\mu_k$) using the test statistic $\overline{\Delta S_k}$. In this study, the default specification for n_{tol} was 3.5. This conformed to the empirical optimal value for all real spectra tested. Additionally, when employing the present procedure to simulated spectra the performance of n_{tol} was fixed between 2.5 and 3.5 which usually made no appreciable difference to the final baseline correction result. Again, the details surrounding the evaluation of model parameters and their selection are presented below.

In general, the acceptance region varies as a simple function of the previous *n* second differences. However, for the first *n* ΔS_j values of the algorithm only, the acceptance region remains unchanged, implying that the first *n*+2 minima of the intermediate difference spectrum are indiscriminately included as baseline. Additionally, after the first *n*+2 segmental minima are surpassed, the statistical interpretation of the acceptance region is complicated because n_{tol} is modified using an additional scale parameter, λ_K. This additional scalar is applied to allow the algorithm to anticipate more abrupt changes in baseline. This twice-scaled σ_k is adaptive in the sense that it allows the algorithm to anticipate marked changes in baseline while still excluding points comprising Raman or IR bands.

The rationale governing the use of the scale parameter λ_K is calculated and explained as follows. First, $\overline{\Delta S_k}$ and σ_k are calculated from the previous differences, containing the *i*+1 through *i*-*n*+3 segmental minima, determined as baseline. Next, least-squares regression of the form

$$\hat{\beta}_{mk} = (\vec{x}_k{}^T \vec{x}_k)^{-1} \vec{x}_k{}^T \vec{y}_k \qquad \text{Equation 2.2}$$

is used to regress the mean-centered values of the past *n* confidence (or threshold) regions, $y_k = \overline{\Delta S}_k + n_{tol}\sigma_k$, onto the past *n* wavenumbers (x_k). This regression results in a linear prediction of the *n*-1 threshold region, which contains the *i*+1 minimum value. Next, an identical regression is performed in which the n^{th} threshold region, containing the *i*+2 minimum, is predicted. Note, the value of *m* in equation 2 corresponds to either 1 or 2 depending on whether regression includes the *i*+1 or *i*+2 point, respectively. The regression coefficients from each estimate are averaged and divided by $\hat{\beta}_{1k}$. The final parameter, defined as λ_k, is used to scale the tolerance parameter for the next threshold region. Formally,

$$\lambda_k = \frac{1}{2} + \frac{\hat{\beta}_{2k}}{2\hat{\beta}_{1k}} \qquad \text{Equation 2.3}$$

Division by $\hat{\beta}_{1k}$ is necessary to create a dimensionless λ_k that scales n_{tol} in a significant way. Therefore, when λ_k is employed, the final equation for the tolerance region is defined by

$$X_{i+1} \leq \overline{\Delta S}_k + \lambda_k n_{tol}\sigma_k. \qquad \text{Equation 2.4}$$

Equation 2.4 acts to either expand or contract the threshold region depending on the relative differences between $\hat{\beta}_{mk}$values. This has the effect of relaxing the assumption that local variations in ΔS_j are always small. For the first iteration, such variations are often large in regions with high curvature and/or substantial noise. The effect of averaging ensures that

neither regression coefficient dominates the value of λ_k. Division by $\hat{\beta}_{1k}$ in equation 3 ensures that the influences of $\hat{\beta}_{2k}$are smaller than $\hat{\beta}_{1k}$ by a factor of 2.

Once all baseline minima are identified, simple linear interpolation is employed and the resulting baseline is subtracted from the original spectrum. After the first identification and interpolation-subtraction is completed, the entire process is repeated no more than fifteen times. Allowing baseline subtraction to be performed fifteen times often results in a better true estimate of the baseline for reasons discussed below. Of the fifteen resulting spectra, the spectrum containing the lowest percentage of negative values is chosen automatically as the best spectrum unless the spectrum is known to contain negative-going features (*vide infra*). This effective stoppage criterion was deemed the most reasonable given the highly localized nature of the baseline correction procedure.

Aside from the standard algorithm, additional user-defined constraints were included and tested to determine if they improved baseline correction performance. For example, regardless of the outcome of the model, the first and last intensity values in a given spectrum were always included to ensure that linear interpolation did not produce spurious (i.e., undefined) values. During algorithm testing on real spectra, it was found that broad stretching regions showed deviations between baseline and band points that were gradual enough to be tolerated as baseline and thus inappropriately removed. Thus another modification involved requiring the user to specify the approximate location of the terminal points on broad stretching modes (e.g., -OH or –NH stretching modes).

During algorithm testing, it was found that employing an iterative procedure to baseline correction occasionally resulted in poorer performance for spectra containing bipolar bands such as some surface-enhanced infrared absorption (SEIRA) spectra that we tested. Thus, a user-specified constraint is toggled specifically for such spectra thereby terminating the iteration after one pass thus bypassing the standard stoppage criterion. Additional user-defined constraints were tested (including a handful of combinations of these constraints) to determine their effects on baseline correction. These constraints included:

1. Toggle (on/off) whether spectrum contains bipolar bands such as SEIRA spectra (i.e., SEIRA = 0/1)
2. Toggle adaptive thresholding (i.e., ADAPT = 0/1)

3. Specify the spectrum as a positive-going only spectrum (i.e., RAMAN = 0/1)
4. Toggle whether to incrementally step through the initial tolerance parameters (n_{toli}) as a function of iteration (n_{tolstp} = 0/1)
5. Specify whether the window size employed in the moving average is fixed or whether the algorithm incrementally steps through window sizes until a global minimum is identified (SW = 0/1)

Constraints 3 and 4 require more explanation. Upon completing the standard algorithm, the RAMAN option performs an additional and final minima search using the already-determined segmental minima as reference values. For example, if there exists another minimum between two predetermined segmental minima the RAMAN constraint activates a supplemental function to locate that minimum between those two values and designates that point as part of the final baseline estimate. Ideally, this implies that no band point has the possibility of retaining a negative value.

When constraint 4 is active, the initial tolerance value of 2.5 is specified and the maximum number of iterations set to ten. As the iteration loop proceeds, n_{toli} is reduced by 0.25 until a final value of zero is reached at the last pass. The rationale governing this constraint relates to the hypothetical possibility that the first pass of the algorithm will incorrectly assign a band point to baseline thereby diminishing the true character of the band. Essentially, this constraint acts as insurance against overcorrecting the baseline. Specifically, if n_{toli} is reduced at each pass, a misidentified band point that was accepted in the first pass might be rejected in the second or later passes because the tolerance criterion becomes stricter as the algorithm proceeds.

The baseline correction algorithm was tested on a total of twelve Raman and SEIRA spectra of varying spectral coverage and resolution. In the early days of FT-IR spectroscopy, when instruments often led to imperfect spectra, the late Tomas Hirschfeld recommended that users should keep a library of unusually bad spectra, that he called the “chamber of horrors”. The spectra that we used to test the baseline correction algorithm were mainly taken from such a collection that had been saved over the years by the Griffiths group. These included the Raman spectra of several explosives including trinitrotoluene (TNT) and Semtex measured on a dispersive spectrometer with a 785-nm laser. These spectra often had

a fluorescent background that could not be readily replicated by a low-order polynomial. A micro-Raman spectrum of electrospun poly(vinylpyrollidone)-titania composite fibers (PVP-TiO_2)[21, 22] was chosen as a proof-of-concept for the baseline correction procedure when applied to noisy micro-Raman spectra. In this case, the spectrum was obtained during an investigation for hyperspectral imaging of nanomaterials with structural features significantly smaller than the laser spot size (< 500 nm). The thickness of the polymeric shell of the ultra-fine fibers was of the order of 10 nm.

Baseline correction of SEIRA spectra measured under conditions where the metal nanoparticles were in electrical contact (percolated) was particularly challenging as many of the absorption bands of molecules at or near the metal surface are bipolar. Other spectra that contain bipolar bands include difference spectra, vibrational circular dichroism spectra and first derivative spectra. Since many automatic baseline correction routines make the explicit or implicit assumption that all bands are positive going, they would not be applicable to such spectra. The spectra that we used to investigate the feasibility of baseline correction in such cases were measured by attenuated total reflection (ATR) where the ZnSe internal reflection element (IRE) had been coated with a 10-nm layer of percolated silver nanoparticles.[23] These spectra included the ATR spectra of a monolayer of *p*-fluorothiophenol (PFTP) covered with a 1-mm thick layer of liquid methyl ethyl ketone and a thick film of liquid methanol.

In general, the Raman spectra were chosen because of the extremely variable nature of their baselines and the fact that several of them had a very low SNR, while the presence of strongly bipolar bands in the SEIRA spectra made them a particularly challenging test case. Upon baseline removal, all spectra were assessed for major baseline component removal and notable inclusion of artifacts due to the implementation of the procedure, as well as the ability to accept negative-going SEIRA bands. Real spectra discussed in the remainder of this paper will be used to emphasize the relative strengths and weaknesses of the present method. All spectra excluded from the discussion are shown as supporting information.

Further analysis was performed using simulated Raman and SEIRA spectra with three major goals in mind. Ultimately, we were interested in objectively assessing the efficacy of baseline correction given (1) a large domain of input scenarios (i.e., large range of parameter and constraint combinations) for the baseline correction algorithm, (2) a seemingly infinite

combination of spectral band shapes, widths, band positions, and band congestion one might encounter in practice, and (3) the effects of noise and nonlinearity on algorithm performance. Of course, a detailed understanding of each of these factors is beyond the scope of the present study. Although preliminary, we believe we have identified a means to probe the influence of each of these factors using some form of repeated random sampling technique. This technique will be described below.

As stated above, the present procedure has a huge array of potential inputs (e.g., window size, starting tolerance, etc.). Therefore, some methodology was required to determine the best combination of inputs for a handful of simulated spectral systems. In our case, we chose three systems mimicking a real (1) Raman spectrum, (2) SEIRA spectrum, and (3) a bipolar-only spectrum. Given the fair degree of uncertainty governing the outcome of baseline correction over a large input domain, a repeated sampling approach was employed. Specifically, a simulated spectrum was constructed by employing the use of well-understood, nonlinear, deterministic functions (e.g., sinusoids for baselines, Lorentzian profiles for Raman bands, etc.) where the behavior of these functions was governed by randomized inputs drawn from a normal probability distribution. After the noisy, baseline perturbed spectrum was constructed, a set of inputs was chosen and used with standard (i.e., unconstrained) baseline correction algorithm to correct the spectrum. The baseline corrected spectrum was then compared to the original baseline-free spectrum and the process of spectrum simulation, correction, and comparison repeated 500 times for the same input scenario (i.e., fixed set of constraints). The details of this process are presented below, beginning with spectrum simulation.

First, the total number of points in the simulated spectrum is set to 1200 with a channel size of 2.5 cm^{-1}. This results in a spectral range of 1-3000 cm^{-1} which, for our present purposes, is approximately on the order (element-wise) of the real Raman and SEIRA spectra tested in this study. Baseline construction begins by defining and summing five pseudo-random sinusoids the sum of which is then added to a cumulative error function. Specifically, a predefined amplitude and frequency for the sinusoids was specified at A_{init} = 100 arbitrary units and f_{init} =1/2500 cycles*$(cm^{-1})^{-1}$, respectively. A random number generator is employed to weight the frequency for every sinusoid (i.e., φ_j) making their behavior random about the predefined values. The amplitude of the baseline is given by a separate summation

of the randomly weighted fixed amplitude (i.e., φ_i). An additional random number generator is employed to generate the weights (θ_k) used to select which fraction of the baseline is comprised of sinusoidal or error function components. These weights are normalized so that their sum is equal to one. Therefore the final baseline takes the form

$$B = \theta_1\left[\left(\sum_{i=1}^{5}\varphi_i A_{init}\right)\sum_{j=1}^{5}\sin\left(2\pi\varphi_j f_{init}\hat{v}\right)\right] + \theta_2\left[\frac{1}{2}\mathrm{erfc}(\hat{v})\right] \qquad \text{Equation 2.5}$$

where B equals the baseline, θ_1, θ_2 represent the random weight fractions for the sinusoids and cumulative error function, the predefined amplitude and frequency are given by A_{init} and f_{init}, and the predefined amplitude and frequency are scaled by the random weights φ_i, φ_j, respectively.

For the sinusoids, the predefined amplitude was chosen to ensure that the magnitude of the baseline was large (on average) thereby simulating fairly horrible baselines. An f_{init} = 1/2500 cycles*$(cm^{-1})^{-1}$, or (1.2 cycles*$spectrum^{-1}$), was chosen to reduce the probability that any one-half period of a given sine was on the order of a Raman bandwidth but still resulted in challenging and highly nonlinear baselines. The cumulative error function was added to the sinusoids precisely to counter the possible generation of overtly high frequency baselines.

Upon baseline construction, a noise-free spectrum was constructed by specifying the number of bands. For all three systems, the number of bands chosen was (1) ten "Raman", (2) five "Raman" and five bipolar, and (3) ten bipolar bands. All Raman bands assumed a Lorentzian profile with bipolar bands assuming a profile generated using the 1st derivative of a Cauchy-Lorentz function. Again, random number generators were employed to weight and scale a predefined band amplitude, width, and position parameter for each Raman and bipolar band. The degree of band congestion was not controlled. A simulated instrument line-shape function was not convolved with the noisy spectrum. The full algorithm for the spectrum simulator showing all pertinent parameters and additional assumptions is presented as supporting information.

Finally, the root-mean-squared (RMS) noise (σ_{rms}) was specified as either one or ten arbitrary units and was used to scale a standard-normal, 1200-element random vector N~(0, σ_{rms}^2). The noise vector, noise-free spectrum and nonlinear baseline were all added to generate the raw spectrum to be baseline corrected.

The success of the baseline correction procedure was assessed by comparing the baseline-free noisy spectrum to the baseline corrected spectrum for a given system and input scenario. Specifically, calculating the root-mean-squared error (RMSE) between the baseline-free noisy spectrum and baseline corrected spectrum acted as the error statistic where a lower RMSE corresponded to a better overall fit. Because of the random nature of the spectrum simulator, baseline correction was performed 500 times for each input scenario resulting in 500 corresponding RMSE values. The RMSE was averaged thus allowing our group to determine which scenario performed the best, on average, for the Raman, SEIRA, and bipolar systems.

For the sake of brevity, selected results of all 48 input scenarios operating for a fixed window (FW) size ($n=5$) are presented in the supporting information section. The full results of 48 runs for the stepped-through window size (SW) is also shown in supporting information. The best and worst performing scenarios for each spectral system using the FW and SW algorithms will be presented below. All programming of the baseline correction procedure and spectrum simulator was performed in MATLAB (2007b, The MathWorks, Natick, MA).

Results and Discussion:

Baseline correction of real spectra

Figure 2.3 shows the raw spectrum (green), an estimate of the baseline for the first iteration (red), and final (blue) Raman spectrum of TNT when (n, n_{tol}) = (5, 3.5). A five-point window size (n = 5) was determined as the empirical best for all spectra and is thus n was held constant at 5 for this study. The Raman spectrum of TNT was a fairly simple test for the baseline subtraction procedure because of the high SNR as well as the gradually varying baseline. The corrected spectrum had only 4.4% of the total counts that exhibited negative values. This is slightly different from the baseline-corrected spectrum acquired after the first iteration. Specifically, Figure 2.4 displays how the sixth iteration was deemed optimal in the sense that the overall percentage of negative values was minimized and thus the spectrum chosen as the best overall.

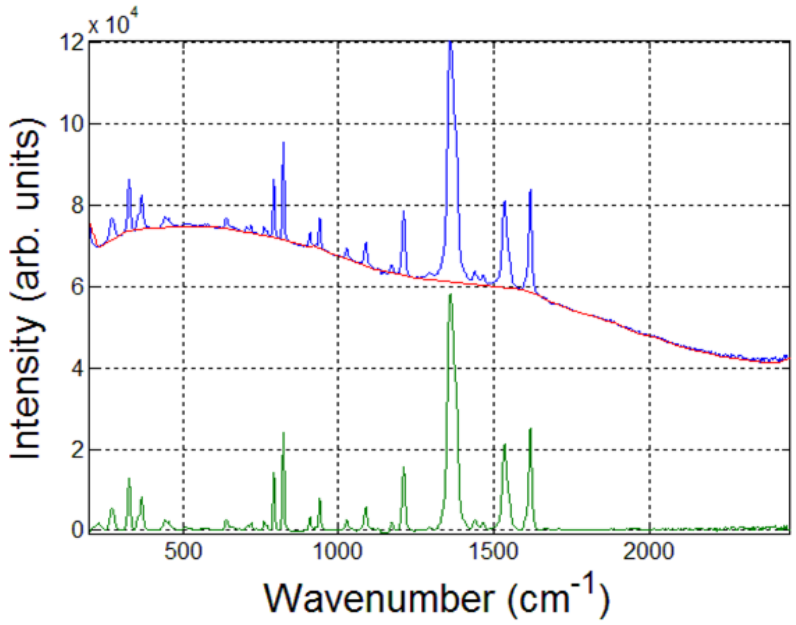


Figure 2.3: Baseline correction of the Raman spectrum of TNT. Baseline correction proceeded with a moving average window size of n=5 and an average value for the adaptive tolerance parameter n_{tol} equal to approximately 5.6. The initial spectrum is shown in green with the baseline estimate shown in red. The final spectrum is shown in blue. After six iterations, the algorithm identified the lowest percentage of all values as negative to equal to 4.4%.

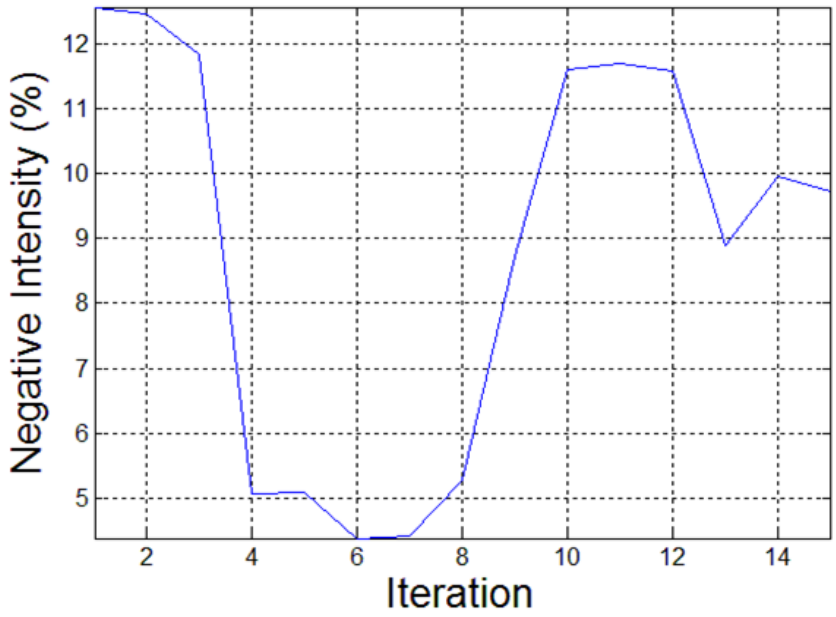


Figure 2.4: Percentage of negative values plotted against the number of iterations for the Raman spectrum of TNT. A minimum is reached at the sixth iteration.

Figure 2.5 displays the baseline-corrected Raman spectrum of Semtex. This spectrum poses a greater challenge when compared to the TNT spectrum shown in Figure 2.3 in that the relatively weak Raman spectrum is superimposed on an intense polynomial-like baseline. In spite of this, the method succeeded at correcting the baseline to within acceptable limits by the second iteration. Figure 2.6 shows an expanded plot of the final corrected spectrum. It can be seen that the algorithm failed to completely remove all negative points as shown at 1050 cm^{-1}, which is a positive feature of this algorithm as these points are clearly due to noise and not baseline.

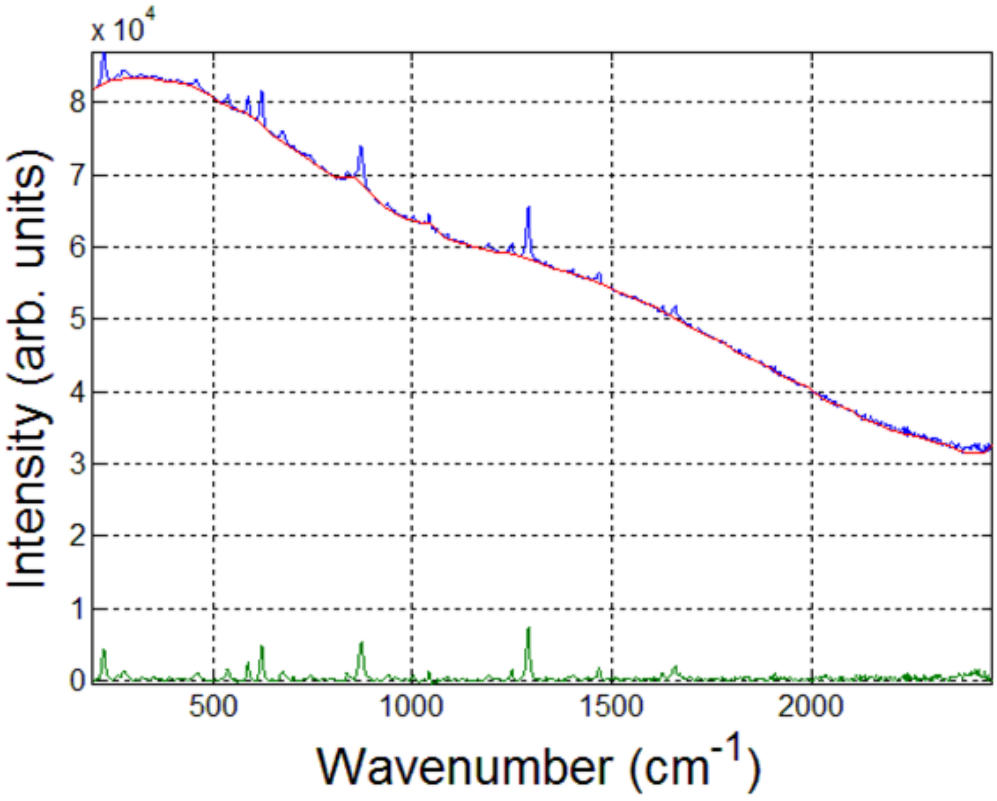


Figure 2.5: Baseline correction of the Raman spectrum of Semtex.

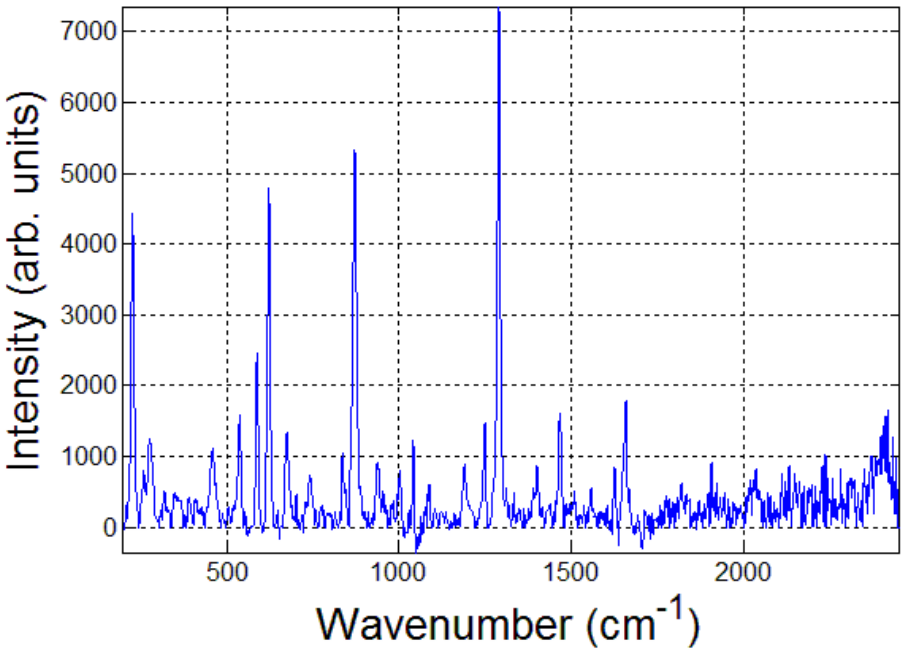


Figure 2.6: Final baseline-corrected spectrum of Semtex.

Notice that the noise in the final spectrum is always biased positive, i.e., it defies the assumption that instrument noise has zero mean. In this respect, the present procedure falls short of the strictest definition of baseline correction in the sense that all spectra, tested or hypothetical, are shifted upward by approximately one-half the peak-to-peak (p-p) noise level. A supplemental modification to the present work would involve employing an automated estimation of noise level and further correction of the vertical offset in the manner of Schultze et al.[24]

Successive iterations only fine-tune an estimate of the first iterated baseline and typically fail to remove any major baseline components missed in the first pass. At each successive pass, segmental minima are redefined but are still constrained in a manner identical to the previous pass with the exception of the adaptive tolerance scalar. For example, in Figure 2.6 the rejected and negatively-going noise-spike at 1050 cm^{-1} was passed over in the first iteration. Since successive iterations tolerate less deviant minima due to the improved constancy of the k^{th} moving average (i.e., flatter baseline and more consistent overall), the iteration step could proceed *ad infinitum* without any meaningful change in baseline correction. Of course, it is possible that the window size, starting tolerance level, or segment size could be modified to yield a different outcome for any spectrum tested.

However appealing, changing any one of those parameters eliminates the major advantage of automation for baseline correction.

The results of the baseline correction of PVP-TiO_2 nanowires are shown in Figure 2.7. Again, the present procedure assigns the negative going spike at 1520 cm^{-1} to noise and not to the baseline. Clearly for Raman spectroscopy, where all the bands are positive-going, the algorithm performs well, even though the SNR in the fingerprint region of the spectrum was very low.

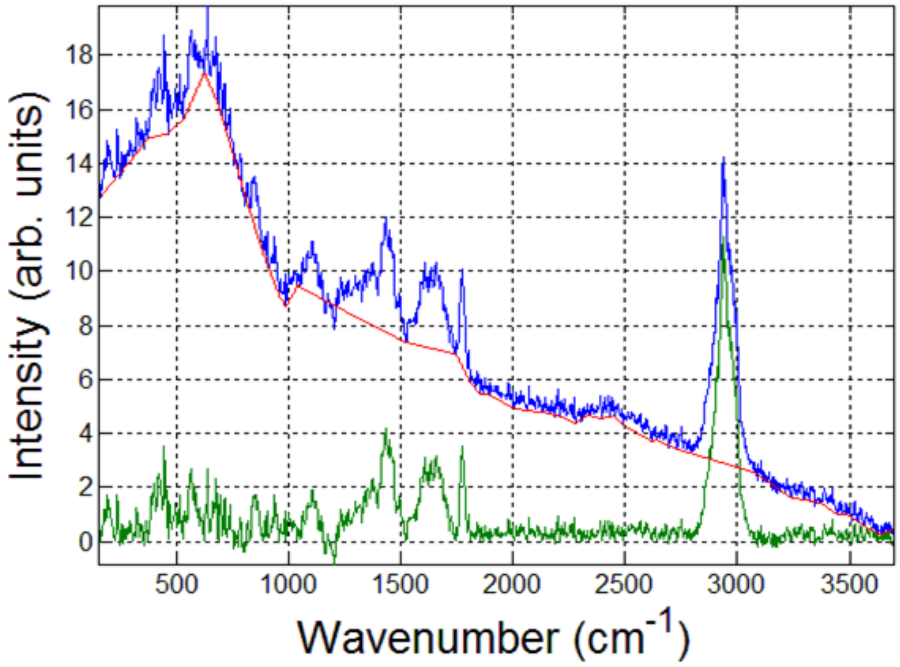


Figure 2.7: Raman spectrum of the original (green), estimated baseline (red), and corrected spectrum (blue) of PVP-TiO_2 fiber (dia. 500 nm).

This method of baseline correction could benefit by the inclusion of a command that specifies that all band points remain positive for Raman spectra. The inclusion of an additional segmental-minima search step, *ex post facto*, would require that the already defined baseline points act as the segment boundaries. These boundaries would then be used to identify negative going artifacts. This optional final step would be independent of the main algorithm thereby allowing its inclusion to be at the behest of the user. Allowing the predetermined baseline minima to define the minima search regions would enable the program to employ linear interpolation without an additional execution of the full algorithm.

The Raman spectra presented in Figures 2.6 and 2.7 would clearly benefit from this modification.

Such a step would not be included for spectra with bipolar bands such as certain SEIRA spectra, which we will now discuss. Figure 2.8 shows the prediction for the first iterated baseline, raw, and final SEIRA spectrum of PFTP chemisorbed on percolated silver nanoparticles and immersed in methyl ethyl ketone (MEK). It is noteworthy that both the PFTP and MEK bands are bipolar even though only the small fraction of the MEK within the penetration depth of the IRE is in contact with the silver coating. This figure demonstrates the efficacy of employing the present baseline correction technique to SEIRA spectra because the bipolar, dispersive bands are neither removed from the spectrum nor seriously distorted. In other words, the spectrum of this sample exemplifies the fact that the present procedure makes no assumptions about the positivity of the vibrational bands, but only whether they deviate from a locally-determined average value. Only the deviation between adjacent points is important for determining the acceptance or rejection of baseline points irrespective of magnitude and sign. The negative going band near 1250 cm^{-1} in Figure 2.8 again shows that relatively intense spikes may not be eliminated when bands are weak and trend in approximately the same direction as the dominant perturbing baseline. Thus considering the complexity of the highly congested, low-wavenumber region of the spectrum in these SEIRA spectra, the method performs exceptionally well.

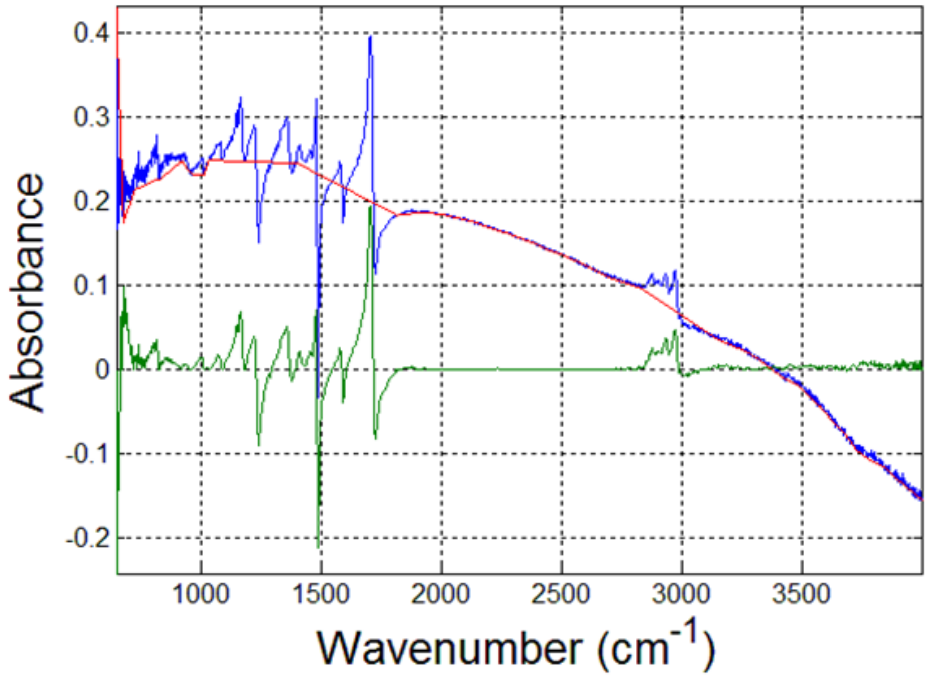


Figure 2.8: SEIRA spectrum of a monolayer of PFTP chemisorbed on percolated silver nanoparticles that had been deposited on a ZnSe IRE and immersed in liquid MEK. Even with automated selection of the "best" spectrum based on the lowest percentage of points exhibiting negative values, the algorithm still manages to include many of the dispersive bands in the spectrum. This shows that the procedure applies equal discrimination to positive- and negative-going bands.

Executing the algorithm in highly band congested regions (especially in the fingerprint region of the spectrum) often leads to better estimates of the final baseline. This is due to the historico-statistical nature of the technique. Specifically, the k^{th} standard deviation was small in regions where the baseline showed low to moderate curvature. This occurred because the n second-differences used to calculate σ_k showed small relative deviations in regions of gradually sloping baseline. Couched in statistical terms, regions of locally high curvature acted as outliers when lumped into a calculation with differences from previous regions of low curvature. The major implication was that if the standard deviation of the past n differences was small prior to encountering a region of inconsistently high curvature or any additional idiosyncrasy, the $i+2$ point was often incorrectly excluded. The adaptive threshold parameter, λ_k, was included precisely to counter the influence of such idiosyncrasies and abrupt changes in local curvature.

Figure 2.9 illustrates the trajectory-like character of the present baseline correction procedure. Although seemly inappropriate, employing the term "trajectory" invokes an

analogy between the present procedure and a dynamical process. This allows one to conceptualize the present procedure as exhibiting path-dependent behavior, i.e., baseline correction shows a path-dependence governed by the history of the past *n* second-differences. To elucidate, when the wavenumber scale of the spectrum is flipped and baseline correction is initiated from lowest to highest wavenumber, the i+2 point considers lower wavenumber points as the "past *n* second-differences." If the wavenumber scale of the process had been reversed (i.e., baseline correction starts in the high wavenumber region), the "past *n* second-differences" would correspond to minima at higher wavenumbers with respect to the present (i+2) point. Thus, whether the spectrum starts at high or low wavenumber prior to differencing has an effect on the baseline correction.

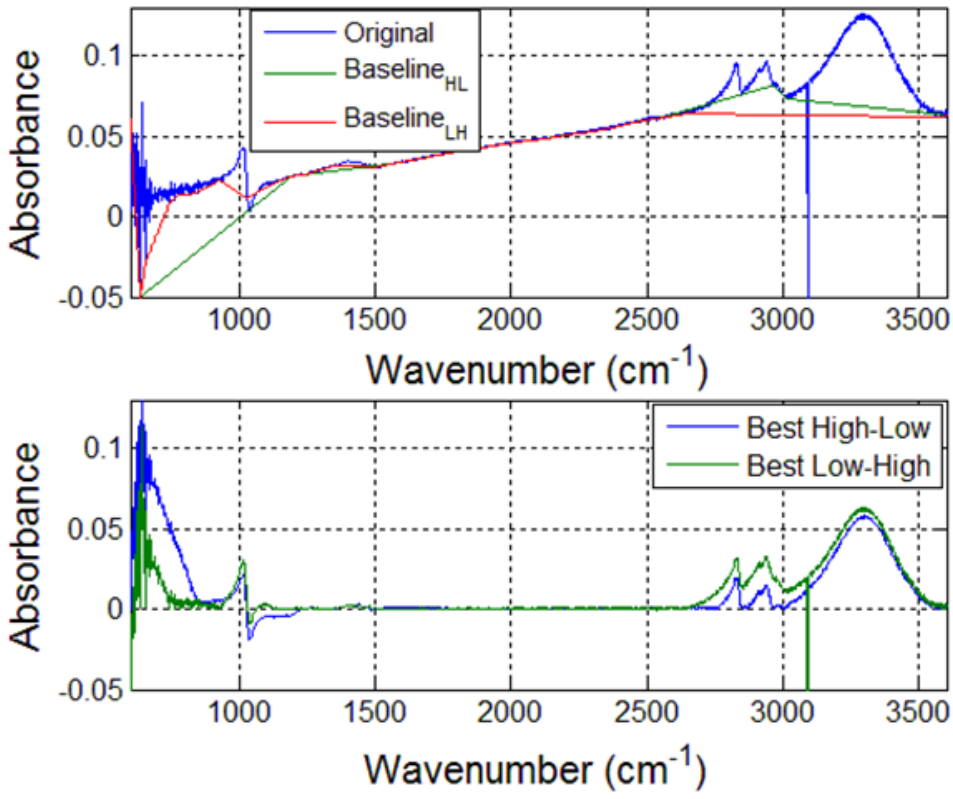


Figure 2.9: Comparison of baseline correction for the SEIRA spectrum of methanol when the algorithm is initiated at high wavenumber (green baseline) and from low wavenumber (red baseline). Note, the "Low-High" baseline corrected spectrum required no manual specification of the terminal endpoints of the broad OH stretching mode at 3250 cm^{-1}. The window size and original tolerance level were 5 and 3.5, respectively.

Figure 2.9 shows an example of a SEIRA spectrum of liquid methanol in contact with percolated silver nanoparticles when differencing and initiation were performed from the

highest to the lowest wavenumber and then from the lowest to the highest wavenumber. As noted above, the baseline-corrected spectra are slightly different whether the procedure is started at high or low wavenumber. This is particularly important when the spectrum shows a broad OH stretching mode as shown in Figure 2.9. Generally, baseline correction was the most effective when the algorithm was initiated from the lowest wavenumber and so this was set as the default orientation.

The dependence of baseline correction on the starting wavenumber can be exploited to the benefit of the user. Figure 2.9 shows that, via a simple modification, the user is given the option to identify a point in the middle of the spectrum where baseline correction is performed from the low-wavenumber end of the spectrum to this point and from this point to the high wavenumber end of the spectrum. The two baseline corrected spectra are then sutured together at this user-defined point. An example of this procedure is shown in Figure 2.10.

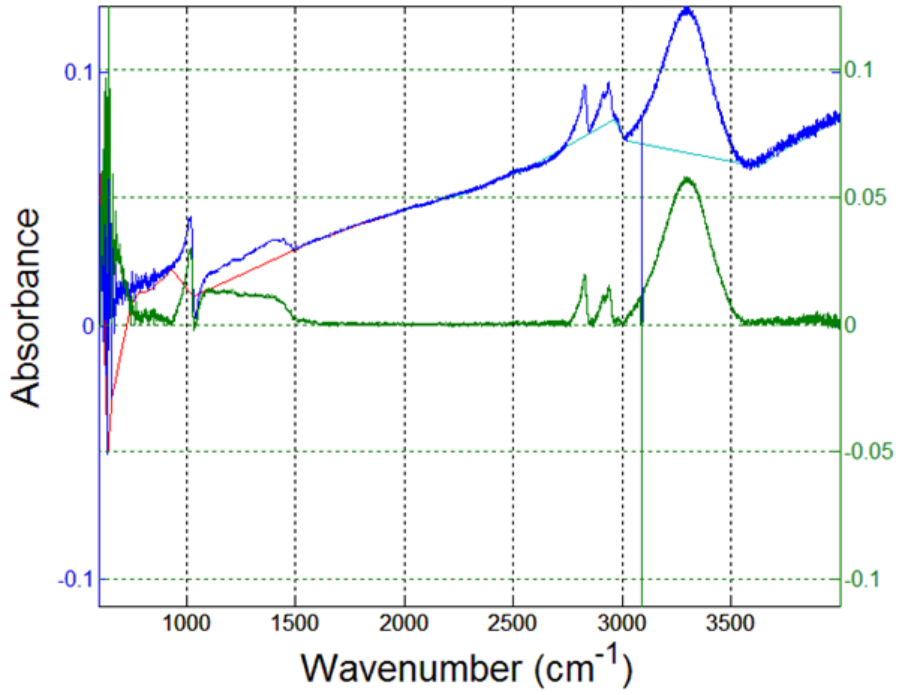


Figure 2.10: Effect of suturing the low-high and high-low baseline estimates at a user-defined value. In this example, the value was selected arbitrarily as 1957 cm^{-1}. Further semi-automation of the method allows the correction procedure to incorporate the best of both orientations into the final baseline corrected spectrum.

A final, but no less important, quality of the present baseline correction procedure is also illustrated in Figure 2.10, where it can be seen that the large negative going noise-spike

at ~3100 cm^{-1} is completely ignored by the baseline correction procedure. This behavior is consistent with rejection of negative-going noise shown in Figures 2.6 and 2.7. Of course, this appropriate rejection has entirely to do with the fact that baseline points are accepted or rejected based on whether they reside inside or outside the threshold boundary, irrespective of sign. Intuitively, if this boundary were to accept the large negative going noise spike in Figures 2.9 and 2.10, it would easily accept minima confined to positive going regions of the same magnitude making baseline correction completely nonsensical.

Baseline correction of simulated spectra

Figure 2.11A shows an example of a bipolar spectrum with the superimposed baseline estimate acquired from the FW algorithm with the "SEIRA" input constraint active. Judging by the results of the baseline correction shown in the corresponding Figure (2.11B), baseline correction was highly successful for this spectrum. This is also indicated by the low RMSE value. This example was taken from a set of 500 randomly generated and corrected spectra for this particular input scenario. In fact, Figure 2.12 shows a histogram of RMSE for this particular simulated system and input scenario. The error is skewed leftward toward zero and appears to take the form of a Weibull distribution. In fact, for every simulated system and scenario tested, this error distribution was observed. Although interesting, this limited our ability to pinpoint a direct relationship between the amplitude and frequency components of the simulated baseline and the final baseline correction outcome using significance tests based on normal distribution theory.

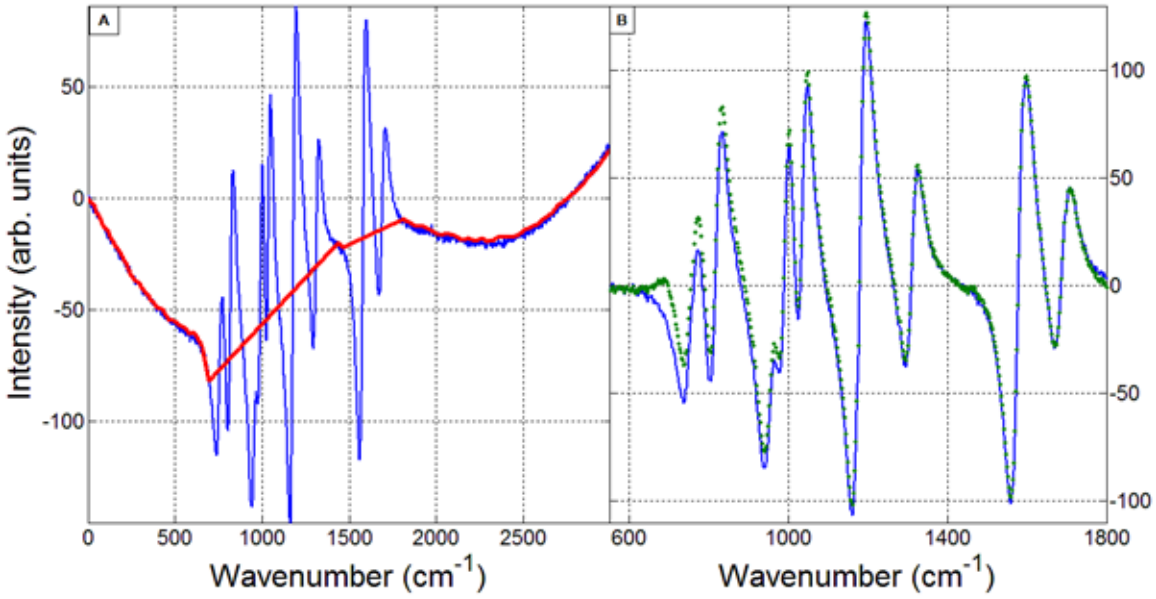


Figure 2.11: Baseline correction of a simulated bipolar band only spectrum using the FW algorithm with the SEIRA constraint active. The raw spectrum (LHS-blue) and superimposed baseline (LHS-red) are juxtaposed with the noisy baseline free spectrum (RHS-blue) and baseline corrected spectrum (RHS-green).

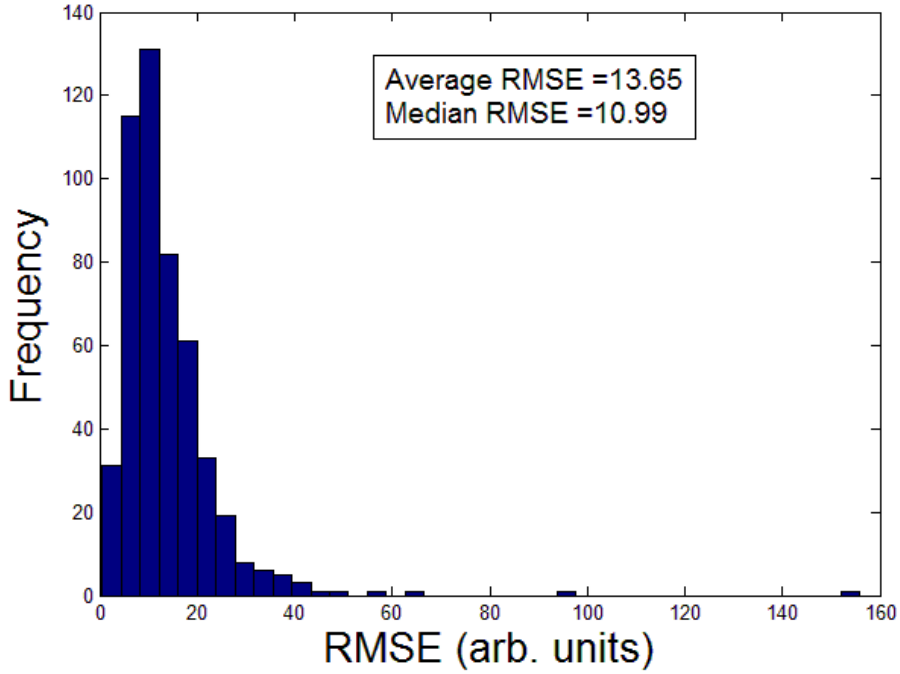


Figure 2.12: Histogram of the RMSE for the 500 simulated bipolar-only spectra. There is a clear leftward skew to the error distribution. This distribution shape was consistent across all system and input scenarios tested.

Tables 2.1 and 2.2 show the six best and six worst results of employing the present baseline correction procedure to forty-eight scenarios for the fixed-window (FW) and stepped-through window (SW) algorithms. As a reminder, the term "best" represents the baseline correction procedure that results in the lowest average RMSE for a given system. In the interest of space, some general trends will be noted.

Table 2.1- Tabulated results of the best (B) and worst (W) results of employing baseline correction to five-hundred simulated spectra for the fixed-window (FW) baseline correction algorithm. Column 1 shows the system characteristics from which each baseline correction was applied. The (#, #, #, B/W) notation signifies (1) the number of positive-going and (2) negative-going bands present in the simulated spectrum, (3) the RMS noise, and (4) whether this was the best or worst scenario in terms of lowest $RMSE_{AVG}$ as a function of parameter/constraints. Other columns show which constraints were active (1) or inactive (0), the average SNR of the simulated spectra, the RMSE statistics, and the average run-time in terms of spectra corrected per second.

Name	n_{tolstp}	Raman	SEIRA	Adapt	SNR_{AVG}	$RMSE_{AVG}$	$RMSE_{MED}$	t_{run}(spec/s)
$FW_{500}^{(10,0,1,B)}$	0	1	0	0	291.3	11.9	9.3	1.9
$FW_{500}^{(10,0,10,B)}$	0	1	0	0	30.5	22.5	20.9	1.9
$FW_{500}^{(5,5,1,B)}$	0	0	0	0	215.8	13.2	11.8	2.0
$FW_{500}^{(5,5,10,B)}$	1	0	0	0	22.5	20.5	19.4	2.0
$FW_{500}^{(0,10,1,B)}$	0	1	1	0	98.6	14.2	11.8	14.7
$FW_{500}^{(0,10,10,B)}$	1	1	1	0	11.3	19.8	19.3	14.3
$FW_{500}^{(10,0,1.W)}$	0	0	0	1	293.4	16.6	13.7	1.9
$FW_{500}^{(10,0,10,W)}$	1	0	0	1	29.9	27.2	26.2	2.0
$FW_{500}^{(5,5,1,W)}$	1	1	0	0	215.9	30.7	27.9	2.0
$FW_{500}^{(5,5,10,W)}$	1	1	0	0	22.7	31.3	29.1	2.0
$FW_{500}^{(0,10,1,W)}$	1	1	1	1	95.1	15.9	13.3	13.2
$FW_{500}^{(0,10,10,W)}$	1	0	1	0	11.5	21.2	20.5	14.7

For the FW algorithm, when the "SEIRA" constraint was active, baseline correction performed the best (row 5 & 6) and the worst (row 11 & 12) for bipolar-only spectra. This might seem paradoxical at first. However, this illustrates the power of using a repeated sampling approach to optimize an algorithm with a large input domain. Specifically, notice that model performance was optimal when the "RAMAN" *and* "SEIRA" constraint were active and the worst when *all constraints were active* or the "RAMAN" constraint was inactive *and* "SEIRA" active for the bipolar-only system. Thus, our brute force approach to

optimization appears essential when the input domain of a baseline correction algorithm is large and the standard algorithm highly localized in nature. Note, in practice it is not entirely reasonable to encounter bipolar-only spectra. However, these results illustrate the potential for employing this method to baseline afflicted, first derivative spectra. Note also that the differences between the best and worst SNR for a given spectrum was always very small, suggesting that a non-optimal choice of starting conditions does not affect the final result too seriously.

In general, the performance of adaptive tolerances was often mediocre as signified by their relative absence from Table 2.1. Therefore, it is entirely likely that the proposed adaptive thresholding scheme is unnecessary or that it introduces artifacts into the final baseline estimate. In fact, for the Raman and simulated SEIRA spectra, a less constrained model (i.e., more inactive constraints) showed the best performance for the FW algorithm. Also, for positive-going and SEIRA spectra, the average RMSE generally doubled when the average SNR was decreased by a factor of ten. The best and worst performance for the bipolar spectra was about the same, indicating that constraining baseline correction prior to algorithm execution affected the model very little for bipolar-only spectra.

The FW algorithm generally performed better than the SW for all scenarios but occasionally under different constraints. In Tables 2.1 and 2.2, grey highlights were used to show which constraint combinations were common (for better or worse) *between* the FW and SW algorithms. Simply, those constraints that performed best in the FW algorithm usually performed better, or nearly as well, in the SW algorithm and vice versa. More importantly, the SW algorithm (not shown) identified that the window size employed the most during algorithm operation was $n = 4$ or 5. This helps justify the use of a five-point fixed window size for all real Raman and SEIRA spectra tested. The starkest difference between the FW and SW algorithm related to the algorithm run time (column 9). The FW algorithm performed 3-5 times faster than the SW for the best and worst performers. A longer run-time for the SW algorithm resulted from forcing the algorithm to repeat the standard baseline iteration loop up to nine times.

Table 2.2- Tabulated results of the best (B) and worst (W) results of employing baseline correction to 500 simulated spectra for the stepped-through window (SW) baseline correction algorithm. All columns in this table follow the same reasoning as Table 1. The SW algorithm performed worse than the FW algorithm as indicted by a higher $RMSE_{AVG}$ for each "best" or "worst" input scenario. This might indict that using the "percentage of negative values" as a stoppage criterion for the algorithm is not an optimal means of assessing when baseline correction should terminate, i.e., the SW probably overestimates the number of minima associated with the baseline and incorrectly includes band points in the final baseline estimate.

Name	n_{tolstp}	Raman	SEIRA	Adapt	SNR_{AVG}	$RMSE_{AVG}$	$RMSE_{MED}$	$t_{run}(spec*s^{-1})$
$SW_{500}^{(10,0,1,B)}$	1	1	0	0	291.3	12.2	9.0	0.6
$SW_{500}^{(10,0,10,B)}$	0	1	0	0	30.0	22.9	21.2	0.6
$SW_{500}^{(5,5,1,B)}$	1	0	0	0	210.8	13.3	11.0	0.6
$SW_{500}^{(5,5,10,B)}$	1	0	0	0	22.7	20.7	19.6	0.6
$SW_{500}^{(0,10,1,B)}$	0	0	1	0	98.3	17.3	13.8	4.5
$SW_{500}^{(0,10,10,B)}$	0	0	1	1	11.3	23.0	21.7	5.2
$SW_{500}^{(10,0,1,W)}$	1	0	0	1	289.2	20.0	17.2	0.7
$SW_{500}^{(10,0,10,W)}$	1	0	0	1	29.9	28.3	26.6	0.6
$SW_{500}^{(5,5,1,W)}$	1	1	0	0	215.5	33.7	29.2	0.6
$SW_{500}^{(5,5,10,W)}$	1	1	0	0	22.4	32.0	28.8	0.6
$SW_{500}^{(0,10,1,W)}$	0	1	1	1	96.8	29.2	26.2	5.1
$SW_{500}^{(0,10,10,W)}$	0	1	1	0	11.5	28.2	26.7	4.1

Understanding the average performance of a baseline correction routine is important. However, it is always pertinent to have some indication of the sources of error between the original and baseline corrected spectrum. Figure 2.13 aims to illustrate where the error contributions arise when using the present baseline correction routine on a single Raman and bipolar band when adaptive tolerances and "RAMAN" and "SIERA" constraints are active, respectively. Note, each band is confined to the center of the spectral range where the baseline was constructed using a single sinusoid with its frequency fixed at f_{init}. Figure 2.13 illustrates that one major source of baseline estimation error comes from the distortion of the Raman and bipolar bands corresponding to the removal of the wings that are characteristic of Lorentzian line shapes. Indeed the shape after baseline correction is closer to that of a Gaussian than a Lorentzian band. Although the role of band congestion as an error source was not evaluated, it is probable that its effect is minimal relative to this form of band distortion.

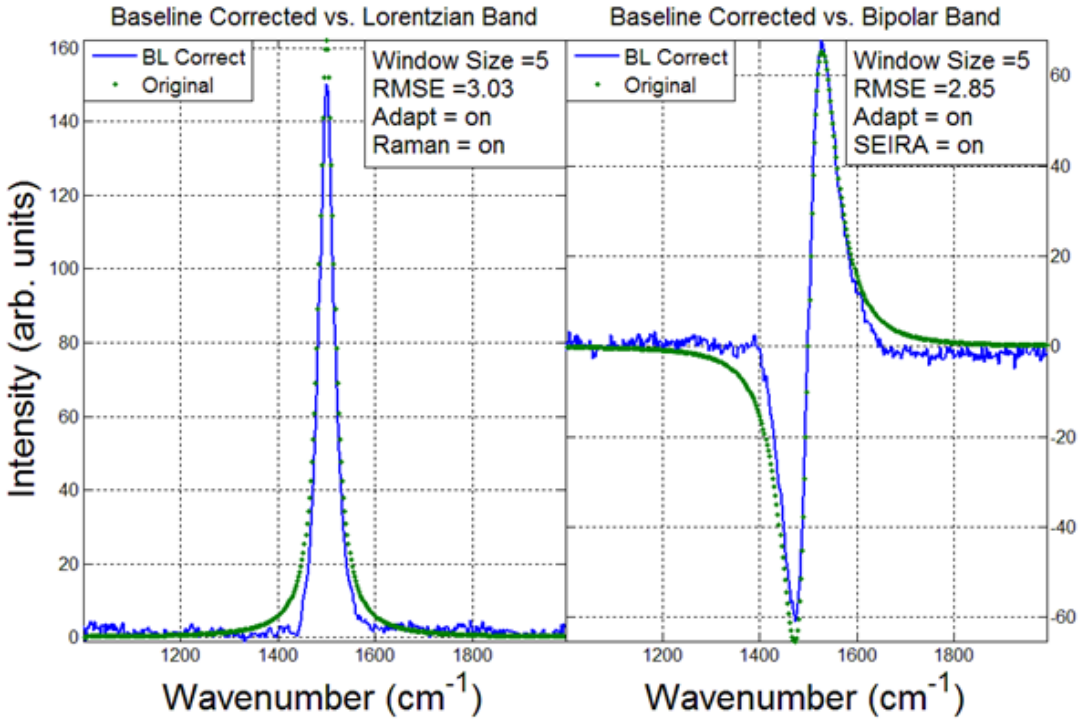


Figure 2.13: Comparison of the original (green) and baseline corrected spectrum (blue) for a Raman and bipolar band. The major sources of error appear to come from the baseline correction routine truncating the heavy tails of each band as well as from the failure to remove the very low frequency, small amplitude components of the baseline.

Figure 2.13 also suggests that a major source of error comes from the estimated baseline itself, i.e., there appears to be large frequency, small amplitude components present in the estimated baseline. These components are inadvertently introduced into the final baseline estimate by the minima themselves. Simply, the minima chosen from any noisy baseline will always correspond to the largest negative going noise spike as long as it is within the local tolerance region. This demonstrates that employing a symmetric, small window smoothing filter (*e.g.*, Savitsky-Golay filter) prior to minima identification will have a positive influence on baseline correction. Fortunately, the repeated sampling approach presented in this paper offers a means to conveniently identify the optimal parameters to use in the smoothing filter.

Conclusions:

A new technique for the semi-automated baseline correction of Raman and SEIRA spectra has been presented. By employing a combination of spectral segmenting, minima searching, differencing, and statistics-based band discrimination, the technique proved successful at removing the major components of the perturbing baseline found in all real spectra tested. The method does not require any smoothing or noise reduction prior to application of the algorithm. The only major requirement is that users input the approximate ($\pm$ 50 cm^{-1}) location of terminal band points for broad –XH stretching modes prior to execution.

Simulated spectra were generated and baselines corrected for three spectral systems and ninety-six input scenarios using a repeated sampling approach. As a proof-of-concept, this technique suggests that for baseline correction algorithms employing a large number of parameters and constraints, a theoretical optimum can be identified using a brute force, repeated sampling approach. Future studies should focus on correlating baseline correction performance directly to the SNR and the properties of all deterministic functions used to simulate each baseline.

Acknowledgements

We thank Dr. Husheng Yang for supplying most of the Raman and SEIRA spectra used in this study with the exception of the ultra-fine titania fibers. We also thank Weston Corporon for preparing the fibers. This work was made possible in part through funding from the National Science Foundation (award DMR-0619310), the US Department of Agriculture (award 2009-34479-19833), and the National Institute of Food and Agriculture (award 2010-34479-20715).

References

(1) G. Schulze, A. Jirasek, M.M.L. Yu, A. Lim, R.F.B. Turner, M.W. Blades. “Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation”. Appl. Spectrosc. 2005. 59(5): 545-574.

(2) C. Rowlands, S. Elliott. “Automated algorithm for baseline subtraction in spectra”. J. Raman Spectrosc. 2011. 42(3): 363–369.

(3) L. Shao, P.R. Griffiths. “Automatic Baseline Correction by Wavelet Transform for Quantitative Open-Path Fourier Transform Infrared Spectroscopy”. Environ. Sci. Technol. 2007. 41(20): 7054-7059.

(4) C.D. Brown, L. Vega-Montoto, P.D. Wentzell. “Derivative Preprocessing and Optimal Corrections for Baseline Drift in Multivariate Calibration”. Appl. Spectrosc. 2000. 54(7): 1055-1068.

(5) L. Keselbrener, M. Keselbrener, S. Akselrod. “Nonlinear high pass filter for R-wave detection in ECG signal”. Med. Eng. Phys. 1997. 19(5): 481-484.

(6) P.A. Mosier-Boss, S.H. Lieberman, R. Newbery. “Fluorescence Rejection in Raman Spectroscopy by Shifted-Spectra, Edge Detection, and FFT Filtering Techniques”. Appl. Spectrosc. 1995. 49(5): 630-638.

(7) E.J. Hasenoehrl, J.H. Perkins, P.R. Griffiths. “Rapid Functional Group Characterization of Gas Chromatography/Fourier Transform Infrared Spectra by a Principal Components Analysis Based Expert System”. Anal. Chem. 1992. 64(7): 705-710.

(8) D.M. Lewis, P.C. Chatwin. “The Treatment of Atmospheric Dispersion Data in the presence of Noise and Baseline Drift”. Boundary-Layer Meteorol. 1995. 72(1-2): 53-85.

(9) R.J. Tervo, T.J. Kennett, W.V. Prestwich. “An Automated Background Estimation Procedure for Gamma Ray Spectra”. Nucl. Instrum. Methods. 1983. 216(1-2): 205-218.

(10) R.P. Goehner. “Background Subtract Subroutine for Spectral Data”. Anal. Chem. 1978. 50(8): 1223-1225.

(11) J. Liu, J.L. Koenig. “A New Baseline Correction Algorithm Using Objective Criteria”. Appl. Spectrosc. 1987. 41(3): 447-449.

(12) C.A. Lieber, A. Mahadevan-Jansen. “Automated Method for Subtraction of Fluorescence from Biological Raman Spectra”. Appl. Spectrosc. 2003. 57(11): 1363-1367.

(13) T. Iwata, J. Koshoubu. “New Method to Eliminate the Background Noise from a Line Spectrum”. Appl. Spectrosc. 1994. 48(12): 1453-1456.

(14) G. Balcerowska, R. Siuda. “Inelastic background subtraction from a set of angle-dependent XPS spectra using PCA and polynomial approximation”. Vacuum. 1999. 54(1): 195-199.

(15) W. Dietrich, C.H. Rudel, M. Neumann. “Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra”. J. Magn. Reson. 1991. 91(1): 1-11.

(16) M.S. Friedrichs. “A Model-free Algorithm for the Removal of Baseline artifacts”. J. Biomol. NMR. 1995. 5(2): 147-153.

(17) M. Pirzer, K.Sawatski. Method and device for correcting a spectrum. US Patent 7359815. Filed 2006. Issued 2008.

(18) S. Wartewig. IR and Raman Spectroscopy. Weinheim, Germany: Wiley-VCH GmbH, 2003. Pp 77-79.

(19) T.V. Karstang, O.M. Kvalheim. “Multivariate Prediction and Background Correction Using Local Modeling and Derivative Spectroscopy”. Anal. Chem. 1991. 63(8): 767-772.

(20) C. Chatfield. The Analysis of Time Series: An Introduction. Boca Raton, Florida: Chapman & Hall/CRC, 2004. 6th ed.

(21) J.M.F. Jabal, L. McGarry, A. Sobczyk, D.E. Aston. “Substrate Effects on the Wettability of Electrospun Titania−Poly(vinylpyrrolidone) Fiber Mats”. Langmuir 2010. 26(16): 13550-13555.

(22) J.M.F. Jabal, L. McGarry, A. Sobczyk, D.E. Aston. “Wettability of Electrospun Poly(vinylpyrrolidone)−Titania Fiber Mats on Glass and ITO Substrates in Aqueous Media”. ACS Appl. Mater. Interfaces. 2009. 1(10): 2325-2331.

(23) D.A. Heaps, P.R. Griffiths. “Band shapes in the infrared spectra of thin organic films on metal nanoparticles”. Vib. Spectrosc. 2006. 42(1): 45-50.

(24) H.G. Schulze, M.M.L. Yu, C.J.Addison, M.W. Blades, R.F.B. Turner. “Automated Estimation of White Gaussian Noise Level in a Spectrum With or Without Spike Noise Using a Spectral Shifting Technique”. Appl. Spectrosc. 2006. 60(7): 820-825.

Chapter 3. Automatic baseline correction of vibrational circular dichroism spectra

Reproduced with permission (Appendix A): Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, Applied Spectroscopy, 2013, **67**(10): 1117-1126.

Abstract

A three phase, computational method for the baseline correction of vibrational circular dichroism (VCD) spectra has been proposed. In the first phase, the raw spectrum is subdivided into *m*-segments (or regions) resulting in *p* rough estimates of the baseline. A second phase uses gradient characteristics to discriminate between baseline and band response for each baseline, in turn. In the final phase all baselines are interrogated simultaneously by assigning the median estimate of each differential response's distribution to the true baseline. Using VCD spectra of (R)-camphor as test cases, this work demonstrated that the accurate removal of baseline components is readily achievable with minimal user intervention. Baseline correction also demonstrated flexibility in that prior information, such as the symmetry of a baseline-free VCD spectrum, is readily used during the correction protocol. Although three adjustable parameters are present in the base algorithm, optimal performance and full automation were attainable following the use of analysis of variance (ANOVA) to analyze simulated bipolar spectra. These ANOVAs suggested that band point discrimination could be discarded and the remaining two default parameters adopted.

Keywords: Automatic baseline correction; VCD spectrum; Vibrational Circular Dichroism; Model-free correction

Introduction

The most popular approach to measure the vibrational optical activity of solution-state chiral molecules is through infrared vibrational circular dichroism (VCD).[1,2] For Fourier transform (FT)-VCD measurements,[2,3] a plane-polarized beam is passed into a photoelastic modulator that alternately generates a left and right circularly-polarized IR beam, with a modulation frequency of typically 25-50 kHz, prior to passing through a solution containing a target chiral molecule. The difference in the absorption of left and right circularly-polarized

radiation, ΔA, is extracted using one or more lock-in amplifiers. Stereo-specific differential absorbance is observed in a VCD spectrum where the strength of each band is contingent upon the combined linear and circular transition moments. The signal observed in a VCD spectrum is typically bipolar about the baseline (ideally zero). Enantiomers are distinguishable by the opposite orientation of corresponding bands. When coupled with ab initio density functional theory calculations [4-8], calculated and measured VCD spectra allow a chiral compound's absolute configuration (AC) to be determined.[1, 2, 9]

The determination of AC is of paramount importance for pharmaceutical research and development where enantiomers may have very different pharmacological properties.[10,11,12] Following AC determination at a given synthesis step, there is usually a need to maximize enantiomeric excess (EE) where the accurate measurement of EE is critical.[7,13-15] Considering the plausible, widespread implementation of VCD spectroscopy to real-time, kinetic monitoring of EE,[16,17] the rapid automated removal of baselines prior to calibration is a pressing concern.[18]

Vibrational spectra are usually baseline corrected by a user who selects wavenumber ranges for which the sample is observably devoid of spectral features and connects them with a best-fit polynomial. The resulting function is subsequently subtracted from the spectrum yielding the corrected data set. This procedure is often time-consuming, tedious, and prone to spectral artifact generation.[19,20] For VCD spectra, baseline point selection is often problematic where one of three procedures requires the availability and use of experimental background spectra.[9] The first and most effective method requires access to the opposite enantiomer's VCD spectrum, which is then subtracted from the sample spectrum and the difference is divided by two. This approach removes the predominant baseline and limits artifact production. The second method requires subtracting the spectrum of the racemic mixture from the VCD spectrum, which generally increases noise. The final approach involves subtracting the solvent's VCD spectrum from the sample spectrum. Although solvents are achiral, their spectra can contribute substantial artifacts and noise to the final result.

Recently, we developed a method of semi-automated baseline correction for Raman and infrared spectra and showed the routine to be applicable for surface-enhanced infrared absorption spectra containing bipolar bands.[21] Our approach employed the use of simple

forward differences followed by a locally-modeled minima search and statistical discrimination routine. Notably, the method demonstrated versatility in isolating baseline components when either unidirectional or bipolar bands were present.

This paper describes a modified baseline correction routine that overcomes the sole reliance on spatially-constrained (i.e., wavenumber-dependent) discrimination of the bands from the baseline. Consequently, this new routine results in smoother baseline estimates, reduced artifact generation due to baseline correction, and a smaller array of input parameters. Considering the character of VCD band-shapes, our modified and purely computational routine offers an alternative to standard experimental approaches for baseline correction.

Experimental

Two sets of VCD spectra of 0.90 M solutions of (R)- and (S)-camphor in CCl_4, and a spectrum of the solvent background were supplied by Jordan Nafie of BioTools, Inc. The solutions were held in a 100-μm path length BaF_2 cell. Spectra were acquired with a Chiral*IR* (Biotools, Inc., Jupiter, FL) spectrometer at a nominal resolution of 4 cm^{-1}. The first set of spectra contained a low level of baseline interference (herein referred to as "better baseline") while the second set of spectra contained moderate baseline interference ("worse baseline"). The availability of the enantiomer and solvent spectra facilitated the direct comparison of our computed baselines to experimental VCD spectral correction.

In addition to the experimental VCD spectra, simulated bipolar spectra were randomly constructed using our simulator.[21] Noise-free spectra were generated by drawing weights randomly from a normal distribution and applied to the position, width, and amplitude of each band. Bands were generated using the first derivative of a Cauchy-Lorentz function with each band's half width at half maximum, γ, and amplitude, ΔA, constrained to $20 < \gamma < 200$ cm^{-1} and $10^{-4} < \Delta A < 10^{-3}$, respectively. To simulate the conditions of our experimental VCD spectra, bands in the synthetic bipolar spectra were plotted over the range of 800-1800 cm^{-1} even though the entire spectral range was interrogated by our routine (0-7898 cm^{-1}). Experimental spectra were also corrected using every channel available to the algorithm.

In an attempt to randomly simulate poor, but still realistic, baselines, frequency information from the "worse baseline" (R)-camphor spectrum was acquired by transforming the raw spectrum to the frequency domain using the fast Fourier transform. The lowest 0.5% of frequencies were taken from the resulting single-sided amplitude spectrum and used to generate sinusoidal baseline components with randomized phase information. Formally,

$$BL = \sum_{i=1}^{n} A_i \sin(2\pi f_i + \varphi_i) \qquad \text{Equation 3.1}$$

where A_i is the amplitude for the *i*th frequency from the "worse baseline" (R)-camphor spectrum, f_i is the corresponding frequency, and φ_i is the phase of the *i*th sinusoid drawn randomly from the standard normal distribution, varying between $-\pi/2$ and $+\pi/2$. In order to suppress the unrealistic generation of wildly fluctuating sinusoidal components, each amplitude was exponentially weighted and scaled. Finally, the raw spectrum was comprised of the pristine spectrum (*S*), baseline, and contaminated with Gaussian-random noise (n_{rms}) with a relative level of $||n_{rms}||_2/||S||_2 = 10^{-1.22}$

Methods:

The present baseline correction technique mirrors our previous approach subject to a few key differences. Phase I of baseline correction proceeds with a regional point-search operation. Specifically, the raw VCD spectrum ($\vec{\boldsymbol{x}}_{\boldsymbol{VCD}}$), which contains *q* spectral channels, is divided into *m* equally-spaced regions requiring *m*+1 boundary points. Differential response (ΔA) in each region of $\vec{\boldsymbol{x}}_{\boldsymbol{VCD}}$ is ranked and sorted from lowest to highest and their channel indices placed in a *m*-by-*p* matrix, C. Under this notation, *p* points are contained in each region. For the C matrix, the element $\{c_{ij}\}$ corresponds to the channel index of the ranked and sorted *j*th differential response in the *i*th bounded region. For example, $\{c_{31}\}$ is the element in C containing the channel index of the regional minimum ($j = 1$) in the third region ($i = 3$) while $\{c_{mp}\}$ is the channel index of the regional ΔA maximum ($j = m$) in the terminal region ($i = p$). It follows that any column vector in C, or $\overrightarrow{\boldsymbol{c}_{\boldsymbol{j}}}$, is a unique set of indices spanning all *m* regions of the VCD spectrum. Hence, the first column vector, $\overrightarrow{\boldsymbol{c}_{\boldsymbol{1}}}$, is the *m*-length set of indices for all regional response minima and the *p*th column vector ($\overrightarrow{\boldsymbol{c}_{\boldsymbol{p}}}$) is the *m*-length set of indices for all regional maxima. Figure 3.1 offers a brief synopsis of the ranking and sorting process that ends phase I of baseline correction.

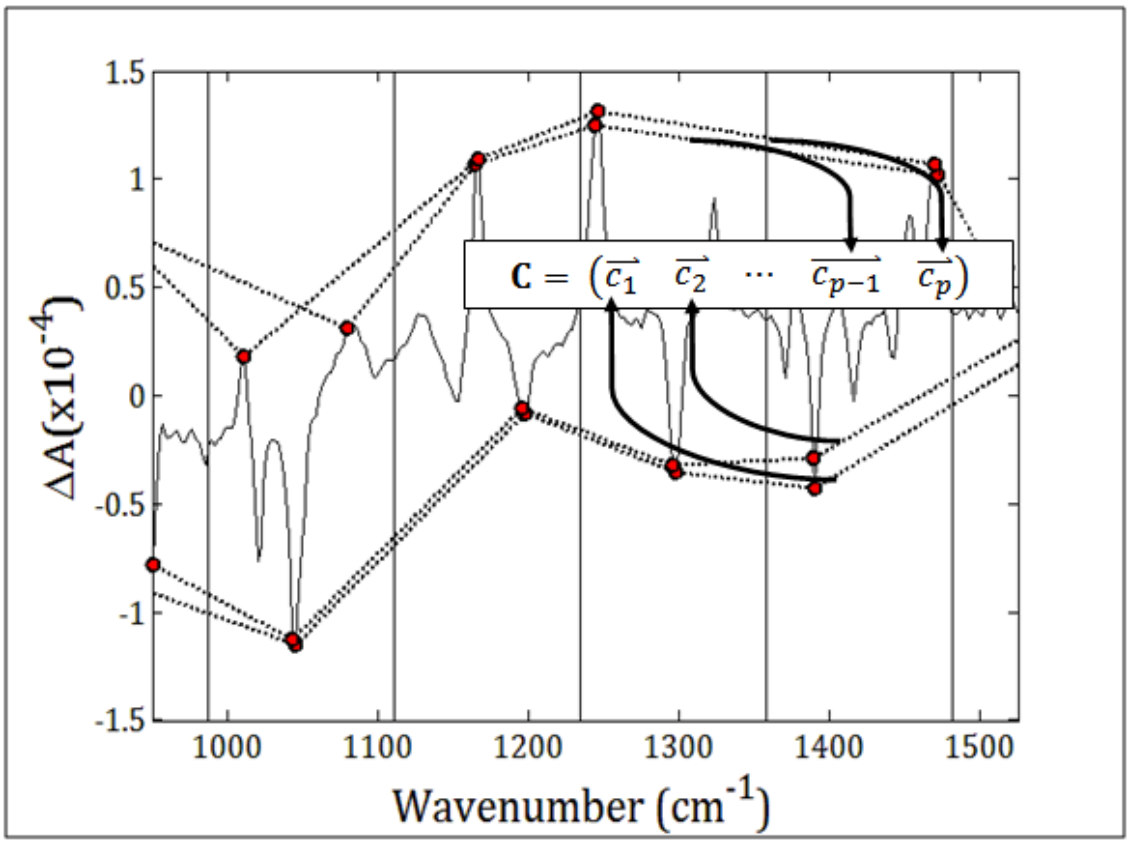


Figure 3.1: Illustration of the ranking and sorting routine used in phase I of baseline correction. Matrix C is constructed after dividing the spectrum into m bounded regions (vertical black lines). Next, all regional points (red circles) are ranked from lowest to highest. Regional minima, shown as points on the lowest dashed line, are located and their indices from $\vec{x}_{VCD}$ placed in C as the column vector $\overrightarrow{c_1}$, and so on. The process ends when regional maxima (uppermost dashed line) have been identified and placed in C as $\overrightarrow{c_p}$.

Next, the indices in C are screened row-wise to locate and remove points that are within four spectral channels. This mitigates noise amplification when a difference spectrum is generated during phase II of baseline correction (vide infra).

Phase II, or the discrimination phase, begins by calling the first column vector of C. The indices contained in $\overrightarrow{\boldsymbol{c_1}}$ map $\vec{\boldsymbol{x}}_{\boldsymbol{VCD}}$ onto an intermediate spectrum of a reduced length *m*+2. This intermediate spectrum is denoted $\vec{\boldsymbol{x}}^{\boldsymbol{1}}_{\boldsymbol{VCD}}$; *e.g.*, "1" is the affiliation with $\overrightarrow{\boldsymbol{c_1}}$. These ΔA values are forward differenced twice, where the spectral range is reduced by a total of two points leaving *m* points; i.e., a difference spectrum is produced using second-order finite differences. Formally, the equation for the second difference operator applied to any intermediate spectrum ($\vec{\boldsymbol{x}}^{\boldsymbol{j}}_{\boldsymbol{VCD}}$) is

$$\Delta_j^2 = \frac{x^j_{VCD}(c_{i+1,j}) - 2x^j_{VCD}(c_{i,j}) - x^j_{VCD}(c_{i-1,j})}{\Delta\tilde{\nu}_1 \Delta\tilde{\nu}_{-1}} \qquad \text{Equation 3.2}$$

where Δ_j^2 is the resulting second difference spectrum, $x^j_{VCD}(c_{i,j})$ is the *i*th regional value of ΔA in $\vec{\boldsymbol{x}}^{\boldsymbol{j}}_{\boldsymbol{VCD}}$, and $\Delta\tilde{\nu}_1 \Delta\tilde{\nu}_{-1}$ is the spacing (in cm^{-1}) between adjacent forward and backward data with respect to $x^j_{VCD}(c_{i,j})$.

Discrimination between baseline and non-zero points in $\vec{\boldsymbol{x}}^{\boldsymbol{1}}_{\boldsymbol{VCD}}$ proceeds by identifying if the terminal point, $x^1_{VCD}(c_{i+1,1})$, used to calculate the current second difference causes the difference in question to reside outside a locally-determined threshold region. The threshold region is constructed using a simple arithmetic back-average of the preceding *n* second differences, where $x^1_{VCD}(c_{i+1,1})$ is rejected if it resides two standard errors away from the back-average. Here, *n* is a user-specified window length for the back-average calculation. If rejected, $x^1_{VCD}(c_{i+1,1})$ is assigned as a band point and ignored in subsequent analyses. This discrimination routine is performed until each point in $\vec{\boldsymbol{x}}^{\boldsymbol{1}}_{\boldsymbol{VCD}}$ is assessed.
Next, $\vec{\boldsymbol{x}}^{\boldsymbol{1}}_{\boldsymbol{VCD}}$ is cleared from memory and the discrimination operation is reinitialized using the following column vector in C. Phase II of baseline correction is terminated when all *p* column vectors in C have been interrogated. Furthermore, this phase results in *p* rough estimates of the VCD spectrum's baseline. Finally, simple linear interpolation returns each of the *p* baseline estimates to the length of the raw spectrum and thus their true spectral range. These baseline estimates are placed row-wise in a $p \times q$ matrix, B, where rows contain the *p* baseline estimates and the *q* columns order the values of ΔA according to their wavenumber. When required, ΔA is denoted $\Delta A(\tilde{\boldsymbol{\nu}}_{\boldsymbol{j}})$ to fix its affiliation with wavenumber and dependence on the columns of matrix B.

The third and final phase aims to select pertinent response information from the rough baseline approximations in B and compress this information into a final baseline estimate. Figure 3.2 shows this accomplished using a statistical filtering scheme on each B column, i.e., the median of each column, denoted Med(B), was one filter used to estimate the final baseline. As a robust statistic, the sample median was the most obvious measure of central tendency because it protected against the influence of outliers. To prevent band and artifact distortion,[23] the final baseline estimate was smoothed using a 16-pt moving-average filter

prior to being subtracted from the raw spectrum. Figure 3.3 offers a concise summary of these three major phases of baseline estimation.

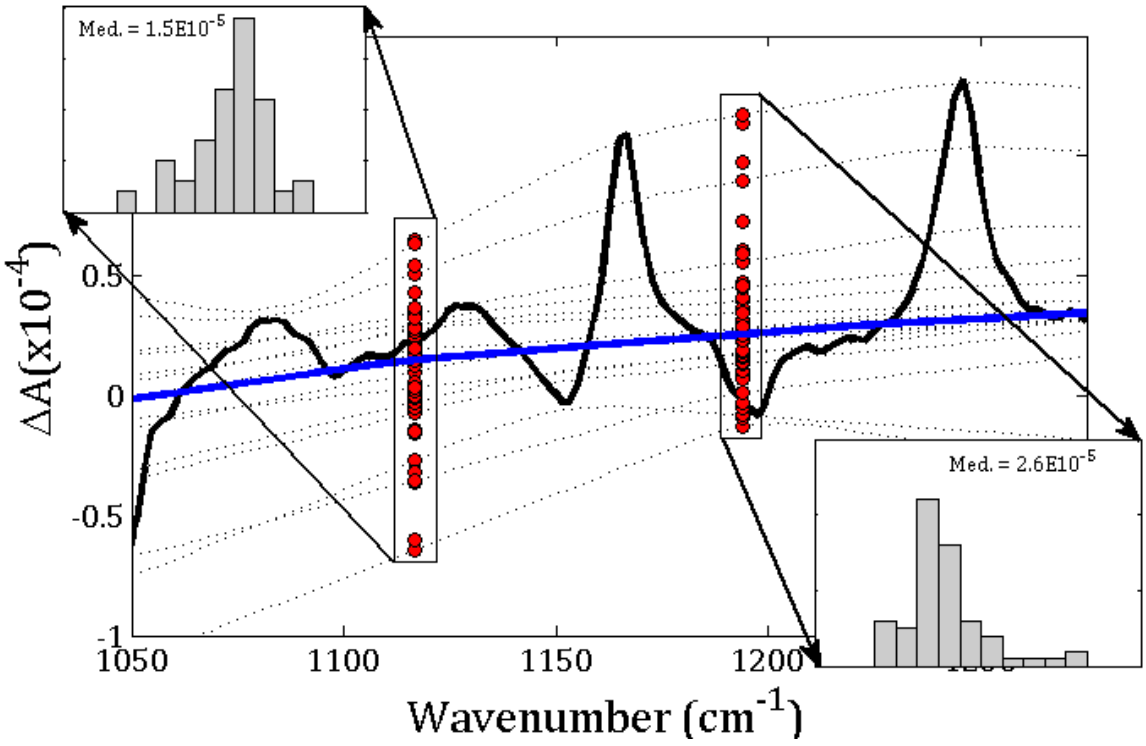


Figure 3.2: Default approach used to approximate the true baseline (blue) of a VCD spectrum (black) from the *p* rough baseline estimates in B (dotted lines). Vertically arranged red circles are two examples of $\Delta A(\tilde{\nu}_j)$ that span each of the *p* baselines. The best estimate for each $\Delta A(\tilde{\nu}_j)$ is chosen as median value of each column.

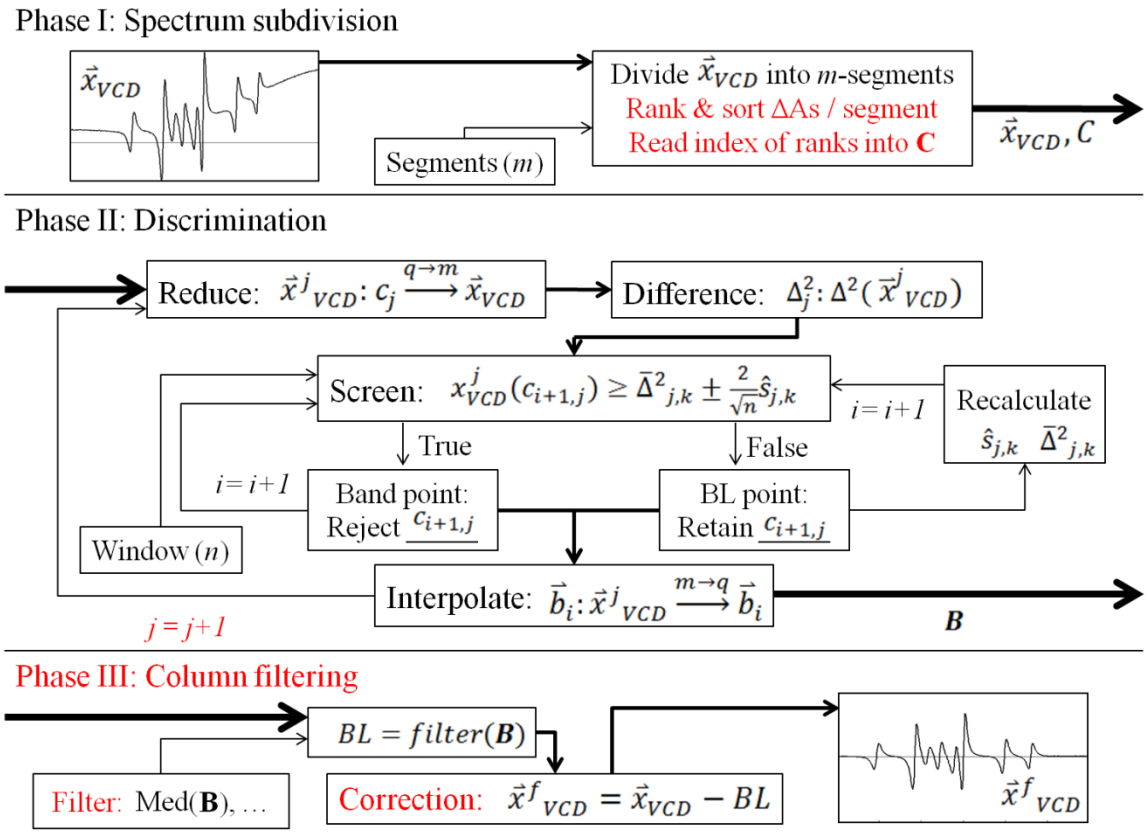


Figure 3.3: Fundamental operations of the baseline correction routine. Three phases partition the algorithm where large, bold arrows connect each. Red text indicates new additions to the routine relative our previous method.[20] Phase I shows subdivision of the raw VCD spectrum ($\vec{x}_{VCD}$) (Figure 3.1). The C matrix and $\vec{x}_{VCD}$ are passed to Phase II where the first column in C is drawn and used to reduce the size of $\vec{x}_{VCD}$ to a *m*-by-1 rough estimate of the baseline (x^{1}_{VCD}). Reduction, differencing, and screening is performed in turn (*j = j+1*) for each of the *p* baselines. Phase III proceeds after interpolation where each column in B is filtered (Figure 3.2). The final estimate of the baseline is attained, smoothed (not shown), and subtracted from $\vec{x}_{VCD}$.

An alternative column-filter was used and compared to the default median filter. In this case, the bidirectional character of VCD band shapes as well as those differential responses residing outside our region of interest were considered and used to remove poor baseline estimates from B prior to filtering. First, the direct subtraction of each rough baseline estimate in B from $\vec{x}_{VCD}$ results in a matrix of residuals, R. Consequently, each one of these residual spectra is a rough approximation to the baseline-corrected VCD spectrum. The symmetry of the distribution of ΔA values about each corrected spectrum was assessed using the residual spectrum's third standardized moment, indicating skewness. Grossly skewed residual spectra effectively indicate a baseline estimate vastly over- or undershooting the true baseline that should be removed. To accomplish this, the absolute values of these

moments were used to sort each row in B from lowest to highest absolute skewness. The first $p/2$ rows in this sorted B matrix are retained and placed into a new matrix B_r while the later $p/2$ rows are discarded. Finally, this row-reduced matrix, B_r, is passed to Phase III of baseline correction and filtered. Baselines estimated using this approach are denoted Med(B_r).

For the VCD spectra of camphor, 4096 spectral channels were available to the algorithm. Each spectrum was divided into 64 segment-regions leaving 64 points/region available for discrimination and column filtering. This choice of spectral subdivision equalized the number of experimental features available to the discrimination and column-filtering phases of baseline correction. The spectral range of interest spanned 800-1800 cm^{-1}. For the search and discrimination phase, a window size of $n = 32$ differences was chosen.

The influence of the near-zero transmittance of the antisymmetrical C-Cl stretching mode, ν_{as}(C-Cl), of CCl_4 from ~700 to 800 cm^{-1} and its significant absorption from 800 to 850 cm^{-1} resulted in a large, nearly discontinuous drop in the baseline from 800 to 850 cm^{-1}. This had a noticeably deleterious effect on the performance of the final computed baseline. To mitigate the unwanted influence of this solvent band, a sigmoid function was fitted using a general linear least-squares model[24] over the range of 806-830 cm^{-1}. If the coefficient of determination (R^2) was above 0.85, the fit was incorporated into the baseline estimate as a fixed, deterministic component, which was always subtracted from the spectrum. All programming for baseline computation and correction was performed in Matlab (2007b, The MathWorks®, Natick, MA).

Analyses of variance (ANOVAs) were performed on 4096-point simulated bipolar-band spectra to assess the confluence of segment division (m), window size (n), and column-wise filtering approaches on baseline correction. Two separate ANOVAs were performed; the first was on a population of test spectra with 64-segment divisions followed by a 32-segment population. Individually, each ANOVA tested the null hypothesis that the average baseline-correction performance was not significantly influenced by either the choice of the point-estimator used for column-filtering or the window size used for point discrimination. The performance metric used as the dependent variable in the ANOVA was the natural log of the root-mean squared deviation (RMSD) between the noisy baseline-free and baseline-corrected spectra. Comparing these two ANOVAs rendered a qualitative evaluation as to whether forcing B to remain a 64×64 matrix optimized the bias-variance trade-off between

the row and column space of B, the importance and theoretical nuances of which are outlined in the Discussion.

Formally, each individual ANOVA tested the mean performance for both *m*-segment populations using two factors containing three and five or six levels, respectively. The first factor assessed the relative importance of the choice of column-filter on baseline-correction (BLC). The three levels of the BLC factor were the sample mean or Mean(B), Med(B), and Med(B_r). Each column filtering regime was tested across our "win-*n*" factor for five or six fixed window sizes including n = 32, 16, 8, 4, and 2 for the 32 segment cases and n = 64, 32, 16, 8, 4, and 2 for the 64 segment populations. The interaction between BLC and win-*n* was also considered. Baseline correction performance for every factor-level combination (for example, Med(B) for a win-*n* of n = 8) was calculated using the mean ln(RMSD) of 300 randomly generated bipolar spectra. We abstained from the formal inclusion of a third factor for "*m*-segment divisions" in the interest of interpretive simplicity. Hence, the 3×5 and 3×6 factorial ANOVAs for the 32- and 64-segements were performed separately for qualitative comparison.

RMSD scores required logarithmic transformations to suppress positive skew, stabilize variance, and normalize their empirical distributions. Of the 300 RMSD values for each factor-level combination, those observations residing outside the empirical quantiles defined at the cumulative probability of 0.02 and 0.98 were removed as outliers. Heteroskedasticity was checked prior to ANOVA using Levene's test[25] under the null hypothesis that all variances were equal (p = 0.05). For significant main effects, multiple pair-wise comparisons between the relevant factor-levels were performed using Tukey's honest significant difference (HSD) *post-hoc* analysis. Briefly, pair-wise comparisons for the BLC factor and BLC:win-*n* interactions were not included herein for either ANOVA. All statistical analysis was performed in R (x64 2.15.2, R Foundation for Statistical Computing®, Vienna, Austria).

Results and Discussion:

Figure 3.4 shows the results of estimating and subtracting the baseline from the “better baseline” (R)-camphor spectrum. The baseline correction of both the better and worse baseline contaminated (S)-camphor spectra are included in supporting information. The large

spike near 1745 cm^{-1}, apparently caused by very strong absorption of the C=O stretching band of camphor, was ignored by the routine. This behavior is consistent with our previous study: spectral features of considerable magnitude are ignored during phase II. The green baseline estimate in Figure 3.4a shows that the simple Med(B) column-filter is capable of removing the predominant baseline components. For this spectrum, a sigmoid function was used to model the large apparent drop in ΔA caused by the solvent ν_{as}(C-Cl) band.

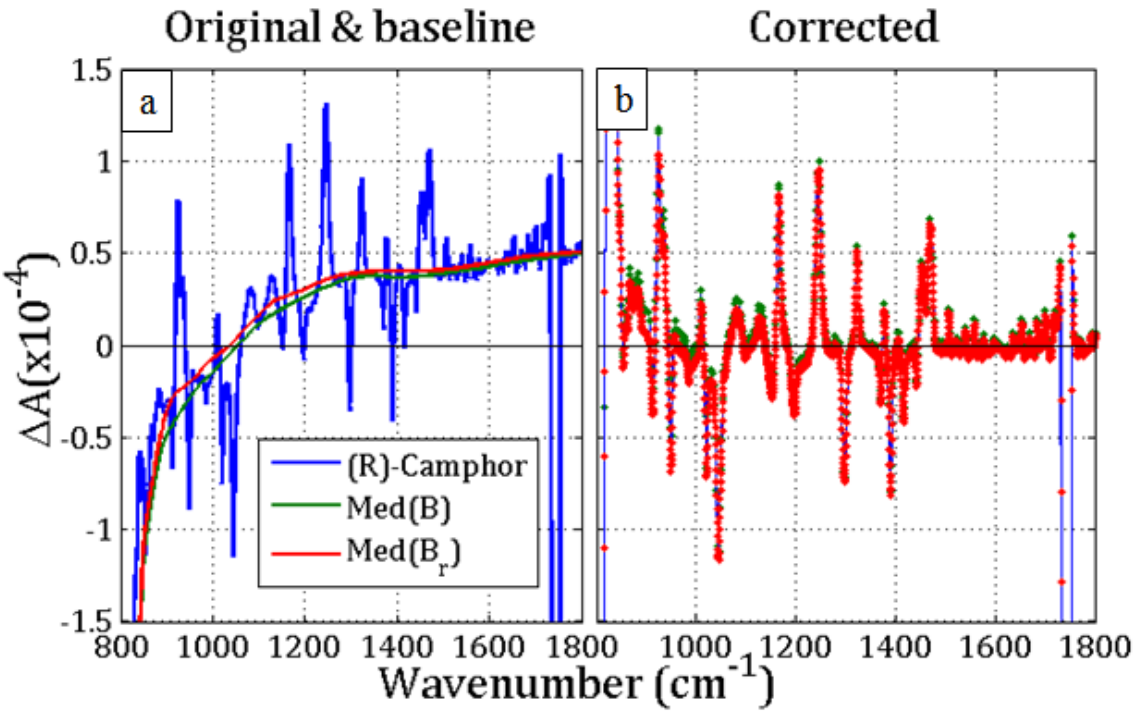


Figure 3.4: Computed baselines and raw spectrum (a) juxtaposed with the corrected spectrum (b) of "better baseline" (R)-camphor. The Med(B) (green) and Med(B_r) (red) filters were used to estimate the baseline (a) resulting in the complementary colored residual (i.e., corrected) spectrum of (R)-camphor (b). Irrespective of filter used, the estimated baseline and corrected spectrum of (R)-camphor show small point-to-point differences.

To illustrate the capability of the algorithm to accommodate various filter designs as well as to compensate for the solvent band, an additional median filtering regime was used on a row-reduced matrix B_r (Figure 3.4, red estimates). Again, this reduced matrix was constructed using row and skewness information from the residual matrix R. Clearly, median filtering columns of B and B_r results in baseline estimates that capture similar trends where a slightly upward bias is observed for B_r. This was particularly beneficial near the ν_{as}(C-Cl) solvent band.

The major drop in baseline caused by the ν_{as}(C-Cl) band effectively pulled all preceding points towards its minimal value(s). In spite of using the robust median measure and sigmoid function, baseline estimates were too skewed to yield approximations for the baseline near ν_{as}(C-Cl)) with accuracy comparable to higher wavenumbers. Figure 3.4a shows that removing rows from B, using skewness information from R, helped limit the influence of the solvent band on baseline estimation. Regardless, an accurate estimate of the baseline was not attained near the discontinuity.

Before the rationale that underpins the reduction of B to B_r is presented, the appropriateness of median filtering in phase III demands theoretical support. First, consider that each rough baseline estimate in B is a realization of the same underlying random process, i.e., each estimate is an attempt to capture the true baseline. Furthermore, we assume that each column vector, $\Delta A(\tilde{\boldsymbol{\nu}}_j)$, is an independent random variable. Hence, each $\Delta A(\tilde{\boldsymbol{\nu}}_j)$ spans our *p* baseline estimates (Figure 3.2) making each observation in $\Delta A(\tilde{\boldsymbol{\nu}}_j)$ part of the same population (Figure 3.2; red circles) and the 64 values in $\Delta A(\tilde{\boldsymbol{\nu}}_j)$ a sample from that population. Thus, the most appropriate point-estimate for each value of the "true" baseline is one that best approximates the center of each column vector's distribution (Figure 3.2; histograms). In this case, a median point-estimate was deemed best for each column distribution given its favorable robustness properties and flexibility where distribution shape is concerned.

The preoccupation with a measure of skewness, contingent upon the neglect of wavenumber dependence, is warranted. For a baseline-corrected VCD spectrum, band points should be arranged almost proportionally (positively and negatively) about a zero baseline. Scatter about the zero baseline in these VCD spectra is ultimately contingent on the real arrangement of bands as well as differential response residing outside our spectral region of interest, i.e., the extent of asymmetry is governed by optical activity. Regardless, removing the rows in B that exhibit the most skewness in their residuals screens unrealistic baseline estimates prior to column-filtering. Consequently, this reduces the influence of the skewed spectra on the final column filtering operation illustrated in Figure 3.2. These baseline estimates were also highly kurtotic, i.e., the distribution was quite narrow. We would expect

this to be characteristic of bipolar spectra in general because the total mass of band points is small relative to baseline points.

It might appear prudent to circumvent the final column-filtering phase altogether and simply choose a baseline that minimizes skewness, i.e., select the first row of B_r. For the (R)-camphor spectra, this resulted in an estimated baseline that substantially distorted VCD bands. Selecting the baseline from the first row of B_r overcompensated for the nuanced and real asymmetric character of the entire spectrum; band shape integrity was traded for a nearly equal portion of spectral points above and below the zero-line. Furthermore, by neglecting column filtering, band distortion was observed. Band distortion results from using linear interpolation and was explored in our previous work.[21] A balance is achieved when the *p/2* least-skewed baselines are column filtered, as is the case with B_r. This implicitly relaxes the requirement for exact symmetry and results in fewer band-shape distortions.

Figure 3.5 shows that the VCD spectrum and estimated baselines for the "worse baseline" spectrum of (R)-camphor were not influenced by the ν_{as}(C-Cl) solvent band. This removed the need to incorporate a sigmoid function into the baseline estimate. We can see from Figure 3.5 that the Med(B) and Med(B_r) baselines perform almost identically in our region of interest with the exception of the small deviation in the 1000-1200 cm^{-1} range. The similar behavior of these filters stems from the fact that the discontinuity experienced at ~800 cm^{-1} is muted relative to the "better baseline" (R)-camphor spectrum.

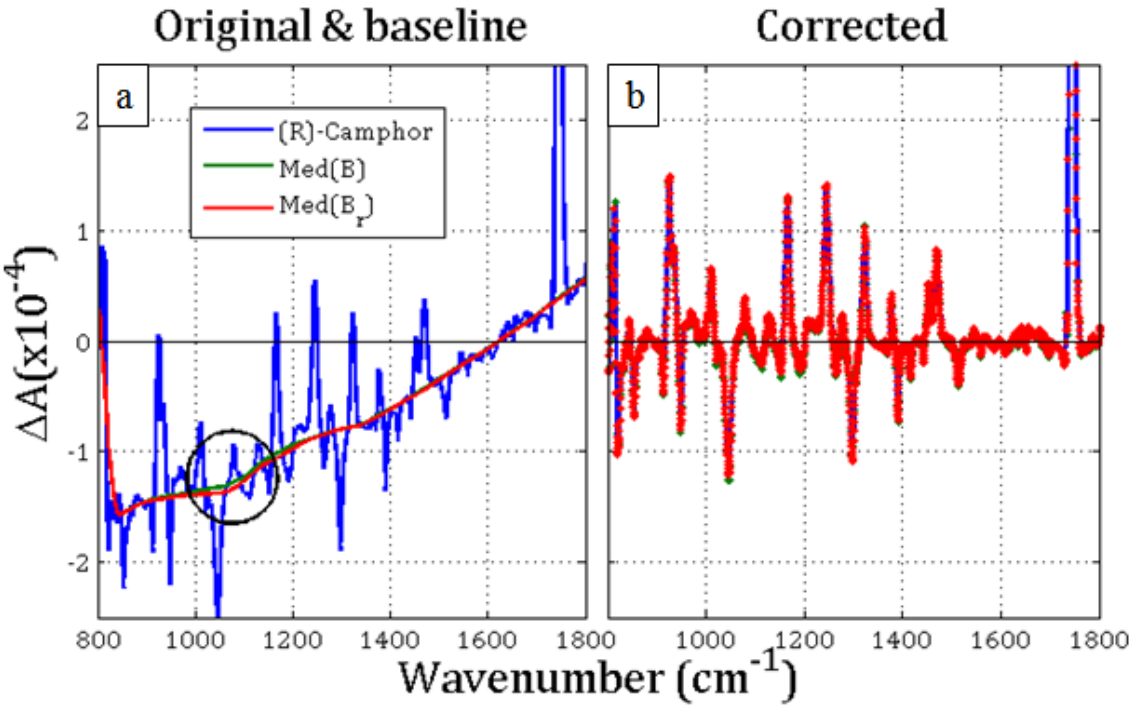


Figure 3.5: Computed baselines and raw spectrum (a) juxtaposed with the corrected spectrum (b) of "worse baseline" (R)-camphor. The Med(B) (green) and Med(B_r) (red) baselines (a) show almost identical behavior where only slight differences exist in their baseline correction performance (b).

The accuracy and universal validity of a proposed baseline correction routine are often judged in two primary ways: (1) the objective assessment of the baseline correction of simulated spectra using hypothesis tests and/or goodness-of-fit statistics[19,26] and (2) a visual/subjective inspection of manually corrected baselines to assess baseline fit acceptability. An objective performance evaluation was possible in our study using simulated spectra. We chose to use simulated spectra in an exploratory and diagnostic manner where the relationship between baseline correction performance as a function of window length (*n*) and segment divisions (*m*) was evaluated. Furthermore, we had access to a unique third option through the availability of experimental enantiomer and solvent background spectra. This facilitated direct comparison between computed baselines and the "gold standard" of VCD baseline correction.

A major qualitative obstacle exists when comparing and evaluating computed baselines and experimental backgrounds. Algorithm developers often judge computed baselines as acceptable when they are smooth and relatively broad, i.e., substantially larger than the full width at half maximum (FWHM) of spectral features. Baseline correction is

often deemed successful when the largest possible proportion of a signal's low-frequency power is removed without distorting band character. For VCD spectra, the baseline is comprised of solvent effects and interferences besides the usual slow variation due to instrument drift. These interfering components may be on the order of the FWHM of the enantiomer's bands. Thus these artifact-features are inaccessible to most computational protocols. Nevertheless, the following results will show that VCD backgrounds are still rife with low-frequency components which can be removed adequately with the present routine. Whether the inability to remove artifact-features is acceptable is up to the user-practitioner.

Figure 3.6 shows a comparison between our two computed baseline estimates superimposed on the enantiomer's spectrum and solvent backgrounds for the "better baseline" (R)-camphor spectrum. The experimental spectra show considerable artifact inclusion relative to our approach. Notable contrasts between the experimental and computed results are observed at the lower and higher wavenumbers. Specifically, the computed baselines appear to perform better than their experimental counterparts at higher wavenumbers and then gradually become worse as the large ν_{as}(C-Cl) band is approached. The Med(B_r) procedure coupled with the incorporation of the sigmoid function tended to reduce but not eliminate the influence of this band. On the other hand, the standard correction techniques performed worse on the experimental spectra at higher wavenumbers but completely removed the ν_{as}(C-Cl) solvent band and major artifact at ~1745 cm^{-1} (not shown).

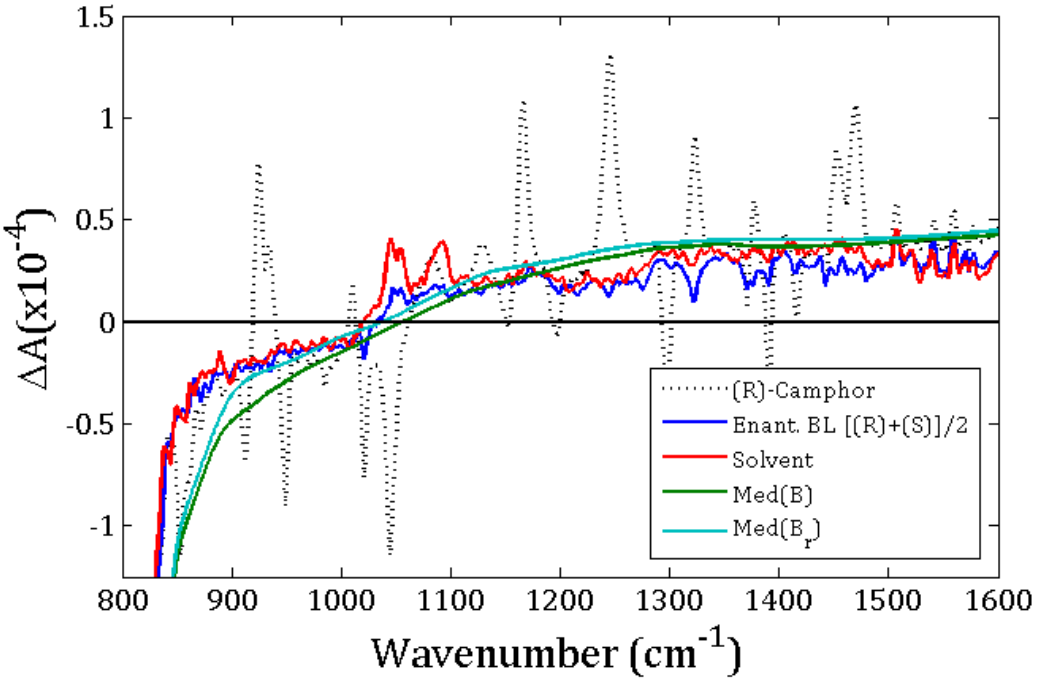


Figure 3.6: Comparison of the computed (green and cyan) and experimental baselines (blue and red) for the "better baseline" spectrum of (R)-camphor. The solvent background spectrum (red) shows a fair degree of artifacts relative to the enantiomer background (blue).

Figure 3.7 shows that computed baselines, represented solely by Med(B), appear to closely approximate a smoothed version of the experimental background spectra. Here, our Med(B) estimate is plotted against the raw and smoothed VCD spectra of the solvent in the absence of the chiral solute. As expected, our baseline estimate captures only the predominant background trends by assimilating a fraction of the information contained in nearest-neighbor ΔA values. Common convolution filters, such as the moving average and Savitzky-Golay (SG) filter employed here, attenuate artifacts in the wavenumber-domain purely by weighted proximity-averaging, limiting the specificity by which select frequencies are attenuated. In other words, these convolution filters always trade resolution for smoothness using band-pass characteristics contingent upon the choice of window length and polynomial order (in the case of the SG filter). The link between resolution and smoothness is tenuous for our computed baselines, i.e., wavenumber-domain points are selectively screened using independent baseline estimates in B as well as consciously discriminated from their nearest-neighbors. Band-pass filters lack the localization benefits[27] that point-to-point

discrimination and statistical inference afford. This heightened selectivity in the wavenumber domain translates into frequency space.

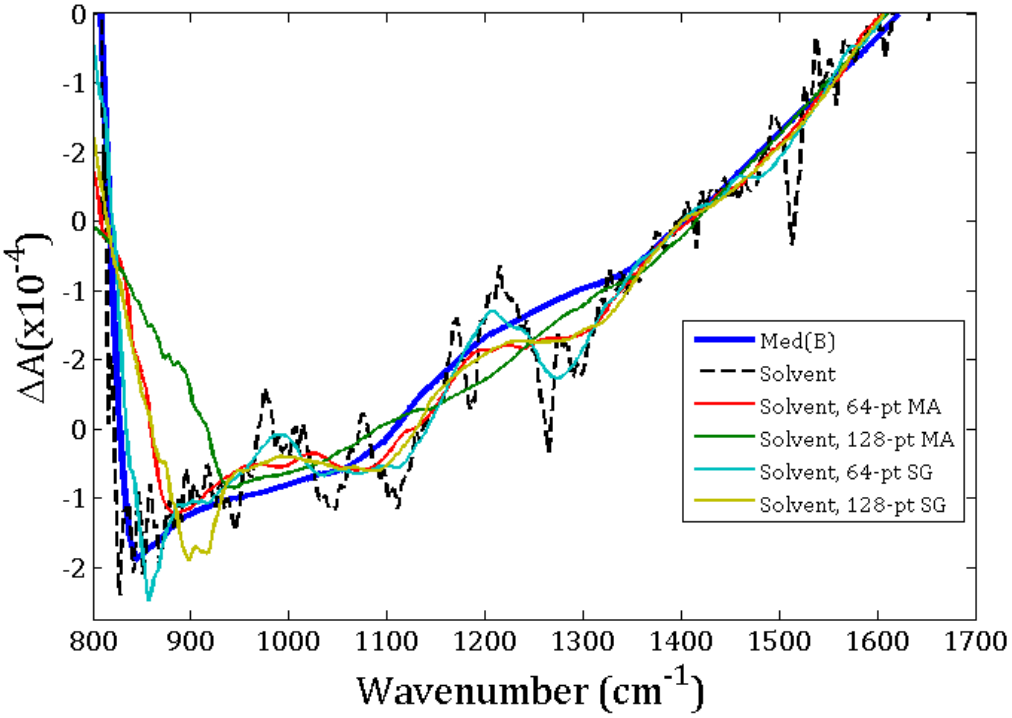


Figure 3.7: Comparison of the computed median filtered baseline to a series of moving average (MA) and Savitzky-Golay (SG) smoothed solvent VCD backgrounds. The raw solvent spectrum (dashed line) is also shown.

Figure 3.8 shows the normalized cumulative periodogram (NCP)[28] of the uncorrected "worse baseline" (R)-camphor spectrum compared to the computationally- and solvent-corrected VCD spectrum. The "better baseline" case is not included here due to its marginal difference in frequency domain behavior. Additionally, a single SG-filtered, solvent-corrected spectrum is used as a representative for the other moving-average filters because of their nearly indistinguishable periodograms. This figure illustrates that in the uncorrected spectrum (blue), 80% of the VCD signal's power is concentrated in the first 0.5% of the frequencies (excluding the zeroth frequency), which is, of course, indicative of a non-zero baseline.[29] Regardless of correction method used, the first 80% of a spectrum's power was confined to approximately 11% of its frequencies.

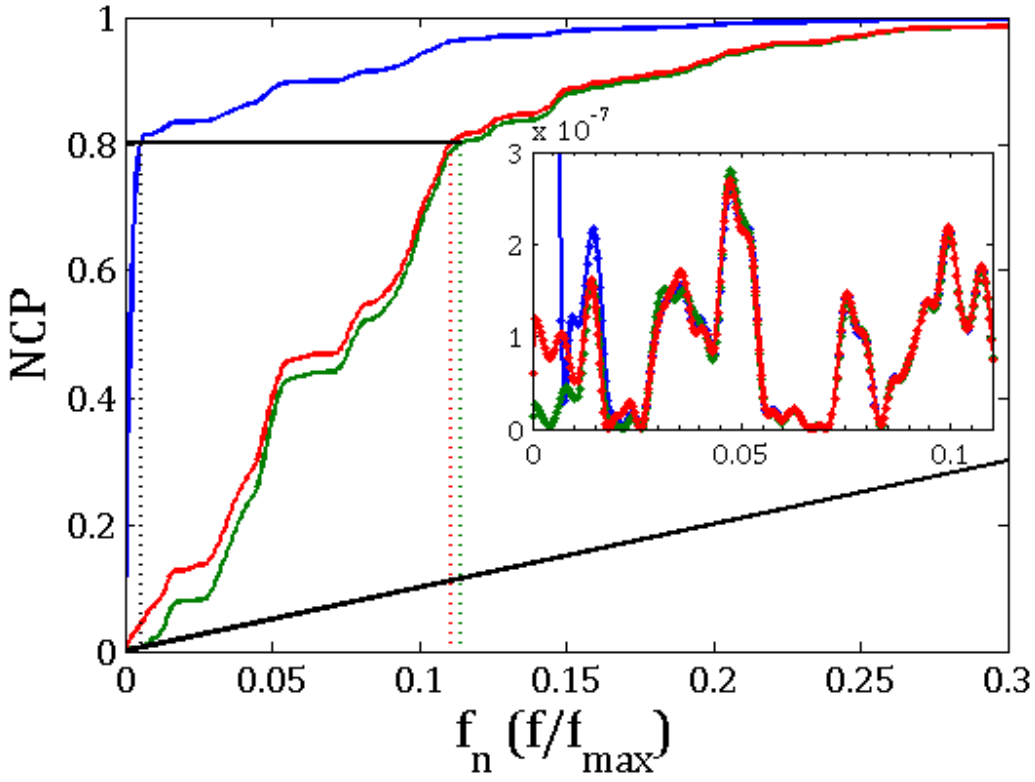


Figure 3.8: Normalized cumulative periodogram (NCP) comparing the uncorrected (blue) and corrected "worse-baseline" spectrum of (R)-camphor for the Med(B) (green) and second-order, 129-point, Savitzky-Golay smoothed CCl_4 solvent spectrum (red). The black line references a periodogram of cumulatively summed white noise. These comparisons, and particularly the inset, show that both baseline correction techniques remove the majority of low-frequency components. Approximately 80% of a spectrum's power was concentrated on the first 0.5% and 11% of frequencies for the raw and corrected spectra, respectively.

Substantial differences exist prior to this 11% point. The inset in Figure 3.9 shows that our computed baseline removed a higher proportion of low-frequency components than the smoothed solvent-corrected spectrum. This inset unambiguously confirms the approach to be a slightly more selective high-pass filter, leaving the intermediate frequencies (> 2.5%), which presumably correspond to vibrational bands, less disturbed. Although significant, it is difficult to generalize these findings due to the underlying non-stationarity of the analyzed signal. Note, this approach, and likely most computational routines, cannot remove solvent-artifact components confined to intermediate frequencies in the pass-band.

Two major issues have yet to be addressed: window size (*n*) for the threshold region and number of segments (*m*) required for accurate baseline estimation. These two choices are intimately interconnected at phase II of baseline correction: the window size is used for the

back-average calculation and explicitly determines the feature selection behavior of the algorithm while the number of spectral segments determines the number of points interrogated. Ultimately, the number of segments, *m*, constrains the window size, *n*, since *n* must be less than the number of *m* points to be evaluated.

Formally, this problem is best addressed by understanding the implications of generating a second-derivative spectrum. Generating a difference spectrum using discrete difference approximations completely removes all constant (p_0) and linear ($p_1\tilde{\nu}$) components from the raw spectrum while higher-order polynomial components and random noise ($\varepsilon(\tilde{\nu})$) are left intact. Although unstated, our baseline correction approach makes the approximation that parabolic components ($p_2\tilde{\nu}^2$) comprise the majority of the interfering baseline. Thus, our discrimination routine uses the fact that p_2 is constant because

$$\frac{d^2}{d\tilde{\nu}^2}[p_0 + p_1\tilde{\nu} + p_2\tilde{\nu}^2 + p_3\tilde{\nu}^3 + \cdots + \varepsilon(\tilde{\nu})] \sim p_2 + p_3\tilde{\nu} + \cdots + \varepsilon''(\tilde{\nu}). \quad \text{Equation 3.3}$$

To relax the expectation that p_2 is constant in a real spectrum, our *n* point mean of prior differences is used to estimate p_2 locally. Ideally, the back-average and corresponding threshold region is chosen not only to account for random noise variations but also to capture higher-order baseline contributions. Formally,

$$p_2 + p_3\tilde{\nu} + \cdots + \varepsilon''(\tilde{\nu}) \sim \bar{\Delta}^2{}_{j,k} \pm \frac{2}{\sqrt{n}}\hat{s}_{j,k} \quad \text{Equation 3.4}$$

where the back-averaged mean $\bar{\Delta}^2{}_{j,k} = \frac{1}{n}\sum_{k=1}^{n}(\Delta^2{}_{j\,,\,i\text{-}k})$ precedes any point to be evaluated and $\hat{s}_{j,k}$ is the sample standard deviation of the *n* preceding points. This concise formulation offers local approximations of the constant component of the difference spectrum. Hence, whether a scrutinized point is included as part of the baseline is contingent upon *n* because the threshold boundaries make explicit use of *n* via the standard error statistic. It follows that a smaller *n* translates into a less confident local approximation for p_2. This instructs the algorithm to accept more points as baseline under the tacit assumption that they correspond to real, higher-order baseline contributions. Conversely, a larger window size a priori instills a greater degree of confidence in the local estimate of $p_2$implying that parabolic baseline components dominate the spectrum.

Under these terms, a convenient means of circumventing any arbitrary choice of *n* is by setting its length equal to a fraction of the total number of *m*-segment divisions. Doing this fixes the proportion of the spectrum used for the threshold region irrespective of the

number of total points available to the algorithm. Fortunately, the pragmatic choice for the number of *m*-segments is much simpler with the value of *n* readily defined ex post facto.

At its most basic, our baseline correction routine is a two-dimensional problem conditioned on the matrix C, the size of which is governed by two considerations: the number of segments (*m*) and the number of points per segment (*p*). For a fixed spectral range, increasing the number of *m*-segments decreases the available *p*-points per segment, and vice versa. The critical dimension in C is *p* since it specifies the number of independent, rough baseline estimates acquired after phase II which are condensed into a single estimate of the true baseline during phase III. Here, the statistical power of the median estimates for the final ΔA values increases as the number of independent baselines generated increases. However, there is a trade-off between accuracy in the full baseline estimate (wavenumber dimension) and certainty in the median estimates (ΔA dimension).

The number of independent baselines generated is maximized when $p = q$ and $m = 1$, i.e., one "segment" exists containing all *q* points in the entire spectrum. Thus, the median estimate for this "segment" would be the median ΔA-estimate for the entire spectrum (located at the center of the spectral range). It is not difficult to imagine that interpolating between a spectrum's median and two end points likely results in a highly equivocal baseline estimate. However, this median estimate would appear superb from a statistics perspective: the point-estimate of the median ΔA uses every available point in the spectrum ($N = q =$ 4096 for (R)-camphor). The improved statistical certainty in the ΔA-dimension comes with the complete loss of any selectivity of spectral features (or spatial relevance). In other words, the bias is too large to adequately distinguish baseline from band point.

Conversely, setting $m = q$ results in $p = 1$, which would effectively divide the spectrum into *q* "segments" leaving only a single point from which to estimate a median statistic per *m*-segment. All beneficial bias required to distinguish baseline from band point is abandoned in exchange for all the variance contained in the spectrum. In other words, the wavenumber-dimension is favored at the expense of any discrimination between baseline and band point (bias). These examples demonstrate the well-understood principle in regularization theory known as the bias/variance trade-off. [30,31] The term bias/variance trade-off is used loosely as an aid to understand the inextricable association between the two dimensions in C since our analysis makes no formal attempt to measure variation or bias.

Given a finite set of spectral channels and ΔA values to predict a baseline, the most pragmatic means of minimizing the trade-off between *m* and *p* is to define C as a square matrix. In this way, a zero-sum game is hypothesized: neither the discrimination phase (*m*-dependent) nor median-estimates (*p*-dependent) have unequal access to the spectrum's resources in relative terms. To accommodate any length, we chose to make *n* a fraction of the number of segments and, for our (R)-camphor spectra, $n = m/2 = 32$.

The choice to use a 32-point window for the baseline correction of the (R)-camphor spectrum was motivated by the observation that it contained only minor baseline errors in even the worst instance. Local baseline variations appeared to be dictated by lower-order polynomial terms, although this was not true near the ν_{as}(C-Cl) solvent band. Furthermore, filtering column-wise appeared to suppress most of the errors when baseline variations were assigned as bands.

The results of the 3×6 ANOVA for the 64-segment baseline correction simulation results are shown in Table 3.1. Very significant main effects for BLC (F =25.7, degrees of freedom (df) = 2, $p < 0.0001$), win-*n* ($F = 3.52 \times 10^3$, df =5 , $p < 0.0001$), and their interaction (F =9.00, df = 10, $p < 0.0001$) were observed. (The 3×5 ANOVA summary table exhibited very significant main effects for both factors and their interaction and was included in supporting information.) Table 3.1 indicates that the populations under scrutiny show substantial performance deviations as a function of column-filtering, window size, and an interaction between column-filtering and window length. We note here that the 64-segment divisions produce a square C-matrix for 4096 spectral channels. Hypothetically, this should indicate whether the aforementioned bias-variance trade-off is optimally balanced when spectral resources are distributed equally across the $\tilde{\nu}$ and ΔA dimensions.

Table 3.1: ANOVA results for 3×6 factorial design using 64-segment population

	df	Sum Sq	Mean Sq	F value	Pr(>F)
BLC	2	11	5.7	2.57E+01	7.74E-12
win-*n*	5	3907	781.5	3.52E+03	< 2e-16
BLC:win-*n*	10	20	2	9.002	7.31E-15

To evaluate the bias-variance trade-off hypothesis, a comparison of each 64- and 32-segment ANOVA is facilitated by the tabular juxtaposition of their mean RMSD scores. Table 3.2 and table 3.3 show that the differences between the 64- and 32-segment populations are large. The grand mean ($\bar{x}_{..}$) between each data set is markedly different, indicating that 32-segment divisions show an improved mean performance in general, irrespective of window size and baseline-correction routine chosen. Using the sample mean as a column filter (row 1) was generally poorer than the more robust median measures (rows 2 and 3) with the exception of low window sizes for both the 32- and 64-segment cases. Regardless, these low window sizes perform four to five times worse than their larger counterparts irrespective of BLC level. Both tables show similar trends as a function of window length parameter, i.e., window length is best set at or nearly equal to *m* segments providing evidence that phase II (particularly, "screening") be left out of the baseline correction routine altogether. These data also suggest that we can safely reject the hypothesis that forcing C to remain square, as a pragmatic compromise between bias and variance, is unfounded. In fact, choosing a 32-segment difference spectrum, and therefore generating 128 independent baselines, appears favorable.

Table 3.2: ANOVA summary table of mean performances for 64-segment population. A lower mean indicates a better estimate of the true baseline. Bold values indicate the lowest RMSD across both factors (interior) and within a given factor (margins). $\bar{x}_{..}$ is the grand mean for the entire sample.

Factor	win-*n*						
BLC	Levels	64	32	16	8	4	2	$\bar{x}_{BLC}(10^6)$
	Mean(B)	1.2	1.2	1.2	1.9	7.4	8.4	3.6
	Med(B)	1.1	1.1	1.1	1.7	8.2	8.7	3.6
	Med(B_r)	1.1	1.3	1.5	2.0	7.7	8.8	3.7
	$\bar{x}_{\text{win}-n}(10^6)$	1.1	1.2	1.3	1.9	7.7	8.6	$\bar{x}_{..}(10^6)$ = 3.6

Table 3.3: ANOVA summary table for 32-segment population. The 64 length level on win-*n* is missing by virtue of the constraint $n \leq m$.

Factor	win-*n*						
BLC	Levels	64	32	16	8	4	2	$\bar{x}_{BLC}(10^6)$
	Mean(B)	-	1.1	1.1	1.1	1.7	4.6	1.9
	Med(B)	-	0.8	0.7	0.8	1.4	4.6	1.7
	Med(B_r)	-	0.7	0.8	1.0	1.7	4.7	1.8
	$\bar{x}_{win-n}(10^6)$	-	0.8	0.9	1.0	1.6	4.7	$\bar{x}_{\cdot\cdot}(10^6)$ = 1.4

Since determining the influence of window length on baseline correction was our second concern, Table 3.4 presents all pair-wise comparisons for our win-*n* factor irrespective of BLC level for the 64-segment populations. Table 3.4 corroborates our qualitative assessment that algorithm performance decreases with decreasing *n*. Specifically, baseline correction performed significantly worse for the lower window sizes (i.e., 2 and 4) relative to the larger window sizes as indicated by the very significant adjusted *p*-value for pair-wise comparisons (Δn_{ij}). A large and steady increase in performance deficits occurred as a fairly stable function of *n*. For example, the magnitude of the deviation in mean performance increased steadily as the difference in window size increased, reaching its highest value for the [2]-[64] comparison. These decrements in performance always singled-out a smaller window length as the source of performance reduction. This behavior suggests the sparing use of band-point screening in phase II, or if desired, its restricted application to a value of *n* close to the number of total segments in each difference spectrum.

Table 3.4: Tukey's pair-wise comparison table for 64-segment population means ($\bar{X}_{ij}$) on the win-*n* levels. All positive $\Delta\bar{X}_{ij}$ indicates that average performance was always worse for a smaller window size but not significantly so for the [32]-[64] comparison ($p_{adj} > 0.05$). For a fixed win-n_i, performance declined as a function of increasing difference in window size.

[win-n_i] - [win-n_j]	$\Delta\bar{X}_{ij}$	p_{adj}
[2]-[4]	0.11	0
[2]-[8]	1.51	0
[2]-[16]	1.91	0
[2]-[32]	1.99	0
[2]-[64]	2.02	0
[4]-[8]	1.41	0
[4]-[16]	1.81	0
[4]-[32]	1.88	0
[4]-[64]	1.92	0
[8]-[16]	0.40	0
[8]-[32]	0.47	0
[8]-[64]	0.51	0
[16]-[32]	0.07	0.02
[16]-[64]	0.11	0
[32]-[64]	0.04	0.6

The poor performance of the small window back-average parameter (*n*) suggests two possibilities, the most intuitive being that band points are inadvertently included as baseline points. Remarkably, this was not found to be the case. Figure 3.9 illustrates that a smaller window length has a very large effect on the viability of the computed baseline.

In Figure 3.9, baseline points are correctly accepted for the small window calculation until confronted with a fair degree of uncertainty near the highly congested region of bipolar bands. This indicates that for many of the rough baseline estimates contained in B the threshold region used during discrimination became fixed near the artificially low-amplitude baseline component at approximately 950 cm^{-1}. This caused the median filter to become incorrectly pulled toward a low ΔA. This behavior was consistent irrespective of the *m*-segment divisions specified. Essentially, the small window ($n < 8$) back-average has the counterintuitive effect of accepting fewer points as baseline in contrast to our theoretical derivation (Equations 3.3 and 3.4). The incongruity between theory and these observations cast doubt on using a simple back-average to simultaneously capture low-order components

and discriminate higher-order baseline from band components (Equation 3.4). Specifically, the real root of these performance issues rests with the point rejection behavior. Specifically, when a point is rejected and assigned as a band point, it is discarded from use in the moving average for the following point. Therefore, when later points are evaluated at substantially higher ΔA (e.g., 2000 cm^{-1}), they are unintentionally neglected as the moving average is confined to low ΔA.

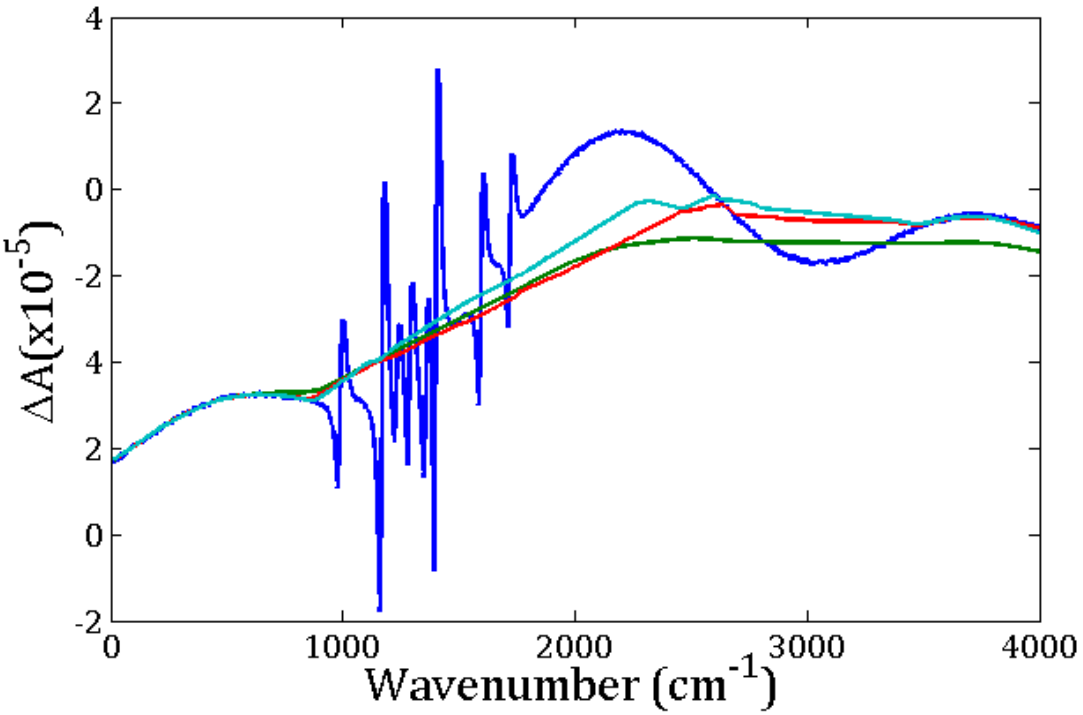


Figure 3.9: Baseline estimates on a random, simulated bipolar spectrum for Mean(B) (green), Med(B) (red), and Med(B_r) (cyan) filtering schemes for *n* = 2. Notice how an accurate baseline estimate becomes corrupted near ~950 cm^{-1}.

ANOVAs and experimental spectra point to a simple and general protocol using the current algorithm to remove baselines from VCD spectra. First, set *m*-segment divisions at or slightly less than the square-root of the number of spectral channels during phase I of the algorithm. Second, ignore the differencing and screening operations in phase II of the algorithm, i.e., generate B directly from the columns in C. Finally, use the default median filter in phase III. Under this parametric regime, our baseline correction algorithm realizes full automation. To illustrate, Figure 3.10 shows the efficacy of using these parameters for the baseline correction of 3a-vinyl-3,3a,4,5-tetrahydro-*2H*-cyclopental[*b*]furan.

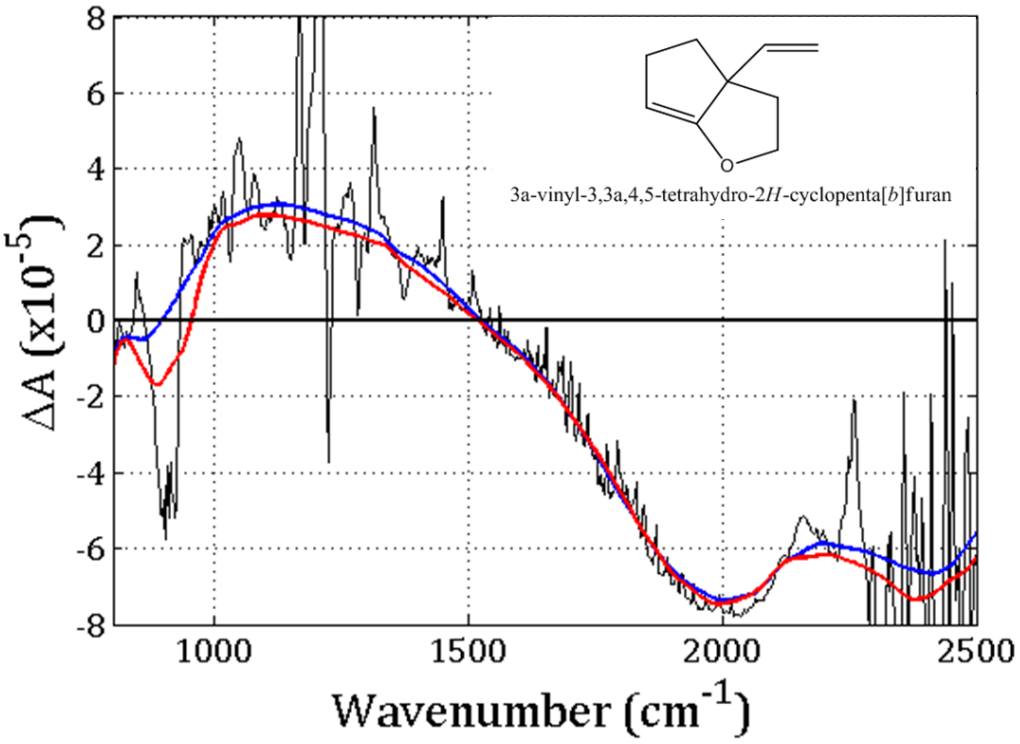


Figure 3.10: An experimental VCD spectrum of 3a-vinyl-3,3a,4,5-tetrahydro-*2H*-cyclopenta[*b*]furan corrected using the default 64-regions with the the Med(B) (blue) and Med(B_r) (red) filters.

Conclusions

A new computational algorithm for the baseline correction of VCD spectra complements the preexisting correction techniques that employ experimental background spectra. Given these promising results, we anticipate that the computational correction of VCD spectra is a cost-effective alternative to experimental approaches.

Careful scrutiny of the present correction routine revealed that limiting the intentional discrimination of baseline and band points during phase II is warranted. As highlighted during ANOVA, at $n > 8$ the sticking problems of the discrimination routine were (Figure 3.9) mostly overcome and lower-order polynomial components were exclusively mapped as baseline. More importantly, this indicates a fortiori that the best choice of n in phase II is either to remove it from the procedure or possibly redesign threshold criteria. More importantly, if the differencing and screening operations of phase II (see Figure 3.3) are removed, baseline correction realizes full automation for default settings i.e., 64-divisions and median filtering.

Broadly speaking, it appears that utility is gained when baseline correction proceeds using a thoroughly statistical, model-free paradigm. For this reason, the present routine realizes remarkable flexibility in terms of column filter design. An intuitive or theoretical grasp of a raw spectrum's baseline-free character affords an experimenter room to incorporate prior information into filter design and optimization. Although a median filter works well by default, knowledge of the particular attributes of B, R, and $\vec{\boldsymbol{x}}_{\boldsymbol{VCD}}$ might benefit from the use of weighted or trimmed mean measures. The use of a sigmoid function to limit the influence of sharp changes in the baseline due to the highly absorbing ν_{as}(C-Cl) was a good example of incorporating rational, prior information into the estimation scheme. Unfortunately, the application of a case-specific model into the correction protocol violates the model-free and fully automated character of the routine. For VCD spectra in particular, we restrain ourselves from advocating for full automation due to the aforementioned inability of most (or any conceivable) routine to remove higher-frequency artifacts and interferences. Discarding full automation to facilitate removing additional interferences remains desirable, i.e., user intervention is reliable when the optical process is well-understood and posed appropriately to the computational routine.

References

1. Y. He, W. Bo, R. K. Dukor, L. A. Nafie, "Determination of absolute configuration of chiral molecules using vibrational optical activity: A review". 2011. Appl. Spectrosc. 65(7): 699-723.

2. L. A. Nafie, "Vibrational Optical Activity". Appl. Spectrosc. 1996. 50(5): 14A-26A.

3. P. L. Polavarapu, G. Shanmugam, "Comparison of mid-infrared Fourier transform vibrational circular dichroism measurements with single and dual polarization modulations". 2011. Chirality 23(9): 801-807.

4. P. J. Stephens, M. A. Lowe, "Vibrational circular dichroism". 1985. Annu. Rev. Phys. Chem. 36: 213-241.

5. P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch. "Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields". 1994. J. Phys. Chem. 98(45): 11623-11627.

6. P. J. Stephens, F. J. Devlin, C. S. Ashvar, C. F. Chabalowski, M. J. Frisch, "Theoretical Calculation of Vibrational Circular Dichroism Spectra". 1994. Faraday Discuss. 99: 103-119.

7. S. Abbate, G. Longhi, E. Castiglioni, F. Lebon, P. M. Wood, L. W. L. Woo, B. V. L. Potter, " Determination of the absolute configuration of aromatase and dual aromatase-sulfatase inhibitors by vibrational and electronic circular dichroism spectra analysis". 2009. Chirality 21(9): 802-808.

8. J. Autschbach, " Computing chiroptical properties with first-principles theoretical methods: background and illustrative examples". 2009. Chirality. 21(1E): e116-e151.

9. L. A. Nafie, Vibrational Optical Activity: Principles and Applications. Hoboken, NJ: John Wiley and Sons, 2011. Pp. 239-241.

10. T. Eriksson, S. Bjöurkman, B. Roth, Å. Fyge, P. Höuglund, " Stereospecific determination, chiral inversion in vitro and pharmacokinetics in humans of the enantiomers of thalidomide". 1995. Chirality 7(1): 44-52.

11. O. McConnell, Y. He, L. Nogle, A. Sarkahian, " Application of chiral technology in a pharmaceutical company. Enantiomeric separation and spectroscopic studies of key asymmetric intermediates using a combination of techniques. Phenylglycidols". 2007. Chirality 19(9): 716-730.

12. D. J. Minick, R. C. B. Copley, J. R. Szewczyk, R. D. Rutkowske, L. A. Miller, "An investigation of the absolute configuration of the potent histamine H3 receptor antagonist GT-2331 using vibrational circular dichroism". 2007. Chirality 19(9): 731-740.

13. B.L. Gao, C.M. Zhang, Y.Z. Yin, L.Q. Tang, Z.P. Liu, "Design and synthesis of potent HIV-1 protease inhibitors incorporating hydroxyprolinamides as novel P2 ligands". Bioorg. Med. Chem. Lett. 2011. 21(12): 3730-3733.

14. B. Kesteleyn, K. Amssoms, W. Schepens, G. Hache, W. Verschueren, W. Van De Vreken, K. Rombauts, G. Meurs, P. Sterkens, B. Stoops, L. Baert, N. Austin, J. Wegner, C. Masungi, I. Dierynck, S. Lundgren, D. Jönsson, K. Parkes, G. Kalayanov, H. Wallberg, Å. Rosenquist, B. Samuelsson, K. Van Emelen, J. W. Thuring, "Design and synthesis of HIV-1 protease inhibitors for a long-acting injectable drug application". Bioorg. Med. Chem. Lett. 2013. 23(1): 310-317.

15. S. Abbate, A. Ciogli, S. Fioravanti, F. Gasparrini, G. Longhi, L. Pellacani, E. Rizzato, D. Spinelli, P. A. Tardella, "Solving the Puzzling Absolute Configuration Determination of a Flexible Molecule by Vibrational and Electronic Circular Dichroism Spectroscopies and DFT Calculations: The Case Study of a Chiral 2,2_-Dinitro-2,2_-biaziridine". 2010. *Eur. J. Org. Chem.* 2010(32): 6193–6199.

16. C. Guo, R. D. Shah, R. K. Dukor, X. Cao, T. B. Freedman, L. A. Nafie, " Determination of enantiomeric excess in samples of chiral molecules using Fourier transform vibrational circular dichroism spectroscopy: Simulation of real-time reaction monitoring". 2004. Anal. Chem. 76(23): 6956-6966.

17. C. Guo, R. D. Shah, R. K. Dukor, X. Cao, T. B. Freedman, L. A. Nafie, " Enantiomeric excess determination by Fourier transform near-infrared vibrational circular dichroism spectroscopy: Simulation of real-time process monitoring". 2005. Appl. Spectrosc. 59(9): 1114-1124.

18. P. Gemperline. "Calibration". In: P. Gemperline, editor. Practical Guide to Chemometrics. Boca Raton, FL: CRC/Taylor & Francis, 2006. 2nd ed. Chap. 5, Pp.140-147.

19. G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, M. W. Blades, "Investigation of selected baseline removal techniques as candidates for automated implementation". 2005. Appl. Spectrosc. 59(5): 545-574.

20. L. Shao, P. R. Griffiths, " Automatic Baseline Correction by Wavelet Transform for Quantitative Open-Path Fourier Transform Infrared Spectroscopy". 2007. Env. Sci. Technol. 4(20): 7054-7059.

21. A. T. Weakley, P. R. Griffiths, D. E. Aston, " Automatic baseline subtraction of vibrational spectra using minima identification and discrimination via adaptive, least-squares thresholding". 2012. Appl. Spectrosc. 66(5): 519-529.

22. P. C. Hansen. Discrete Inverse Problems: Insight and Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2010. Chap. 3, Pp. 50.

23. M. Friedrichs, " A model-free algorithm for the removal of baseline artifacts". 1995. J. Biomolecular NMR 5(2): 147-153.

24. B. J. T. Morgan. Applied Stochastic Modelling. Boca Raton, FL: CRC/Taylor & Francis, 2009. 2nd ed. Chap. 8, Pp. 238-239.

25. D. C. Howell. Statistical Methods for Psychology. Belmont, CA: Thomson Wadsworth, 2009, 7th ed. Chap. 7, Pp. 214-215.

26. H. G. Schulze. R. B. Foist, K. Okuda, A. Ivanov, R. F. B. Turner," A small-window moving average-based fully automated baseline estimation method for Raman spectra". 2012. Appl. Spectrosc. 66(7): 757-764.

27. M. Frazier. An Introduction to Wavelets Through Linear Algebra. New York, NY: Springer, 1999. Chap. 3, Pp. 165-167.

28. P. C. Hansen. Discrete Inverse Problems: Insight and Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2010. Chap. 5, Pp. 98-101.

29. C. Chatfield. The Analysis of Time Series: An Introduction. Boca Raton, FL: CRC/Taylor & Francis, 2004. 6th ed. Chap. 6, Pp. 110.

30. C. Chatfield. The Analysis of Time Series: An Introduction. Boca Raton, FL: CRC/Taylor & Francis, 2004. 6th ed. Chap. 7, Pp. 142.

31. P. C. Hansen. Discrete Inverse Problems: Insight and Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2010. Chap. 5, Pp. 86-87.

Chapter 4. Multivariate analysis of micro-Raman spectra of thermoplastic polyurethane blends using principal component analysis and principal component regression

Reproduced with permission (Appendix A): **Andrew T. Weakley**, P.C. Temple Warwick, Thomas E. Bitterwulf, D. Eric Aston, Applied Spectroscopy, 2012, **66**(11): 1269-1278.

Abstract

Probing the specific hydrogen-bonding behavior of thermoplastic polyurethane (TPU) blends using vibrational spectroscopies remains the *sin qua non* for understanding the link between hydrogen-bonding and phase segregation behavior. However, current literature holds to more traditional univariate approaches when studying the morphologically interesting normal molecular vibrations of TPUs. In the present study, multivariate analysis, including principal component analysis (PCA) and principal component regression (PCR), is used to scrutinize the relevant Raman bands acquired from a binary mixture of analogous TPUs copolymer blends. Considering the near identical behavior of selected spectral regions, PCA was capable of isolating linear and nonlinear composition dependent trends on PC-scores plots. From here, the PC-scores, extracted from wavelengths comprising the carbonyl stretching region (1681-1764cm^{-1}), CH_2 deformations (1380-1500 cm^{-1}), and aromatic stretch from the hard segment (1617 cm^{-1}), and Amide II mixed band (1540 cm^{-1}), were used to explicitly predict the mole fraction of hard segment present in each blend using PCR. Spectral preprocessing, wavelength selection, and variable scaling were a major factor in PCR accurately predicting the weight fraction of each copolymer in spite of the clearly evident, blend-specific spectroscopic behavior.

Key Words: Thermoplastic polyurethane blend; Raman spectra; principal component analysis; principal component regression

Introduction

Polyester-based, thermoplastic polyurethane (TPU) elastomers are an important class of segmented, multi-block copolymers in which unique mechanical properties are obtained by the regulation of chemical and physical networks.[1] Mechanical and viscoelastic properties are controlled by the careful selection and thermal manipulation of glassy, semicrystalline

and soft, rubbery TPU blocks.[2-4] Rubbery blocks consist of a flexible and relatively long polyester segment. Conventionally, this has been referred to as the TPU soft segment or polyol. The much more rigid and shorter hard segment typically consists of *a,b*- aromatic diisocyanates with short diol chain extenders.[5, 6] The physical cross-linking of TPU takes place between the proton donating N-H of the hard segment and the hard and/or soft segment carbonyl (C=O) groups, which can result in a phase-separated polycrystalline structure.[7, 8] Many thermodynamic and kinetic factors influence microphase separation.[9, 10] These factors include but are not limited to (i) the composition of the soft and hard segments, (ii) the length distribution of the soft and hard segments, (iii) branching and atypical soft segment side-chain content, (iv) crystallization of the hard segment, (v) hard segment mobility in the soft segment matrix, and (vi) processing and annealing conditions. [2, 4, 6, 10-15]

Studies employing infrared and Raman spectroscopies attempt to probe the physical cross-linking of the urethane amines with the proton-accepting C=O groups.[1, 16-21] Depending on the chemistry and processing conditions, these morphologies include a variety of amorphous, semicrystalline, and ordered crystalline microstructures that appear as specific bands in the C=O and N-H stretching envelopes. The carbonyl stretching region, in which these morphologies have a specific band shape, is typically found in the range of 1624-1735 cm^{-1} for polyamides, model polyurethanes, and TPU elastomers. The N-H stretching modes are present in the 3130-3445 cm^{-1} region of infrared and Raman spectra, although in a Raman spectrum, the non-hydrogen bonded (or "free") N-H stretching mode is rarely present or extremely weak because of insufficient polarizability.[22-25] These regions in infrared and Raman spectra have important implications for qualitative and quantitative polyurethane research. For example, in model, hard segment-only polyurethanes, it has been demonstrated that the fraction of "free," disordered, and ordered hydrogen-bonded carbonyl groups can be estimated by measuring the linear changes in intensity and relative proportions of the carbonyl stretching modes as the temperature of the sample cell is varied.[17] In polyester-based TPU elastomers, it is often difficult to resolve the hard and soft segment C=O contributions to the carbonyl stretching envelope due to the existence of at least five active modes (i.e., three corresponding to hard segment morphologies and two to the soft segments).[4]

In spite of the comfortable marriage of chemometrics and spectroscopy, there remains a clear deficit in the literature investigating the role of multivariate analysis in studying the infrared and Raman spectra of TPUs. Multivariate methods such as principal component analysis (PCA) are prime candidates for the investigation of TPU and TPU blend-spectra because they assume that the underlying system of interest contains a set of latent, linearly additive chemical factors (i.e., principal components) derived directly from, in this case, spectroscopic data.[26] For example, the most dominant spectral features acquired in an infrared thermal study is the thermodynamically-controlled band shape and intensity change due to the shift of the hydrogen bonded to "free" C=O and N-H stretching modes with increasing temperature. The linear change in infrared band shape would undoubtedly express itself (i.e., "load") substantially onto the first few principal components (PCs) acquired from a PCA. If distinct, individual Raman active modes would load onto further PCs thus facilitating interpretation. PCA has added benefits of data compression, noise reduction, and data transformation from which the transformed spectra (a.k.a., PC-scores) can be projected onto low-dimensional principal subspaces where their relationships can be visualized.[27]

In this study, it is hypothesized that applying PCA to the micro-Raman spectra of TPU blends will demonstrate a proof-of-concept in which the morphology of TPUs can be more efficiently investigated through Raman spectral features. Specifically, by applying PCA to regions of the micro-Raman spectra that show the most hydrogen-bonding sensitivity, spectra from TPU blends containing identical composition should distinctly cluster as points on PC-scores plots. Principal component regression (PCR) will then assess whether the Raman spectra of TPUs can be used to directly estimate the copolymeric components in each sample blend.[28]

Experimental Section

Sample Preparation

Two polyester-based TPUs were received from Lubrizol, Inc. (Cleveland, USA) each having a Shore A hardness rating of 85 and 92, respectively. Furthermore, each TPU samples were abbreviated EST85 and EST92 and their blend compositions represented as the weight ratio of EST85:EST92. Due to proprietary protections applied to commercial TPUs, we are unable to publish the trade names of the polyurethanes analyzed in this study.

Each TPU was blended in a small quantity, laboratory-scale, mixing molder (Dynisco Polymer Test LMM) at 50 RPMs, 170°C for 7 minutes. The samples were pressed into a circular mold and had a final diameter of 1 inch and thickness of 1/16 inch. Each blended sample had an initial mass of 2.2 g and final mass after molding of approximately 2 g. The loss of 0.2 g was the result of material remainders during pressing and molding. The composition of the samples were 100% EST85, blends of 80:20, 60:40, 50:50, 40:60 and 20:80, and 100% EST92. The samples were then annealed at room temperature (23°C) for 24 hours prior to scanning. A total of four batches of samples were blended. The first three batches were used in the model training phase of PCR and the final batch was used for model validation phase of PCR.

^{13}C NMR analysis: polymer weight fraction estimation

^{13}CNMR spectroscopy was employed for each as-received TPU to identify each chemical constituent present in the polyurethanes, to estimate the mole ratio of the hard and soft segments, and to estimate the fraction of hard and soft segment present in each TPU blend. ^{13}C NMR spectra were acquired using a 300 MHz Bruker AVANCE 300 operating at 75.5 MHz and 25 °C. Each sample was dissolved in hot, d-DMSO (99.9%, Cambridge Isotope Laboratories) for 2 hrs. At least 500 scans were signal averaged to ensure adequate signal-to-noise ratio. All further processing and analysis was performed using SpinWorks 3.1.8 beta.

Raman Spectroscopy: Instrumentation and Procedures

Confocal Raman spectroscopy was performed using the WITec™ alpha300 R Raman instrument. A 100 mW, frequency doubled, Nd:YAG (λ = 532 nm) laser was utilized and focused using a 20× Nikon objective (NA = 0.4, WD = 3.9 mm) with a spot size of ~10 μm. A Thorlabs™ PM100A optical power meter was used to measure the incident power. The laser power was attenuated to reduce fluorescence background and achieve the highest signal-to-noise ratio (SNR). This gave a final incident sample irradiation power of 3.76 mW. Single spectrum "spot" scans were acquired using the alpha300 R via a UHT-300 spectrometer (grating = 600 grooves/mm, entrance aperture = 50 μm), with the Andor™ DU970N-BV, 1600 × 200 pixel, CCD array detector. The sample surface was brought into optical focus and the fluorescence background reduced by exposing for 30 seconds prior to

acquiring each spectrum. Raman spectra were acquired with an integration time of 3 seconds averaged over 8 spectra for 10 random spots on each sample. The first three batches of TPU blend samples contained a total of 210 spectra for the model training spectra (i.e., 10 scans × 7 sample blends × 3 batches). Thirty scans per blend type were necessary to offset the possible effects (on average) of heterogeneous, local-level mixing and to reduce the influence of instrument noise on PCA. The final batch of TPU samples was scanned at 5 random spots per sample type, for a total of 35 Raman spectra for the PCR model validation set.

Processing of TPU Blend Spectra

WITec™ Project software was used to perform fluorescence background subtraction and cosmic ray filtering of Raman spectra. The background of each spectrum was approximated by a 9^{th}-order polynomial fit and removed. Additional outliers were removed using leverage statistics where observations with a leverage value higher than three times the average value were removed.[29] The final number of spectra used in the training and validation set was 193 and 34, respectively.

PCA is not scale invariant and highly susceptible to inconsistently removed backgrounds.[30, 31] Additional fine-tuning of spectra was performed in Matlab (v. 7.5 R2007b) with a user-created program that subtracted a linear baseline from selected spectral regions. This ensured that differences in baseline intensities did not artificially influence the clustering of spectra on PC-scores plots.[32] Additionally, each spectrum was area-normalized by dividing each intensity measurement by its spectral average intensity. This had the "effect of equalizing the spectral area," [33, 34] helping to eliminate the influence of spectral intensity on PCA and helping to emphasize the differences between spectral features.

All spectra were loaded into an $m \times n$ matrix in Matlab, where m rows corresponded to spectra and the n columns to spectral intensities at a given wavenumber (rel. cm^{-1}). The n^{th} dimension of the data matrix was bounded and reduced to incorporate only those variables comprising the morphologically interesting, Raman active, TPU group frequencies. Three regions were analyzed and included the carbonyl stretching envelope (1681-1764 cm^{-1}), the CH_2 deformations and Amide II combination band (1370-1573 cm^{-1}), and the N-H stretching envelope (3184-3463 cm^{-1}).

Principal Component Analysis (PCA)

PCA was performed on each reduced data matrix using the singular value decomposition (SVD) algorithm in Matlab of the form

$$[X] = [U][L][V]^T \quad \text{Equation 4.1}$$

where *[X]* is the data matrix, *[U]* is the matrix of standardized eigenvectors of the matrix $[X][X]^T$, *[L]* is the matrix of singular values, and *[V]* are the eigenvectors of $[X]^T[X]$.[27] The eigenvector matrix *[V]* contains the orthogonal "loadings" used in PCA and, mathematically, corresponds to the set of coefficients of linear combination that define the principal components.[30] In this case, the matrix *[V]* is used as a linear transformation matrix in which the original spectra are defined on a new, variance-maximized basis and take the form $[Z] = [X][V]$.[35] The column vectors of *[Z]* resulting from the action of *[V]* on *[X]* define the PC-scores.

The performance of two types of PCA was assessed, each being referred to as covariance (COV-) PCA and correlation (CORR-) PCA. In COV-PCA, the data matrix was simply column-centered by subtracting each variable by their column mean prior to SVD. For CORR-PCA, the data matrix was centered and each column scaled to unit variance.[36] Although all variables were on the same scale, both PCAs were analyzed to assess the best clustering of spectra on PC-scores plots. Leave-one-out cross-validation was performed to determine the rank of the data that corresponds to the precise number of chemically significant factors present in the data.[37, 38]

Principal Component Regression (PCR)

In general, PCR is a linear least-squares regression technique used to correlate the PC-scores extracted from PCA with an external material or chemical property.[29] In our case, seventeen outliers were pruned from the model training set leaving a 193-element vector containing the measured weight fraction of each TPU blend. Employing the hard-to-soft segment ratio estimated using ^{13}CNMR, we could directly estimate the fraction of the hard and soft segments present in each blend using elementary material balances. The final 193-element vector of the hard segment mole fraction was used in the PCR. Specifically, this vector was regressed against the corresponding PC-scores taken from the reduced data matrix of Raman band intensities located in the carbonyl, hard segment aromatic stretch of 4,4'-MDI, Amide II stretching, and CH_2 deformation regions of the Raman spectra. Cross-

validation was again employed to estimate the number of PCs needed to model the data. The details of PCR and the seemingly redundant, but necessary, application of cross-validation are outlined elsewhere.[27, 39]

The performance of PCR was evaluated for the training set by calculating the root-mean-squared error of calibration (RMSEC) and the coefficient of determination (R^2). The predictive performance of the trained model was further assessed using the independent validation set in which the root-mean-squared error of validation (RMSEV) and coefficient of determination (Q^2) were calculated. For this study, PCR model building was an iterative procedure that involved changing the degree of preprocessing, e.g., scaling, band smoothing and filtering, etc., until the best PCR model was attained. The final model was chosen based on (1) its ability to deal with nonlinearity in the model residuals, (2) minimize the number of PCs required to model the data, (3) optimize the R^2 and Q^2 statistics, and (4) minimize the RMSEC and RMSEV.

Results and Discussion

Qualitative analysis revealed that the EST92 hard segment was synthesized using 4,4'-methylenediphenyl diisocyanate (MDI) and 1,4, butane diol (BD) precursors while the soft segment generated using a poly(butylene adipate) (PBA) precursor.[40] Due to the low signal-to-noise ratio of the EST85 NMR spectrum, it was only possible to infer that the soft segment consisted of PBA and the hard segment was chain extended with 1,4 BD. Regardless of the ambiguity surrounding the diisocyanate block of EST85, Raman bands relevant to hard segment chemistry indicated the a 4,4'-MDI precursor was also synthesized to form the EST85 hard segment.

^{13}CNMR was used to estimate the mass fraction of the hard and soft segments. By exploiting the fact that the spin-lattice relaxation time constants of the interior methylene carbons of 1,4-BD and PBA are identical, the integral-area of these chemical shifts was used to directly estimate the ratio of the hard and soft segments present in each polymer.[41] This technique resulted in a hard to soft segment estimate of 0.41 and 0.56 for EST85 and EST92, respectively.

Figure 4.1 shows the group-averaged Raman spectra for pure EST85 and EST92, respectively. These spectra were generated by averaging all single micro-Raman spectral

scans for each pure EST85 and EST92 sample. Their visual similarity is not surprising considering that each TPU has identical hard and soft segments where only the hard-to-soft segment mole fractions differ between pure species. Therefore, the scrutiny of bands sensitive to hydrogen-bonding behavior is warranted. Additionally, analyzing these mixture-spectra helps to consolidate the local morphology effects (i.e., heterogeneous microscale behavior) of each sample-scan into a spectrum that better represents a spatially averaged TPU. This fostered comparison between the Raman spectra and other analytical instruments including DSC and ^{13}CNMR.

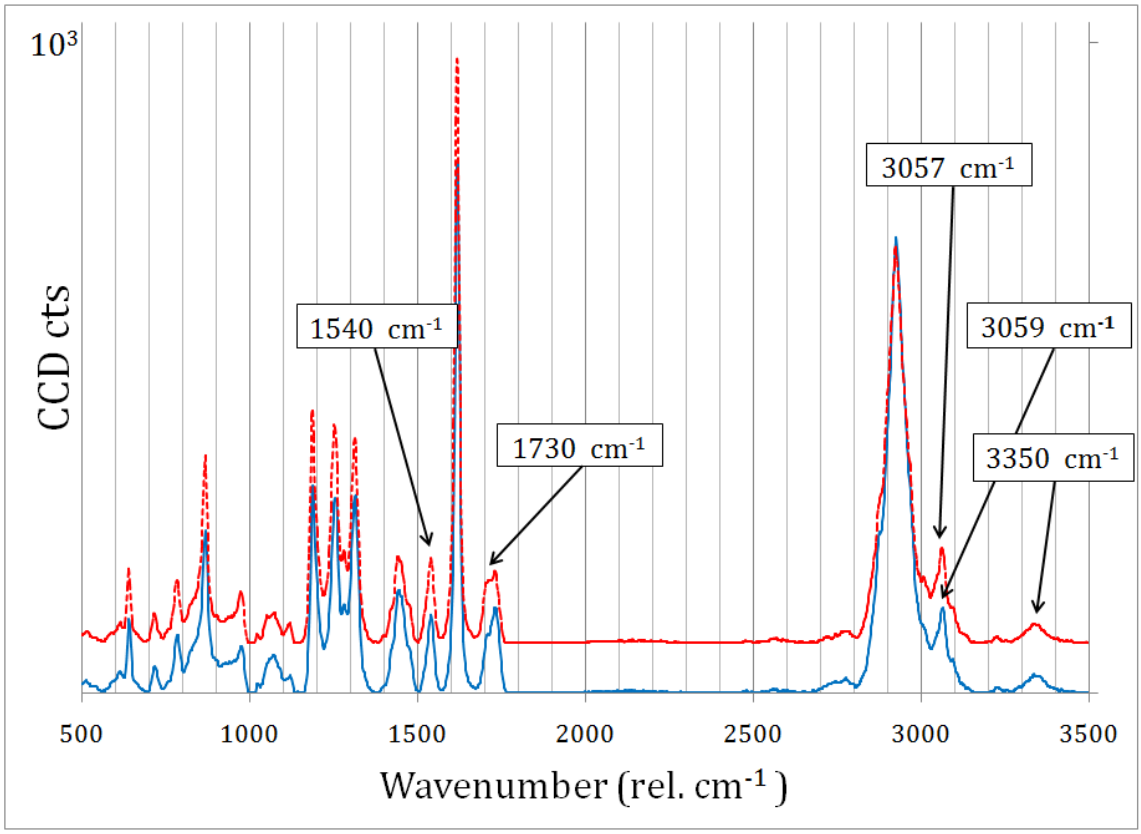


Figure 4.1: Raman spectrum of EST85 (blue) and EST92 (red) from 500 cm^{-1}-3500 cm^{-1}. Literature is typically focused on Raman vibrations active at 1540 cm^{-1}, 1730 cm^{-1}, and 3350 cm^{-1}. Bands at 3057 cm^{-1} and 3059 cm^{-1} were labeled to illustrate how spectroscopically similar the two polymers are prior to mixing.

Bands located in the N-H stretching region were too weak to adequately describe sample-to-sample variation using PCA. Naturally, no trace of useful clustering on PC-scores plots was identified. When the Amide II band (1540 cm^{-1}) was analyzed with the CH_2 deformations (1380-1500 cm^{-1}), unique clustering of spectra on PC-scores plots was

observed. Although interesting, the chemical factors governing the activity of the Amide II band are very complex, thus making simple interpretation untenable.[17] Specifically, the Amide II band is a mixed mode primarily comprised of the C-N stretch and N-H in-plane bend.[23] Using vibrational spectra to infer the magnitude of each active contribution is difficult, if not impossible, when hydrogen-bonding is evident. To a practical end, the wavelengths comprising the Amide II and CH_2 deformations result in a much simpler and accurate PCR than those wavelengths comprising the carbonyl stretching envelope alone. The resulting calibration and reasons for this will be described in extensive detail below.

Figure 4.2 shows the average mixture-spectrum of each TPU blend corresponding to the carbonyl stretching envelope. Peaks suspected of being either hydrogen-bonded or "free" carbonyl stretching frequencies are labeled. According to literature, bands appearing at 1703 cm^{-1}, 1718 cm^{-1}, and 1732 cm^{-1} likely correspond to ordered crystalline hard segment domains, disordered amorphous hard segment domains, and the "free" carbonyl stretching frequencies of the hard and soft segments, respectively.[6, 7, 17, 24] Additional bands at 1708 cm^{-1} and 1710 cm^{-1} correspond to hydrogen-bonded C=O modes that defy clear identification due to the complexity of the present system. Although, Teo and Colleagues have attributed the hydrogen-bonded urethane stretch at ~1707 cm^{-1} to disordered hard segments of poly(urethane-urea)s dispersed in the soft segments phase.[19] As stated in the introduction, there are at least five (unresolved) carbonyl stretching frequencies consolidated within this region for each individual polyester-based TPU corresponding to a unique hard segment-to-hard segment, hard segment to soft segment, or "free" carbonyl vibrations.[4] Hence, it is reasonable that Raman shifts indicative of similar coordination vibrate at slightly different frequencies.

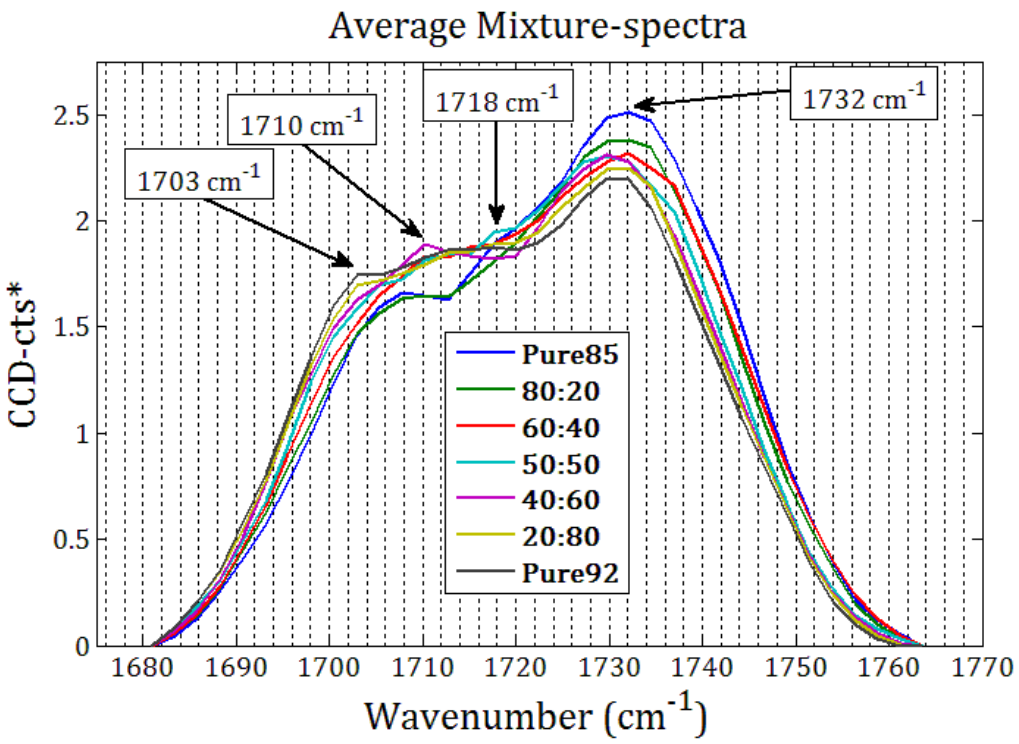


Figure 4.2: Area-normalized, TPU mixture-spectra for the carbonyl stretching envelope. Although complex, it appears that as the concentration of EST92 increases in the sample-mixture the intensity of the "free" carbonyl band at ~1732 cm^{-1} decreases (and possibly shifts) while the intensities of all hydrogen-bonding modes increase (1703 cm^{-1}, 1710 cm^{-1}, 1718 cm^{-1}). The asterisk on y-axis label denotes scaling.

Regardless of the composition of the TPU blends, the most intense stretching always occurred at the "free" carbonyl mode (1732 cm^{-1}). Additionally, Figure 4.2 shows that the peak at 1708 cm^{-1} is distinct for EST85 and the 80:20 blend while absent or not resolvable for EST92. For blends containing a higher composition of EST92, a shoulder on the hydrogen-bonded envelope appears at 1703 cm^{-1}. The disproportionate intensity of this shoulder relative to the other mixture-spectra indicates that EST92 has better self-associating hard segment-to-hard segment compatibility. This is clearly attributable to the higher fraction of hard segment present in EST92 relative to EST85. This implies that a larger fraction of the total carbonyl groups present in EST92 is ordered relative to EST85 thus explaining the greater hardness of as-received EST92. Interestingly, the peak at 1718 cm^{-1} is the most intense for the 50:50 blend and weak or not resolved in pure EST85 and EST92 spectra, respectively. This signifies that hydrogen-bonding is occurring across polymers as well as within each individual polymer. An additional shoulder at 1710 cm^{-1} and a possible

wavenumber shift in the "free" carbonyl stretching mode for the 40:60, 50:50, and 60:40 mixture-spectra support the likelihood of interactions across polymers as well.

Even at this preliminary stage of analysis some concentration-contingent trends can be seen. For one, the "free" carbonyl stretch (1732 cm^{-1}) becomes progressively less intense as the blend concentration of pure EST85 decreases. Conversely, the stretching envelope corresponding to the hydrogen-bonded modes becomes more intense on average as the concentration of EST85 decreases. This trend confirms the expectation that, for this system at least, TPU hardness is linked to a greater degree of hydrogen-bonding between N-H and C=O.

PCA was initiated by performing cross-validation. From this, it was determined that four significant PCs for COV-PCA and three for CORR-PCA were the main sources of useful variation in the spectral matrix. Pareto/scree plots of the individual and cumulative percent of variance explained by the first ten PCs were reviewed (not shown). The first four PCs explain roughly 90% of the variance for COV-PCA and the first three PCs explain roughly 75% for CORR-PCA. PC1 is very much dominant in both cases. Loading and PC-score's clustering of CORR- and COV-PCA were very similar. Therefore, only the results of CORR-PCA will be discussed in detail. These results are generally anticipated. Specifically, the first eigenvalue, which represents the proportion of explained variance of PC1, is very dominant in each PCA. Although, in general, scaling changes the meaning and type of information contained on a principal component,[30] it appears that the same composition-dependent information was isolated to a similar degree on the significant PCs. This makes further comparison between column-centered and autoscaled models redundant.

PC-scores plots were constructed using wavenumbers comprising the carbonyl stretching envelope. Figure 4.3 illustrates a projection of these bands onto the first two PCs. The larger black datum in Figure 4.3 represent the average score values for each TPU mixture-spectrum. These spots are located where the average spectra (as shown in Figure 4.2) would lie if they were reconstructed using only the first two PCs. Remarkably, there is a clear, almost one-dimensional distribution of mixture-spectra ranging from pure EST85 to pure EST92 in PC1 with significant point-to-point scatter in the PC2 direction. The almost linear distribution along the PC1 axis reflects the aforementioned increase and decrease in intensity of the hydrogen-bonded and "free" carbonyls as we move to a high EST92

containing blend, respectively. Additionally, there is overlap between spectra of the 20:80 and 40:60 blends. This suggests either an influence of imperfect mixing or likely a true morphological similarity between the two groups. Additional detailed studies employing small-angle x-ray scattering could be used to suggest a proper link between spectroscopic behavior and morphology, particularly phase separation.[11, 42-44]

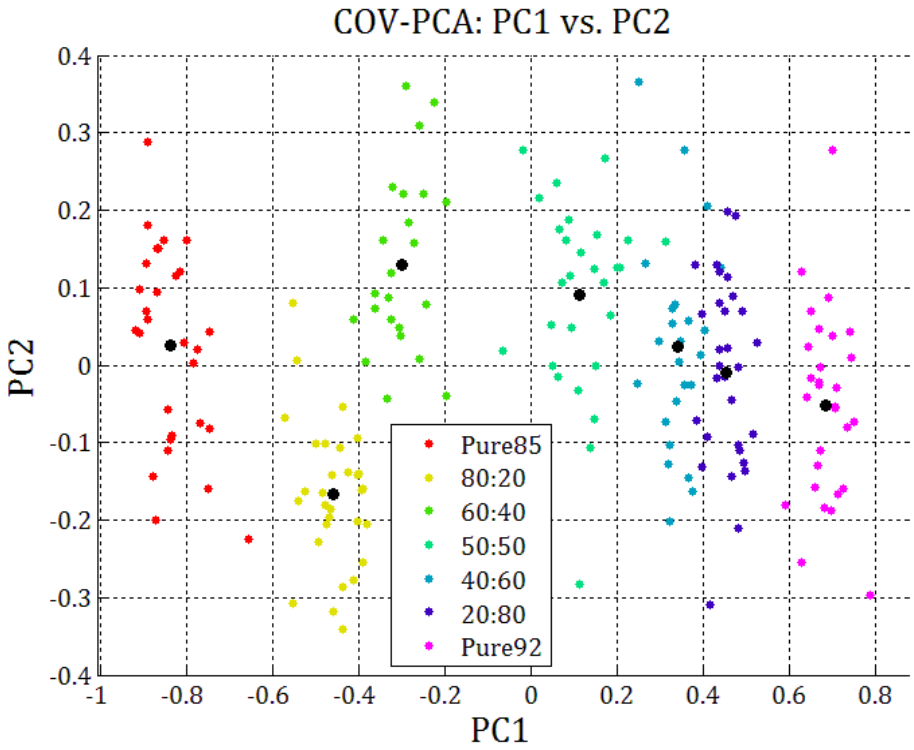


Figure 4.3: TPU mixture-spectra projected onto the first two PCs for CORR-PCA. A near one-dimensional distribution of points on the PC1 axis exists for each blend-spectrum. This is likely due to the simultaneous decrease and increase in the "free" carbonyl and the hydrogen-bonded C=O region as the composition of EST85 decreases and EST92 increases, respectively.

Figure 4.4 shows the results of reconstructing spectra using the first PC. This plot supplements the more traditional (but sometimes less interpretable) loadings plots frequently encountered PCA-spectroscopic literature. Figure 4.4 presents a surprisingly lucid picture of the band variation present in the carbonyl stretching envelope. For example, Figure 4.4 shows that when the average mixture-spectra are reconstructed using only PC1, the spectra are dominated by a nearly linear increase and decrease in the intensities of the hydrogen-bonded and "free" carbonyl stretching envelopes contingent upon the weight fraction of EST92, respectively. Furthermore, the loadings, marked on the secondary axis (blue lines), show that PC1 is positively correlated with the hydrogen-bonded stretching envelope and

negatively correlated with the "free" carbonyl envelope. Chemically, this implies that score clustering on the PC1 axis is largely dominated by peak narrowing, broadening, and intensity changes (on average) as a result of the changes in TPU blend composition.

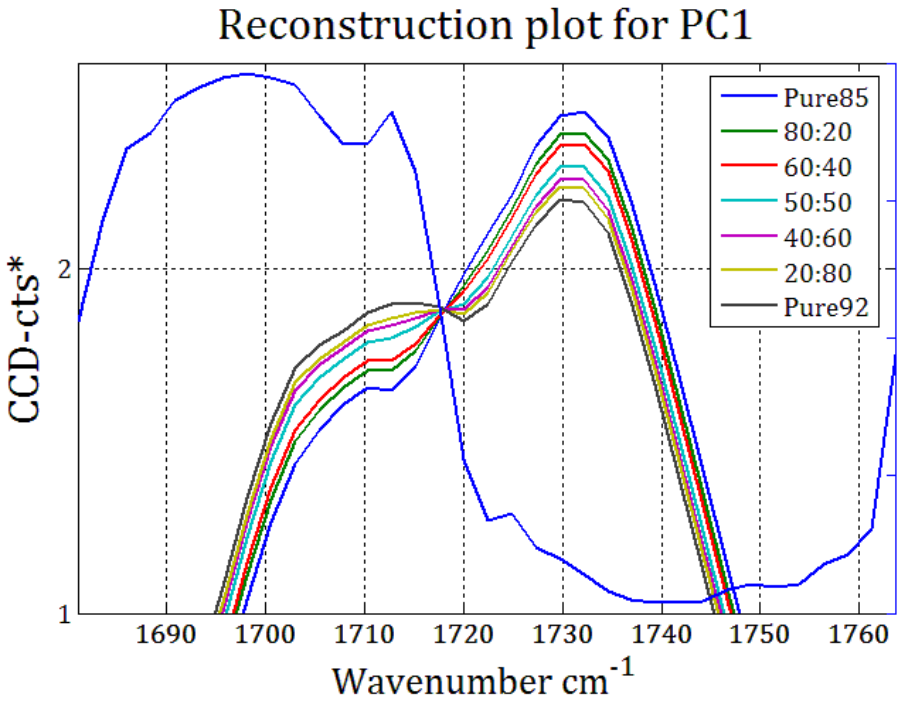


Figure 4.4: Single PC-Reconstruction plot for CORR-PCA. The PC1 loadings (blue) exhibit large-positive and large-negative correlations with wavenumbers on either side of the carbonyl stretching envelope. This has some influence on the meaning of PC1 as a chemical factor and the predictive performance of PCR.

The marked absence of the 1718 cm^{-1} band appears as an "isosbestic-like" crossover point in Figure 4.4. For spectral-reconstruction using PC2 only (not shown), the 1718 cm^{-1} peak appears as the dominant spectral feature. Remember, this confirms the preliminary observation that the most amorphous polymer is likely the 50:50 blend, suggesting that cross-polymer solvation is the strongest for this blend, on average. As we will see, this does not exclude the possibility of short or long-range ordering in the 50:50 blend but only indicates a reduced degree of coordination. It is important to mention that the term isosbestic has been borrowed from absorbance spectroscopy to describe the point at which the spectral behavior of two mixed species cross at a given wavelength and therefore have an identical molar absorptivity (ε).[45] Since in Raman scattering, any discussion of molar absorptivity is not relevant, these points should be referred to, more generically, as isopoints when encountered.

Because cross-validation isolated three significant factors, it is customary to review spectra projected onto the remaining PCs. Figure 4.5 shows the result of projecting spectra onto PC1 and PC3. Notice that the spectra cluster very distinctly onto these two components. Again, the nearly one dimensional clustering is almost due entirely to the in dominance of PC1.

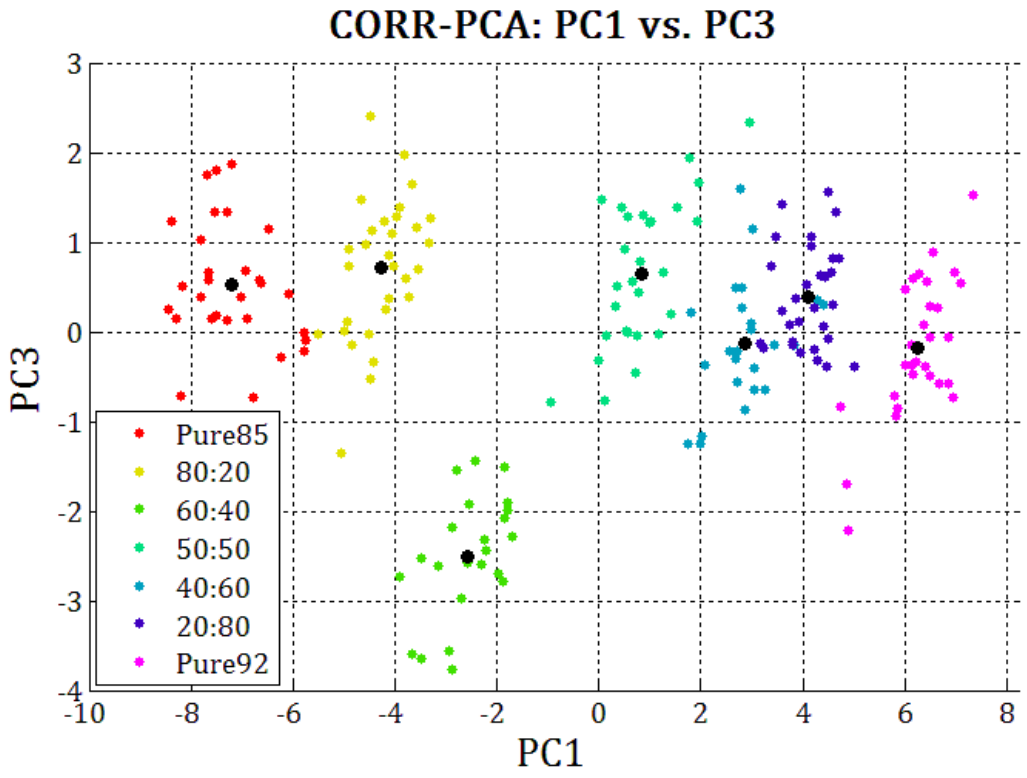


Figure 4.5: TPU mixture-spectra of carbonyl stretching region plotted onto PC1 and PC3 subspace for CORR-PCA. Clustering is similar to the PC1 and PC2 plots with the exception of the 60:40 blend.

The obvious "outlying" 60:40 blend spectra cluster signficantly lower than the other blends on PC3. The clustering of the 60:40 blend in the PC1-PC3 subspace is related to the unique loading of the 1710 cm^{-1}, 1715 cm^{-1}, and 1732 cm^{-1} peaks onto PC3 as shown in figure 4.6 (below). The 60:40 blend appears to have a much lower loading of the 1732 cm^{-1} band as well indicating a probable morphology difference relative to the other blends.

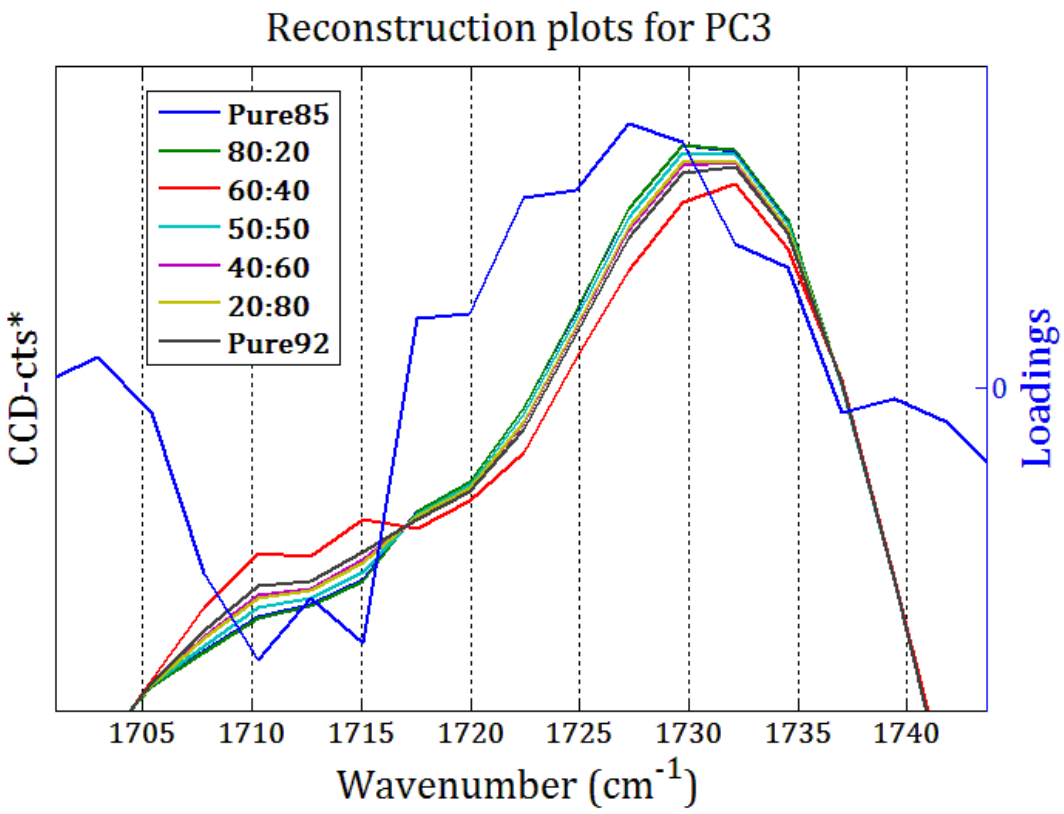


Figure 4.6: Spectrum reconstruction plot using the third PC. The 60:40 blend exhibits reduced intensity in the "free" carbonyl band and uncharacteristically large amplitude features in the hydrogen-bonding region of the carbonyl envelope. The complimentary loadings (blue) also emphasis the strong, group-dependent influence of the 60:40 blend, particularly for the 1710 cm^{-1}, 1715cm^{-1}, and left hand side of the "free" carbonyl band.

Figure 4.7 presents DSC curves for each blend. To elaborate, endotherm I (50-70 °C),[8] is common to all samples, endotherm II (120-175 °C) is only well represented in EST92 and the 50:50 blend, and the large endotherm at 96 °C is specific to the 60:40 blend. Morphologically, endotherm I represents the melting of hard segments exhibiting short-range order for room temperature annealed polyurethanes.[8] Essentially, we should expect that the existence of short-range order in EST85 signals that all TPUs will show similar melt behavior in this region (which is indeed the case). Endotherm II indicates the presence of long-range hard segment ordering.[6] As mentioned previously, Figure 4.4 provides spectroscopic support for an average increase in segment interactions via hydrogen-bonding where the areal increase of the hydrogen-bonded carbonyl envelope is discernible as a function of increasing EST92. Again, PC1 only represents an average increase in the magnitude of hydrogen-bonding, with no discrimination as to orientation or segment

interaction behavior. Using DSC, we can explain this increase in the hydrogen-bonding envelope as an overall increase in order, where the presence of endotherms I and II in concentrated EST92 samples provides qualitative assurance. Unfortunately, it is impossible to infer the degree of soft segment, carbonyl involvement from DSC and Raman spectra.

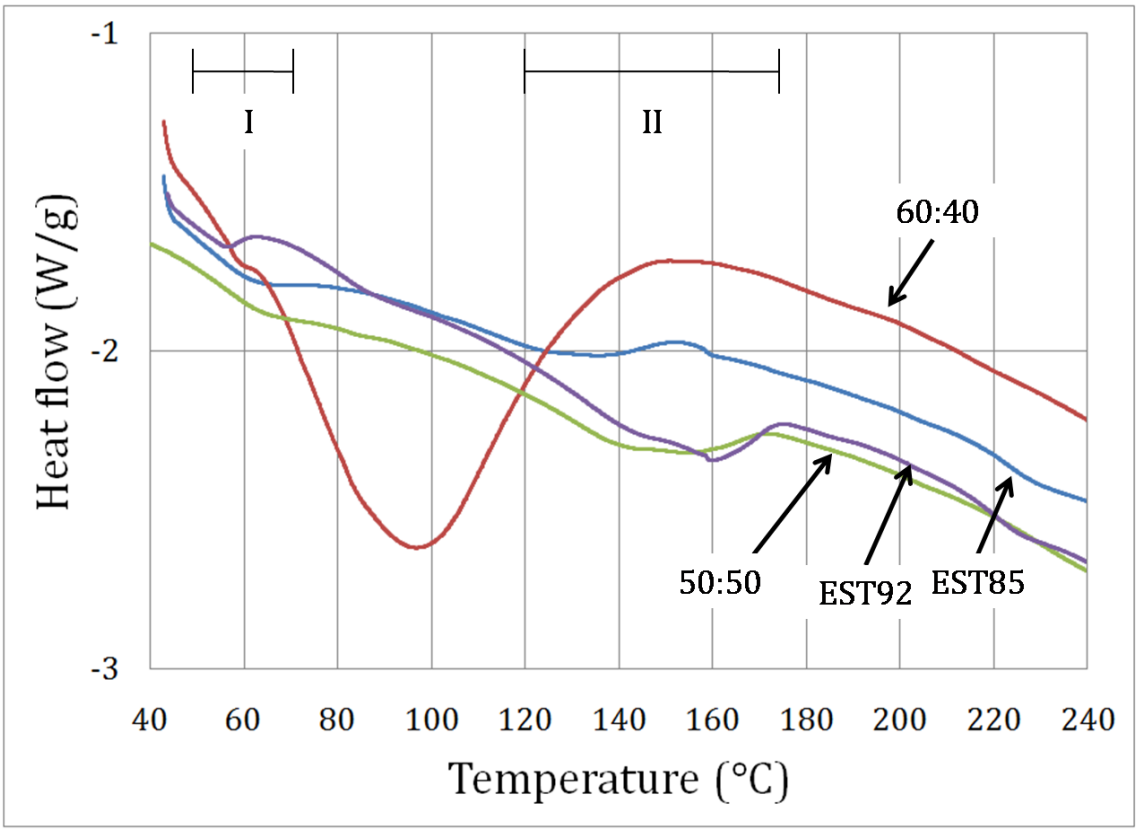


Figure 4.7: DSC curves for select TPU-blends that exhibited important feature variations on PC-scores plots. Endotherm I (50-70 °C) is common to all TPU and represents short-range, hard segment ordering while endotherm II (120-175 °C) corresponds to long range ordering of the hard segments. The 60:40 blend shows distinct melt behavior as indicated by the large endotherm at 96°C.

Returning to Figure 4.3 with the DSC behavior in mind, we observe large, non-spherical, point-to-point scatter on the PC2 direction within each TPU-blend. Notably, this scatter appears to roughly increase in magnitude as the concentration of EST92 increases. It is tempting to infer that this scatter indicates that each TPU is *a priori* phase separated for the following reason. Essentially, when taking into account the small laser spot size, the population of ordered domains residing under the beam-probe should vary from spot-to-spot in each sample-scan where the magnitude of this variation is larger for TPUs exhibiting greater order. Specifically, blends exhibiting increased ordering (i.e., blends with more

EST92) show an increased magnitude of scatter on PC2. The DSC curves shown in Figure 4.6 exclude the possibility of ordered micro-phase separation in our blends because the hard segment endotherm III corresponding to crystalline domains is absent. Regardless, it appears reasonable that sub-micron scale heterogeneities are present in most blends for two reasons. Specifically, as-received EST92 is optically turbid and samples containing at least 50 wt.% EST92 show endotherm II, the implications of which have already been discussed. Minding spatial-resolution limitations, the observation that morphology effects are identified and isolated in multivariate analysis suggests that such models are natural supplements for micro-Raman hyperspectral imaging, i.e., PCA could easily be employed to assist in micro-Raman mapping for polyurethanes.

The endotherm unique to the 60:40 blend is initially baffling in that it resides just outside the traditional bounds of endotherm I or II. In fact, it has been observed that endotherm I can move "upscale" as a function of annealing temperature to an intermediate region such as the one observed here.[8] However, endotherm I is present in the 60:40 blend in the expected range (56 °C) and no annealing was performed on the present sample set. Thus, a certain ambiguity is inherent in trying to explain the location and strength of the observed endotherm.

Regardless, Figure 4.6 suggests that an additional hydrogen-bonding interaction is occurring for the 60:40 blend (red) followed by the 80:20 (green) to a lesser degree. It is possible that these active vibrations signal a unique cross-polymer interaction with distinct ordering (likely short range due to its proximity to endotherm I). This theory would explain the presence of an additional endotherm in the 60:40 blend. Furthermore, Figure 4.6 shows that the pure polymers and 50:50 blend show identical line-shapes upon reconstruction using PC3. This spectroscopic behavior tracks closely with the 'normal' melt behavior of the pure polyurethanes and 50:50 blend (i.e., absence of a large, ambiguous endotherm). This nonlinear, blend-specific behavior will significantly challenge the interpretation of the following PCR.

The results from four separate PCR calibrations and complimentary predictions are shown in Table I. Initially, only the carbonyl stretching envelope was used to correlate the hard segment fraction to Raman spectra (see, row 1). However, a heuristic approach was adopted whereby the CH_2 deformations (1380-1500 cm^{-1}), Amide II (1540 cm^{-1}), and

aromatic (1617 cm^{-1}) bands were included in PCR. This led to a reduced factor space in all cases with comparable results. For each spectral range tested, PCR was repeated multiple times using a variety of scaling regimes and smoothing filters. Finally, it was determined that area-normalizing each Raman intensity by the average of all intensities in the spectrum,[46] autoscaling variables, and employing a nine-point, 2nd-order Savitzky-Golay smoothing or derivative filter achieved the best PCR model for predicting hard segment content.

Table 4.1: Results of four separate calibrations and predictions using PCR. In general, the results are remarkably similar irrespective of the spectral range used. The key differences resided in the size of the factor space used to construct the model.

Spectral Range (cm^{-1})	Bands	Filter*	PC	RMSECV ($x10^3$)	RMSEC ($x10^3$)	ε_c(%)	R^2	RMSEV ($x10^3$)	ε_p(%)	Q^2
1681-1736	ν(C=O)	SG(0,2,9)	11	2.96	2.82	0.87	0.985	5.06	1.56	0.940
1384-1573	δ(CH_2) Amide II	SG(0,2,9)	3**	4.15	4.12	1.26	0.966	4.50	1.38	0.957
1350-1764	δ(CH_2) Amide II ν(Ar) ν(C=O)	SG(1,2,9)	4***	3.46	3.41	1.05	0.977	4.49	1.36	0.959
1377-1503, 1578-1676	δ(CH_2) ν(Ar)	SG(1,2,9)	3**	4.90	4.85	1.49	0.953	4.54	1.40	0.958

* - SG(#,#,#) - Savitzky-Golay filter of given derivative order, polynomial order, and window size

** - RMSECV plot levels-off at specified factor

*** - RMSECV achieves a local minimum at specified factor

Standard error measures (e.g., RMSEC) and error percentages (ε_c,ε_p) are nearly identical across trials, indicating that when preprocessed appropriately the micro-Raman spectra of these TPU-blends are fitted reasonably well using linear models. Although the R^2 and Q^2 values are not as large as desired, the accuracy and reproducibility of the results is somewhat surprising considering the influence of nonlinear composition-dependent effects of each polymer blend. In particular, employing the CH_2 deformation and Amide II band in the PCR resulted in the best model as judged by the low RMSEV, RMSEC, RMSECV as well as, and most importantly, the smallest factor space. Wavelengths comprising the CH_2

deformations and aromatic stretch performed similiarly on standard error measures and R^2 statistics. However, the analysis of residual plots showed substaintially more blend-specific behavior as compared to the other two-band calibration. Therefore, the former set of wavelengths was deemed the most valid of the two.

Figure 4.8 shows fit residuals for wavelengths pertaining to the best performing, two-band regression. Figure 4.8 illustrates that even the best performer showed substantial nonlinear behavior for certain blend-groups (e.g., pure EST92 in this case). Initially, we see a clear improvement in model accuracy and randomness in the residuals. Beyond the third factor, it is unclear whether the additional PCs substaintially mitagate blend-specific effects. In fact, the minor improvement in the error statistics and non-verifiable improvement in residual randomness illustrates the folly of placing too much faith in standardized error measures when nonlinear behavior is present.

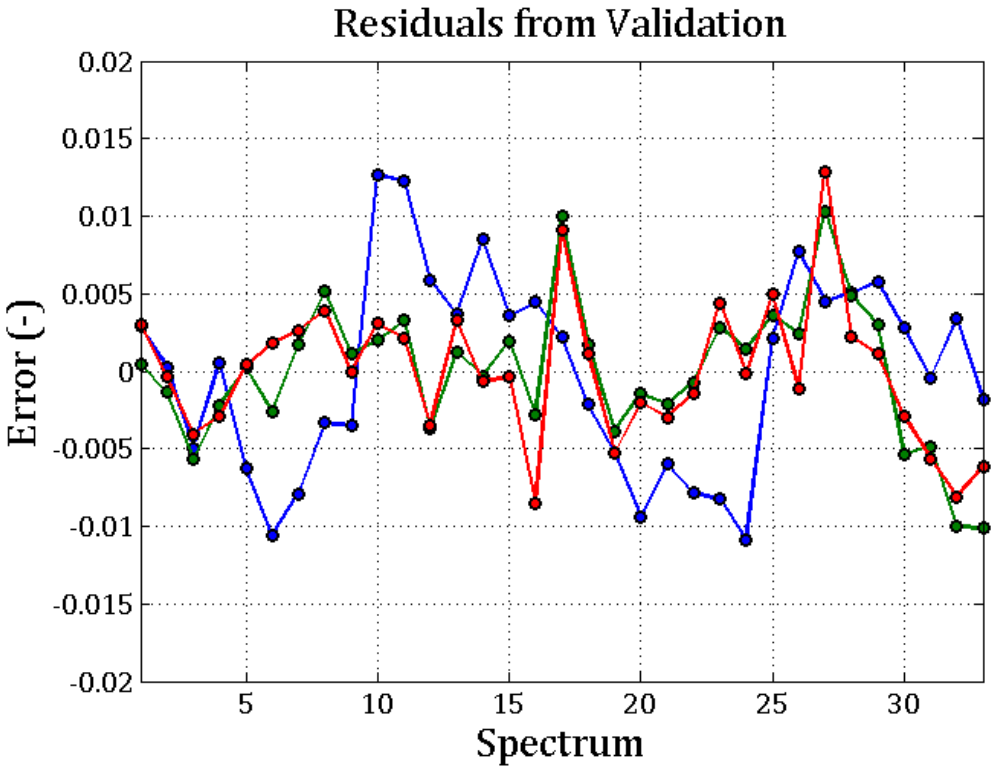


Figure 4.8: Residual plot for validation set spectra. A 1-PC model (blue) shows clear blend-specific nonlinear behavior while a 3-PC model (red) shows marked improvement for low-EST92 containing blends (< 17) with adverse clustering behavior for the high-EST92 samples (31-35). Notably, the 13 factor model (green) shows that increasing the size of model tends to reduce the standard error but fails to ameliorate most nonlinear behavior.

Next, in an effort to combat the nonlinearities, we removed the most egregiously offending EST92 mixture-spectra and repeated the regression. Remarkably, this had a negative impact on the calibration error and R^2 statistics (RMSEC = 0.0072, R^2 = 0.8530) while the prediction error and Q^2 statistics (RMSEV = 0.0036, Q^2 = 0.9695) improved, *ceteris paribus*. In both cases, blend-specific anomalies appeared in the residuals for the training and validation data albeit for different blends. These, as well as Figure 4.9 (below), highlight that the better-than-expected results achieved for the PCR are somewhat incidental, i.e., the group-specific nonlinearities tend to balance each other out overall thereby skewing error measures. In the interest of simplicity, all blend spectra (with the exception of outliers) remained in the PCR where visual approval of random residuals were priveledged over a completely minimized RMSECV.

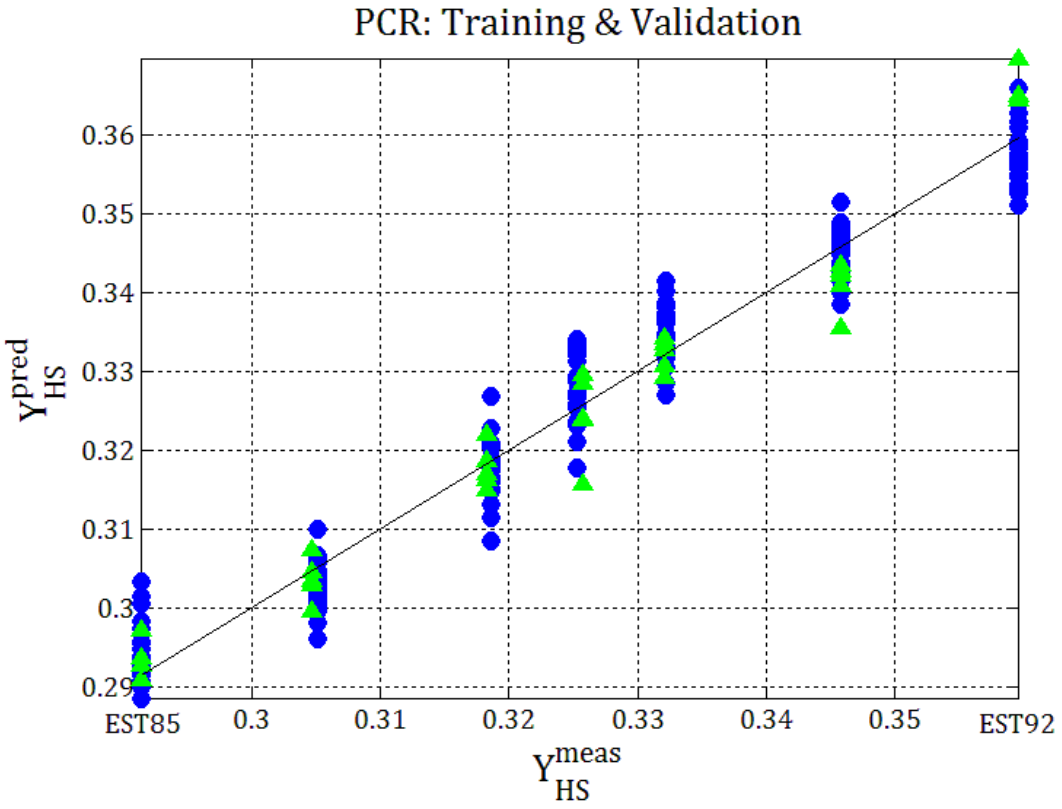


Figure 4.9: Calibrated (blue) and predicted (green) versus measured hard segment fraction for the best performing two-band PCR model. Specifically, the CH_2 deformations and Amide II band were used as predictor variables in the PCR. For the EST92 and 20:80 validation spectra in particular, we can see that blend-specific behavior is balanced thereby falsely improving the RMSEV.

Although variable-scaling is a major factor in a PCR, our results were achieved to the credit of a fairly aggressive (i.e., large window, low-order) Savitzky-Golay filtering. For example, in the case of the carbonyl region, filtering played a significant role in convolving the unresolved, nonlinear features (e.g., shoulder at 1703 cm^{-1}) into the broader stretching envelope. In general, it is unfavorable to reduce the resolution at which a vibrational spectrum is measured much below the full-width at half-height of the narrowest band. However, convolving the morphologically interesting spectral features (e.g., 1718 cm^{-1} band) with the broad, hydrogen-bonded carbonyl stretching envelope was paramount to successful calibration, i.e., the sacrifice in resolution led to a more closely linear model and a smaller factor space for each spectral range tested. Or simply, smoothing resulted in more variance being allocated to the first PC, eliminating many nuanced structural effects related to hydrogen-bonding. Regardless, eleven PCs were still required correlate the carbonyl stretching behavior to hard segment content.

A potential conceptual difficulty arises when attempting to rationalize how PCR operates given reduced band resolution. Specifically, filtering influences resolution while still leaving the PCR apparently teeming with PCs. Particularly, the number of resolution elements (after smoothing) is smaller than the size of the factor space identified by cross-validation. This empirical outcome militates against intuition, where band resolution is ultimately constrained by the number of resolution elements and should *a fortiori* constrain the possiblity of isolating chemical effects larger than the sum of all resolution elements in the carbonyl stretching envelope.

Recognize that for homogeneous and non-interacting chemical mixtures, the number of factors isolated using PCA should correspond exactly to the number of species present.[47] Intuitively, the existence or gradual disappearance of each species is confirmed via clear resolution changes in relevant spectral bands as a function of composition. In this ideal case, PCA treats each species' band resolution changes (across experimental observations) as independent, linearly additive sources of variation.[26] Regardless, reality often dictates that even perfectly mixed and non-interacting chemical measurements suffer from a number of practical shortfalls thereby inflating the factors space (as determined by cross-validation); the most common culprits being noise, spectral artifacts (e.g., spectrometer drift[32] or cosmic rays), and inadequately corrected baselines.

To address these concerns one-by-one, we note that PCA is traditionally employed for noise reduction and therefore noise effects should rarely lead to additional factors given adequate signal-to-noise ratio. Although not quantified, the signal-to-noise ratio of our spectra were high prior to smoothing suggesting that this would not negatively impact the regression (see Figure 4.1). In the absence of nonlinear behavior, only spectral artifacts, poor baseline correction, and an overzealous adherence to cross-validation can lead to factor space inflation. Briefly, neither artifact nor baseline effects are probable because (1) manual baseline correction was performed twice, (2) major artifacts were screened visually, (3) less obvious artifacts were screened using PC-score plots, (4) overtly influential observations were removed via leverage statistics, and finally (5) spectrometer drift was countered by continuously recentering the Rayleigh peak to 0 cm^{-1} during acquisition.

The propensity for cross-validation to overfit multivariate models has been reported.[48, 49] As emphasized for the best performing, two-band PCR, we relegated conclusions derived from cross-validation to the status of practical guide as opposed to the purveyor of absolute truth. Again, Table I should alert the reader to the prudence of this approach. Notably, we can see that when the wavenumbers comprising hydrogen-bonding insensitive bands are chosen as predictors the number of PCs needed to model the data decreases sharply. A clear minimum was observed on a plot of RMSECV versus PCs at eleven factors for the carbonyl envelope and likely led to PCR overfitting the data. For TPUs, as well as other interacting copolymer blends, the use of wavenumbers corresponding to inert molecular vibrations, as opposed to those representing nuanced morphology effects, is warrented. Note, it is possible to temper the tendency for cross-validation to overfit by employing further statistical aids such as Monte Carlos approaches[48] or supplementary plots that account for bias/variance trade-off as a function of PC inclusion (e.g., harmonious plots).[50] In the interest of simplicity, we disregarded the use of more statistics and instead paid careful attention to residual plots.

Having accounted for the prime candidates responsible for factor space inflation, it is safe to assume that band variation identified by PCA is due primarily to composition changes as well as hydrogen-bonding behavior corresponding to morphologically random, short-range, or long-range coordination (see Figure 4.5,4.6 and 4.7 in particular). Concerning morphology, our primary observation is that PCA parses out slight, but consistent, spectral

variations that represent these phenomenon. This implies that only feature variations and not band resolution is an explicit requirement for chemical factor identification (sans overfitting issues). Naturally, this statement might appears as overtly generous at best and sophistic at worst. However, our position it is not wholly untenable when considering the interplay of nuanced morphology effects (introducing nonlinear behavior into the data) and their effects on factor space inflation.

To emphasize, when PCA is used to analyze our TPU blends, we've shown (in Figure 4.6) that the third PC was not isolated due to a clear changes in band resolution but distinguished via simultaneous, lineshape variations confined to the 1710 cm^{-1}, 1715cm^{-1}, and 1732 cm^{-1} features. These spectroscopic features are clearly not resolved based on any strict resolution criteria but we believe that a strong case was made favoring a physico-chemical interpretation via DSC. Furthermore, the changes in Figure 4.6 were extraordinarily blend-specific; showing only significant group differences for the 60:40 and possibly 80:20 blends and appearing to behave similiarly for the pure polymers. Thus, the third PC represents a (partial) morphological factor that does not scale linearly with hard segment fraction and therefore copolymer blend composition. Thus, additional factors identified in the cross-validation phase of PCR are either truly independent morphological effects, nuanced combinations of spectral features residing separately and consistently within the carbonyl stretching envelope, or artifacts of the statistical nature of cross-validation. Furthermore, future experiments employing multivariate regression to correlate Raman spectra to viscoelastic and mechnical behavior are likely to be fruitful.

Conclusions

Regardless of fundamental problems associated with direct interpretation of all significant PC axes, linear and composition-dependent trends exist in the carbonyl stretching region and are primarily confined to PC1, i.e., EST92 contains a higher fraction of hydrogen-bonded C=O groups *ab initio* whose influence diminishes (on average) as its concentration in the blend decreases. Note, these linear changes in blend composition, which are directly expressed as shape and intensity changes in the broad carbonyl stretching envelopes (centered at approximately 1715 cm^{-1} and 1732cm^{-1}) offer little information as to the specific manner in which these copolymers were interacting. As for other composition-dependent

trends, PCA appeared capable of isolating the strong hydrogen-bonding interactions for the 1718 cm^{-1} band onto PC2 as well as unique morphology effects for the 60:40 blend (and to a lesser extent the 80:20 blend) on PC3. Additional interactions were not predominately confined to a single PC and were responsible for introducing nonlinear behavior into the PCR model and residuals. Filtering of spectra prior to PCR eliminated portions of the group-dependent behavior present in the fitted residuals and reduced the factor space necessary to sufficiently model the data. Proper background subtraction, careful spectrometer alignment, and, most importantly, spectral preprocessing where the major factors in the success of PCA and PCR in the analysis of TPU mixture-spectra.

Overall, PCR modeling proved capable as a proof-of-concept for correlating hard segment molar fraction (and by association, weight fraction) with the micro-Raman spectra of TPU blends thereby demonstrating that linear multivariate models can be employed successfully to heterogeneous, strongly associating copolymer blends. Furthermore, by rationally selecting morphologically interesting band prior to analysis, multivariate methods show further promise for correlating the micro-Raman spectra of TPU blends to polymer mechanical and viscoelastic behavior as well as fostering easier interpretation of phase-separated domains in Raman hyperspectral imaging applications.[51]

Acknowledgements

The authors gratefully acknowledge the support of the National Science Foundation (award DMR-0619310), the University of Idaho Biological Applications of Nanotechnology (BANTech) research initiative, and helpful discussions with Professor Peter R. Griffiths, Chemistry Dept., UI.

References

1. V. P. Volkov, K. V. Nel'Son, É. N. Sotnikova, N. P. Apukhtina, L. I. Potepun. "IR-Spectroscopic Investigation of Molecular Interactions in Segmented Polyurethanes". J. Appl. Spectrosc. 1982. 35(5): 557-561.

2. R. G. Goddard, S.L. Cooper. " Polyurethane Cationomers with Pendant Trimethylammonium Groups. 1. Fourier Transform Infrared Temperature Studies". Macromolecules. 1995. 28(5): 1390-1400.

3. S. Velankar, S. L. Cooper. " Microphase Separation and Rheological Properties of Polyurethane Melts. 2. Effect of Block Incompatibility on the Microstructure". Macromolecules. 2000. 33(2): 382-394.

4. P. J. Yoon, C. D. Han. "Effect of Thermal History on the Rheological Behavior of Thermoplastic Polyurethanes". Macromolecules. 2000. 33(6): 2171-2183.

5. G. Odian. Principles of Polymerization. Hoboken, NJ: John Wiley & Sons, 2004. 4th ed. Pp. 130-132, 140-142.

6. K. Chen, T. L. Yu, Y. Chen, T. Lin, W. Liu. "Soft- and Hard-Segment Phase Segregation of Polyester-based Polyurethanes ". J. Polym. Res. 2001. 8(2): 99-109.

7. R. W. Seymour, G. M. Estes, S. L. Cooper. "Infrared Studies on Segmented Polyurethan Elastomers. I. Hydrogen Bonding". Macromolecules. 1970. 3(5): 579-583.

8. R. W. Seymour, S. L. Cooper. "Thermal Analysis of Polyurethane Block Copolymers". Macromolecules. 1973. 6(1): 48-53.

9. H. S. Lee, Y. K. Wang, W. J. MacKnight, S. L. Hsu."Spectroscopic Analysis of Phase-separation Kinetics in model Polyurethanes". Macromolecules. 1988. 21(1): 270-273.

10. Y. Li, W. Kang, J.O. Stoffer, B. Chu."Effect of Hard-segment Flexibility on Phase Separation of Segmented Polyurethanes". Macromolecules. 1994. 27(2): 612-614.

11. S. L. Chang, T. L. Yu, C. C. Huang, W. C. Chen, K. Linliu, T. L. Lin. "Effect of polyester side-chains on the phase segregation of polyurethanes using small-angle X-ray scattering". Polymer. 1998. 39(15): 3479-3489.

12. D. J. Martin, G. F. Meijs, P. A. Gunatillake, S. J. McCarthy, G. M. Renwick."The Effect of Average Soft Segment Length on Morphology and Properties of a Series of Polyurethane Elastomers. II. SAXS-DSC Annealing Study". J. Appl. Polym. Sci. 1997. 64(4): 803-817.

13. J. A. Miller, S. B. Lin, K. Hwang, K. S. Wu, P. E. Gibson, S. L. Cooper. "Properties of Polyether-Polyurethane Block Copolymers: effects of Hard Segment Length Distribution". Macromolecules. 1985. 18(1): 32-44.

14. H. J. Tao, X. Y. Meuse, W. J. MacKnight, S. L. Hsu. "A spectroscopic Analysis of Phase Separation behavior of Polyurethane in restricted geometry: Chain Rigidity effects". Macromolecules. 1994. 27(24): 7146-7151.

15. N. J. Clayden, C. Nijs, G. Eeckhaut. "Study of the Polymer Morphology in Urethane Eastomers by Solid State ^{2}H NMR and Small Angle X-ray Scattering". Macromolecules. 1998. 31(22): 7820-7828.

16. C. M. Brunette, S. L. Hsu, W. J. MacKnight. "Hydrogen-Bonding Properties of Hard-Segment Model Compounds in Polyurethane Block Copolymers". Macromolecules. 1982. 15(1): 71-77.

17. M. M. Coleman, K. H. Lee, D. J. Skrovanek, P. C. Painter. "Hydrogen Bonding in Polymers. 4. Infrared Temperature Studies of a Simple Polyurethane ". Macromolecules. 1986. 19(8): 2149-2157.

18. L. I. Maklakov, V. L. Furer, V. V. Alekseev, A. L. Furer."Investigation of the Vibrational Spectra of 2,6- and 4,6-Polyurethanes and Hexamethylenedimethylurethane". J. Appl. Spectrosc. 1979. 31(4): 1285-1289.

19. L. Teo, C. Chen, J. Kuo."Fourier Transform Infrared Spectroscopy Study on effects of Temperature on Hydrogen Bonding in Amine-Containing Polyurethanes and Poly(urethane-urea)s ". Macromolecules. 1997. 30(6): 1793-1799.

20. C. Wilhelm, J. Gardette. Polymer. "Infrared Analysis of the Photochemical Behaviour of Segmented Polyurethanes: 1. Aliphatic Poly(ester-urethane) ". 1997. 38(16): 4019-4031.

21. V. W. Srichatrapimuk, S. L. Cooper."Infrared Thermal Analysis of Polyurethane Block Polymers ". J. Macromol. Sci., Part B: Phys. 1978. 15(2): 267-311.

22. D. W. Mayo, F. A. Miller, R. W. Hannah. Course Notes on Interpretation on Infrared and Rama Spectra. Hoboken, NJ: John Wiley & Sons. 2003. Pp. 22-23.

23. D. J. Skrovanek, S. E. Howe, P. C. Painter, M. M. Coleman. "Hydrogen Bonding in Polymers: Infrared Temperature Studies of an Amorphous Polyamide". Macromolecules. 1985. 18(9): 1676-1683.

24. S. K. Pollack, D. Y. Shen; S. L. Hsu, Q. Wang, H. D. Stidham. "Infrared and X-ray Diffraction Studies of a Semirigid Polyurethane". Macromolecules. 1989. 22(2): 551-557.

25. S. Parnell; K. Min, M. Cakmak. "Kinetic Studies of Polyurethane Polymerization with Raman Spectroscopy". Polymer. 2003. 44(18): 5137-5144.

26. E. R. Malinowski. Factor Analysis in Chemistry. New York, NY: John Wiley & Sons. 2002. 3rd ed.

27. H. Abdi, L. J. Williams. "Principal Component Analysis". Wiley Interdiscip. Rev. : Comp, Stat. 2010. 2(4): 433-459.

28. I. Hamerton, H. Herman, A. K. Mudhar, A. Chaplin, S. J. Shaw. "Multivariate Analysis of Spectra of Cyanate Ester/bismaleimide Blends and Correlations with Properties". Polymer. 2002. 43(11): 3381-3386.

29. T. Naes, T. Isaksson, T. Fearn, T. Davies. A User-Friendly Guide to Multivariate Calibration and Classification. Chichester, UK :NIR Publications. 2002. Pp. 5-33, 177-190.

30. K. V. Mardia, J. T. Kent, J. M. Bibby. Multivariate Analysis. New York, NY: Academic Press Inc. 1979. Pp. 213-253.

31. C. Chatfield, A. J. Collins. Introduction to Multivariate Analysis. London, UK: Chapman & Hill. 1980. Pp. 68-71.

32. J. M. Shaver. "Chemometrics for Raman Spectroscopy". In: I. R. Lewis, H.G.M. Edward, editors. Handbook of Raman Spectroscopy. New York, NY: Marcel Dekker. 2001. Vol. 28, Chap. 7, Pp 275-306.

33. S. Šašić, T. Itoh, Y. Ozaki. "Classification of Single-Molecule Surface-enhanced Resonance Raman Spectra of Rhodamine 6G from Isolated Ag Colloidal Particles by Principal Component Analysis". Vib. Spectrosc. 2006. 40(2): 184-191.

34. E. R. Malinowski."Factor Analysis for isolation of the Raman Spectra of Aqueous Sulfuric Acid Components". Anal. Chem. 1984. 56(4): 778-781.

35. G. M. Arnold, A. J. Collins. J. R."Interpretation of Transformed Axes in Multivariate Analysis". Stat. Soc. Ser. C Appl. Stat. 1993. 42(2): 381-400.

36. S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjostrom. "Multivariate Data Analysis in Chemistry". In: B. R. Kowalski, editor. Chemometrics: Mathematics and Statistics in Chemistry. Series C: Mathematical and Physical Sciences. Dordrecht, Holland: D. Reidel Publishing Company. 1984. Vol. 138, Chap. 2, Pp 17-95.

37. S. Wold."Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models". Technometrics. 1978. 20(4): 397-405.

38. H. T. Eastment, W. J. Krzanowski. Technometrics."Cross-Validatory Choice of the Number of Components From a Principal Component Analysis". 1982. 24(1): 73-77.

39. R. G. Bereton. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. Chichester, UK: John Wiley & Sons Ltd. 2003. Pp. 313-333

40. A. J. Brandolini, D. D. Hills. NMR spectra of polymers and polymer additives. New York, NY: Marcel Dekker. 2000. Pp. 470-471.

41. K. Hatada, T. Kitayama. NMR Spectroscopy of Polymers. New York, NY: Springer. 2004. Pp. 43-71.

42. A. Saiani, W. A. Daunch, H. Verbeke, J. W. Leenslag, J. S. Higgins."Origin of Multiple Melting Endotherms in a High Hard Block Content Polyurethane. 1. Thermodynamic Investigation". Macromolecules. 2001. 34(26): 9059-9068.

43. K. Bagdi, K. Molnár, B. Pukánszky, B. J. Pukánszky. "Thermal Analysis of the Structure of Segmented Polyurethane Elastomers". Therm. Anal. Calorim. 2009. 98(3): 825-832.

44. T. K. Chen, S. T. Shieh, J. Y. Chui. "Studies on the First DSC Endotherm of Polyurethane Hard Segment based on 4,4'-Diphenylmethane Diisocyanate and 1,4-Butandiol". Macromolecules. 1998. 31(4): 1312-1320.

45. J. T. Edward, S. C. Wong. "Ionization of Carbonyl compounds in Sulfuric Acid. Correction for medium effects by Characteristic Vector Analysis". J. Am. Chem. Soc. 1977. 99(13): 4229-4232.

46. T. Bocklitz, A. Walter, K. Hartmann, P. Rosch, J. Popp. "How to pre-Process Raman Spectra for Reliable and Stable Models?". Anal. Chim. Acta 2011. 704(1): 47-56.

47. P.J. Gemperline. "Principal Component Analysis". In: P.J. Gemperline, editor. Practical Guide to Chemometrics. Boca Raton, FL: CRC Press. 2006. 2nd ed. Chap. 4, Pp. 86-89.

48. Q. Xu, Y. Liang. "Monte Carlo Cross validation". Chemom. Intell. Lab. Syst. 2001. 56(1): 1-11.

49. K. Baumann. "Cross-validation as the objective function for variable-selection techniques". Trends Anal. Chem. 2003. 22(6): 395-406.

50. J. H. Kalivas, P.J. Gemperline. "Calibration". In: P.J. Gemperline, editor. Practical Guide to Chemometrics. Boca Raton, FL: CRC Press. 2006. 2nd ed. Chap. 5, Pp. 144-147.

51. H. Janik, B. Palys, Z.S. Petrovic. "Multiphase-Separated Polyurethanes Studied by micro-Raman Spectroscopy". Macromol. Rapid. Commun. 2003. 24(3): 265-268.

Chapter 5. Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least-squares regression

Reproduced with permission (Appendix A): Andrew T. Weakley, Arthur Miller, Peter R. Griffiths, Sean J. Bayman, Anal Bioanal Chem, 2014, DOI: 10.1007/s00216-014-7856-y (available online)

Abstract

The feasibility of measuring airborne crystalline silica (α-quartz) in noncoal mine dusts using a direct-on-filter method of analysis is demonstrated. Respirable α-quartz was quantified by applying a partial least squares (PLS) regression to the infrared transmission spectra of mine-dust samples deposited on porous polymeric filters. This direct-on-filter method deviates from the current regulatory determination of respirable α-quartz by refraining from ashing the sampling filter and redepositing the analyte prior to quantification using either infrared spectrometry for coal mines or x-ray diffraction (XRD) from noncoal mines. Since XRD is not field portable, this study evaluated the efficacy of Fourier transform infrared spectrometry for silica determination in noncoal mine dusts. PLS regressions were performed using select regions of the spectra from non-ashed samples with important wavenumbers selected using a novel modification to the Monte Carlo unimportant variable elimination procedure. Wavenumber selection helped to improve PLS prediction, reduce the number of required PLS factors, and identify additional silica bands distinct from those currently used in regulatory enforcement. PLS regression appeared robust against the influence of residual filter and extraneous mineral absorptions while outperforming ordinary least squares calibration. These results support the quantification of respirable silica in noncoal mines using field-portable infrared spectrometers.

Key Words: Partial least squares, Monte Carlo unimportant variable elimination, silica measurement, FT-IR, mine dust

Introduction

The US Mine Safety and Health Administration (MSHA) mandates the monitoring and quantification of occupational exposure to airborne respirable silica in US mines [1]. The

most abundant polymorph of crystalline silica, α-quartz, is internationally recognized as a carcinogen [2-4]. Failure to quantify and mitigate worker exposure to respirable quartz leads to a decrease in lung function [5] and may result in respiratory diseases such as silicosis [6-9].

Since no field-portable method exists to monitor silica exposure, a current research goal of the National Institute for Occupational Safety and Health (NIOSH) Office of Mine Safety and Health Research (OMSHR) involves evaluating the efficacy of a direct-on-filter method for potentially measuring silica in the field. The two analytical techniques presently used to quantify silica in respirable samples collected in mines are mid-infrared (IR) spectrometry and X-ray diffraction (XRD), which are used for coal mines and noncoal mines, respectively [10-12]. The ubiquity, speed, and diminishing cost of instrumentation makes IR spectroscopy the preferred analytical technique for the purpose of field-portable measurements. However, the success of on-site assessment using portable FT-IR spectrometers is contingent upon overcoming critical implementation barriers associated with end-of-shift silica assessment [13, 14].

Field-portable IR methods will require procedures that anticipate filter-substrate mishandling and limit the need for specialized training in quantitative spectroscopic analysis [12, 15]. Regulatory agencies currently utilize multiple preparation steps prior to acquiring an IR spectrum of airborne dust. These include pre- and post-weighing filters, removal of the filter along with any organic contaminants by plasma ashing, and finally redeposition of the treated dust on a clean polymeric filter prior to measurement. Sample preparation and spectral post-processing both afford distinct stages to introduce variability into the prediction [16, 17]. Technical aptitude in these domains still dictates the accuracy of quartz determination by experts.

Acquiring spectra from samples collected directly on the filter substrate would facilitate the rapid determination of α-quartz by circumventing the ashing and redeposition steps. This approach has been tested in the recent past [15, 18-19]. A consequential improvement in prediction accuracy is expected as opportunities for sample loss or mishandling are removed from the protocol [13, 17]. Unfortunately, this approach leaves the filter in place along with any trace organic and mineral interferents within the IR sampling volume. Kaolinite clay is the most serious and common interference when using the MSHA-

regulatory IR method for quartz prediction [1, 20], since its presence inflates the estimated mass of quartz [21, 22]. Although kaolinite-correction methods are commonly used, they can add to the inaccuracy of the determination of α-quartz, since there are several types of kaolin with slightly different IR spectra [23]. Thus the application of chemometric techniques, such as partial least squares regression [24-26], may produce calibrations resistant to confounder and substrate interference when these effects are appropriately modeled or at least suppressed by carefully selecting only stable predictors (channel-wavenumbers) quintessential to determining a target analyte.

Partial least squares (PLS) regression demonstrates remarkable selectivity and prediction accuracy when applied to complex multicomponent spectra [25-27]. A major appeal of PLS regression in spectroscopy relates to its handling of predictor collinearity and explicit modeling of the data structure (i.e., latent physicochemical effects). In other words, PLS regression assumes that a few latent variables, articulated mathematically as linear combinations of predictors (absorbances at each wavenumber), underscore the relationship between absorbance changes and analyte concentration. By isolating a common subspace between the concentration y-vector and the matrix of IR absorbances, PLS regression inherently minimizes the influence of irrelevant species on calibration while maximizing the linear relationship(s) between analyte concentration and target absorbance.

Using a PLS approach, improvement in the quantification of airborne silica is expected [28, 29]. Given a host of known background interferences and probable unknown day-to-day variation in geological interferences, an integrated approach to PLS calibration will better suit the direct-on-filter quantification of airborne silica. Furthermore, this feasibility study assesses the role of sample subset partitioning on PLS calibration, the gains in precision, and aid to latent variable interpretation imparted via a novel approach to Monte Carlo unimportant variable elimination (MCUVE) [30], and the ultimate basis for selecting a viable PLS model. Candidate PLS models generated from variable elimination are compared to an ordinary least squares (OLS) regression derived from the MSHA P-7 method.

Experimental

Quartz sampling

Filter samples of noncoal mine dusts were obtained from field surveys in three active mines in Idaho, New York, and Ohio. Samples were acquired from mines with sedimentary

and igneous rock formations. Seventeen samples were acquired from a hard rock silver mine on two separate occasions, twenty-one samples from a granite mine, and eight samples from a limestone mine. A total of 46 samples were available for analysis.

Airborne dust was collected using sampling trains standardized for noncoal mines [11]. The sampling pump flow rate was set at 1.7 L min^{-1} and large particulate matter was removed using Dorr-Oliver style cyclones, which separate the respirable fraction. Respirable particles were deposited onto preweighed 37-mm polyvinyl chloride (PVC) filters with 5-µm pore size (SKC Corp., Inc.).[1] Filters were mounted within three-piece plastic cassettes. After collection, all filter samples were postweighed to determine the mass of loaded dust. For each group of filter samples collected during one mine visit, three unused filters from the same lot were set aside as controls.

FT-IR instrumentation and acquisition parameters

Spectra were collected in transmission mode at 4 cm^{-1} resolution by averaging 40 scans using the Bruker Optics model Alpha FT-IR spectrometer. Interferograms were processed using Blackman-Harris 3-term apodization prior to Fourier transformation. Spectra were saved from 399.5 to 3,998.5 cm^{-1} so that each spectrum contained 2,542 spectral channels, i.e., discrete wavenumbers, since the data spacing for the Bruker Alpha spectrometer operating at a nominal resolution of 4 cm^{-1} is 1.417 cm^{-1}. Each individual spectrum was ratioed against the spectrum of a blank filter to remove the filter background. Multiplicative scatter correction (MSC) was applied to remove scattering effects from each individual spectrum as well [31]. Fig 5.1 shows the resulting average spectra from 2,200 to 400 cm^{-1} for samples from the silver mine (see Supplementary Material Figs E.13 and E.14 for the granite and limestone mine examples, respectively).

[1] Since particles larger than 4 µm were removed by the Dorr-Oliver cyclone and the pore size is specified as 5 µm, it may be asked why all the particles did not pass through the pores. In practice, the PVC from which the filters were fabricated is porous and the pathways through it are sufficiently tortuous that small particles contact and adhere to the structural features via diffusion, interception, and impaction.

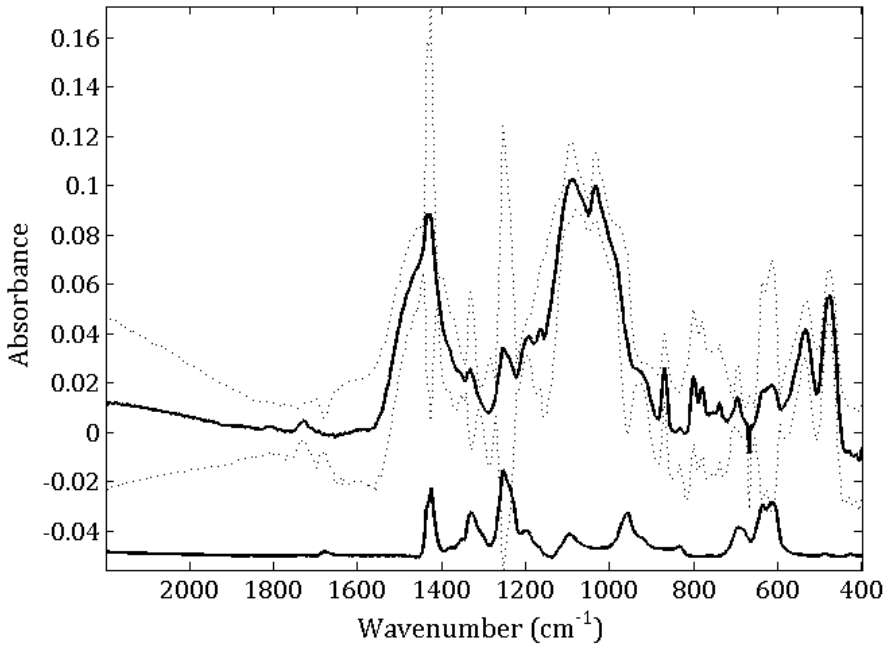


Fig. 5.1 Average spectrum calculated from 17 silver mine dust samples (top, solid) with accompanying +/- one standard deviation on each absorbance (*dashed*). A baseline-corrected [62] and scaled (to 5% original absorbance) spectrum of PVC is plotted with -0.05 absorbance offset. The peaks of the α-quartz doublet are clearly visible at 780 and 800 cm^{-1}.

To compensate for the nonuniform spatial distribution of dust particles, each filter was carefully mounted using a stainless steel holder to ensure that the 6-mm diameter IR beam always interrogated the center of the filter. Miller et al. showed that the estimation of silica using only this center portion of the filter adequately captured data representative of the silica deposited onto the PVC filter, when samples were collected using any one of three common sampler fixtures (filter cassettes) and a Dorr-Oliver-style respirable cyclone [14]. After analysis, all samples were sent to an independent laboratory for XRD analysis, which is the primary analytical method used for regulatory measurement of silica in noncoal mines.

X-ray diffraction

Quantitative XRD analysis (NIOSH 7500 method) was performed independently (RJ Lee Group Inc, Monroeville, PA) to generate the empirical α-quartz standards for PLS and manual FT-IR calibration [11]. The filter and any organic matter in the dust was eliminated using a low-temperature radiofrequency plasma asher. The appropriate procedures were then followed to wash, suspend, and disperse intact particulate matter in solution. Thin film

deposition of the suspended solid onto a qualified 25-mm silver membrane filter substrate followed.

This was then placed in an XRD sample holder. Working standards of α-quartz (NIST SRM 2950) were prepared in triplicate and used to generate a calibration curve from the referenced-normalized, α-quartz using the primary diffraction peak at 26.66° (2θ). Pertinent diffraction peaks collected from respirable α-quartz were also normalized and the concentration of silica (µg SiO_2/filter) estimated using this calibration curve. The PANalytical Cubix Pro X-ray diffractometer equipped with rotating anode was used to gather all necessary diffractograms.

Spectral preprocessing for PLS regression

Although the single-beam spectra were ratioed against that of a blank filter, the spectra often included residual PVC bands because of differences in filter thicknesses. Low-frequency baseline perturbations due to substrate and particle scattering were also observed. Given the minimal baseline complexity, a formal baseline correction algorithm was not used. Rather, a smoothed first-derivative spectrum was produced using a Savitzky-Golay 21-point, second-order polynomial filter. This operation simultaneously removed the predominant baseline and (because of the smoothing effected by the filter) increased the signal-to-noise ratio (SNR) of each spectrum with a minimal sacrifice in resolution. Filtering removed the first 10 and the last 10 channels from each IR spectrum leaving 2,522 points in the first-derivative spectra.

Sample set partitioning based on joint *x-y* distances (SPXY) [32] was used to select calibration samples that optimally spanned the calibration range (0-281 µg-SiO_2/filter). From the original 46 samples, 2 were first removed as outliers, using principal component analysis (PCA) [27, 33], and SPXY was applied to select the best 29 (N_c) samples for training, and the remaining 15 spectra (N_p) were assigned for validation. Because 42 out of the 44 samples contained less than 200 µg-SiO_2/filter, SPXY selected all SiO_2 compositions in the higher concentration range. To ensure comparable coverage in the validation samples, one sample from the granite mine data (251 µg-SiO_2/filter) was exchanged with a low concentration SiO_2 sample from the calibration set.

Model training proceeded using two wavenumber ranges for PLS regression as follows: the wavenumber range containing the α-quartz doublet (751-857 cm^{-1}) and the

spectral range between 415 and 2,201 cm^{-1}(referred to subsequently as the half spectrum). It can be assumed *a priori* that most channels in a full first-derivative mid-IR spectrum (415-3,986 cm^{-1}) will not be useful to PLS regression. Furthermore, all of the fundamental lattice vibrations of silica are active below ~1,400 cm^{-1}. Confining the chemometric analysis to absorbance within the 415-2,201 cm^{-1} range accomplished three things. First, it left enough redundant/useless absorbance available to challenge our variable selection algorithm (discussed in detail below). Secondly, it accelerated the required runtime for the variable selection algorithm. And finally, this spectral range facilitated an exploration of additional silica vibrations outside the fairly narrow α-quartz doublet.

Model training and testing spectra, using the masses acquired by XRD as the primary calibration values (Y), were mean-centered prior to regression to stabilize the PLS1 algorithm. Preprocessing, PLS modeling, and feature selection were performed in MATLAB® (2007b, The Mathworks, Natick, MA) using either the open source libPLS (v. 1.6, Changsha Nice City, China) package or custom software.

Latent variable and feature selection in PLS regression

A major practical task of PLS regression is to estimate the number of latent variables best representing the relationship between the spectra and analyte composition. In this study, Monte Carlo cross-validation (MCCV) employed a 60-40 calibration-validation split with a calibration set resampling number set to 1000 [35]. A minimized root mean squared error of cross-validation (RMSECV) identified the optimal number of latent variables to use in all PLS models tested.

There is a growing interest in modeling the latent structure while including only the best (and fewest) spectral features [30, 36, 37]. Wavenumber selection routines often apply an objective metric and/or heuristic device to eliminate those variables least important to the modeling problem. Data reduction often removes noise and excessive redundancy. This generally improves performance and, more importantly, assists latent variable interpretation.

In this study, a modified version of Cai, Li, and Shao's Monte Carlo unimportant variable elimination is implemented in a novel backward elimination manner [30]. In this BMCUVE routine, the latent variables are estimated at each pass of the algorithm (using MCCV) [34, 37], followed by the selection of the best subset of wavenumbers needed to predict α-quartz. The process of latent variable estimation and wavenumber elimination

proceeds until a termination criterion is reached (or the number of variables available for regression is exhausted). Similar to other feature-selection methods, only calibration data are used in model reduction. Prediction testing and cross-validation inform the choice of the final PLS model.

A single pass of the MCUVE algorithm used in this study proceeds as follows: 60% of the calibration data is randomly sampled to develop a temporary training set, a PLS regression is performed using this training data, and a regression coefficient for each predictor variable is calculated. This entire process is repeated for 1,000 trials resulting in 1,000 regression coefficients for *each* predictor variable. Note that the number of latent variables remained fixed over the 1000 trials where only the composition of the training samples was varied by random sampling. Additionally, Cai and colleagues recommend only 100 trials for an individual pass of MCUVE [30]. This was deemed inadequate for this particular data due to the small training set.

After resampling, each variable's relative importance is assessed using a statistic known as a reliability index (RI) [38]. Formally,

$$RI_j = \frac{mean(\beta_{ji})}{std(\beta_{ji})} \qquad i = 1, 2, \dots, N_r \quad j = 1, 2, \dots, p \qquad \text{Equation 5.1}$$

where $mean(\beta_{ji})$ is the average regression coefficient for the 1,000 trials (=N_R) on the *j*th wavenumber, $std(\beta_{ji})$ is the *j*th wavenumber's standard deviation, and *p* are the total number of wavenumbers available to BMCUVE. Due to some sampling-distribution asymmetry, a robust form of the RI criterion was used where the $median(\beta_{ji})$ and interquartile range, $IQR(\beta_{ji})$, was substituted for each regression coefficient's mean and standard deviation, respectively.

Typically, variables with an *RI* value above some estimated noise level are retained (i.e., deemed important) while those below this cutoff value are discarded [38]. Given a reasonable estimate of the cut-off value, wavenumber elimination ceases after discarding unimportant variables once. This approach is adequate at removing redundant wavenumbers from spectra containing only additive noise and minimal artifacts. For this study, we assumed that mid-IR transmission spectra were nonideal on both accounts, i.e., spectra were rife with absorbance redundancies, contaminated with scattering artifacts and interferences, and contained spectroscopic biases contingent upon varying geological factors.

As opposed to operating a single pass of MCUVE, the quantity of variables available for modeling was successively reduced by 10% until only a small number of wavenumbers remained for PLS regression. For example, if approximately half the spectral channels (e.g., 415-2,201 cm^{-1} range = 1,262 channels) were evaluated by BMCUVE, the first pass of the algorithm would remove the 126 least important variables according to their low *RI* values. The remaining 1,136 channels are then evaluated in the second pass of the algorithm resulting in the subsequent removal of another 114 least important variables, and so on. Starting with a given number of predictor (p_i) and fixed percentage of wavenumber removed at each pass (q), it can be easily shown that the number of passes of the MCUVE algorithm required to reach one remaining wavenumber is expressed as follows:

$$M = {-log_{10}(p_i)}\Big/{log_{10}\left(1 - \frac{q}{100}\right)}. \qquad \text{Equation 5.2}$$

Therefore, equation (2) estimates that the half-spectrum elimination will take 68 passes to reach a single remaining wavenumber (after rounding M up to the nearest integer). Note, rounding creates a situation in which a handful of these passes are redundant. To counter a repetitive evaluation, auxiliary code is executed to reduce the number of wavelengths by one for a repetitive pass. This influences the accuracy of equation (2).

This entire backward elimination routine was repeated five times in order to identify a class of viable models that showed good performance according to the root mean square error of calibration (RMSEC), RMSECV, root mean square error of prediction (RMSEP), and the cross-validated coefficient of determination (Q^2). The results of PLS regression with and without wavenumber selection were compared for the two aforementioned wavenumber ranges: 751-857 cm^{-1} and 415-2,201 cm^{-1}. To investigate whether the mass of silica could be predicted without the α-quartz doublet, the 415-2,201 cm^{-1} range was evaluated using BMCUVE with the 751-857 cm^{-1} range removed prior to analysis.

Three parameters were supplied to the MCUVE function (libPLS v.1.6, Changsha Nice City, China) prior to execution: the size of a temporary training set, the resampling number, and the number of latent variables. As described above, 60% of the available calibration samples were randomly sampled, the resampling number was set to 1,000, and the latent variables estimated using MCCV.

Manual Calibration

The manual calibration was similar to that used in previous work [13], which entailed using OPUS spectral analysis software (Bruker Optics) to perform peak integrations on IR spectra from the dust samples. We note here that band integration is susceptible to errors, especially for spectra with curved baselines. In a manner similar to that used in an earlier study [13], the approach included silica quantification via manual integration of the α-quartz doublet in the range 767-816 cm^{-1}. A correction protocol for kaolinite interference involved subtracting the potential contribution of kaolinite to the doublet using a peak-ratio technique previously published along with the kaolinite band at 915-cm^{-1} [39]. The area of the corrected doublet bands was used as a single predictor in an ordinary least squares (OLS) regression and regressed against the mass of XRD estimated silica. Although correction for kaolinite absorption was employed, the level of kaolinite in these samples was generally negligible as indicated by the absence of a strong peak at 915-cm^{-1}.

To foster simple comparisons, the samples were partitioned according to the prescription of the SPXY algorithm for the PLS regression in the region of the α-quartz doublet. A calibration curve was developed using 2/3 of the available samples for model training and the other 1/3 for prediction testing. OLS calibration, prediction, and model diagnostics were also performed using MATLAB packages.

Results

After 42 wavenumbers were removed using BMCUVE, PLS regression results were obtained using 34 wavenumbers from the quartz doublet region (Fig 5.2). Performance statistics for select regression models are presented in Table 1. It may be noted that the calibration error was often higher than the prediction error for the 11 models shown in Table 1. This indicates that the training data often best spanned the calibration range while the validation set understated the model performance. This was probably caused by the manner in which the SPXY algorithm was used in this study and will be discussed in detail below.

Returning to the 34 predictor PLS model, visual inspection of figure 5.2 confirms the strong Q^2 statistic—that is, a strong linear relationship between the predictors and response variable. Remarkably, only one latent variable (LV) was required to predict quartz in the presence of the PVC substrate when the narrow wavenumber range was tested.

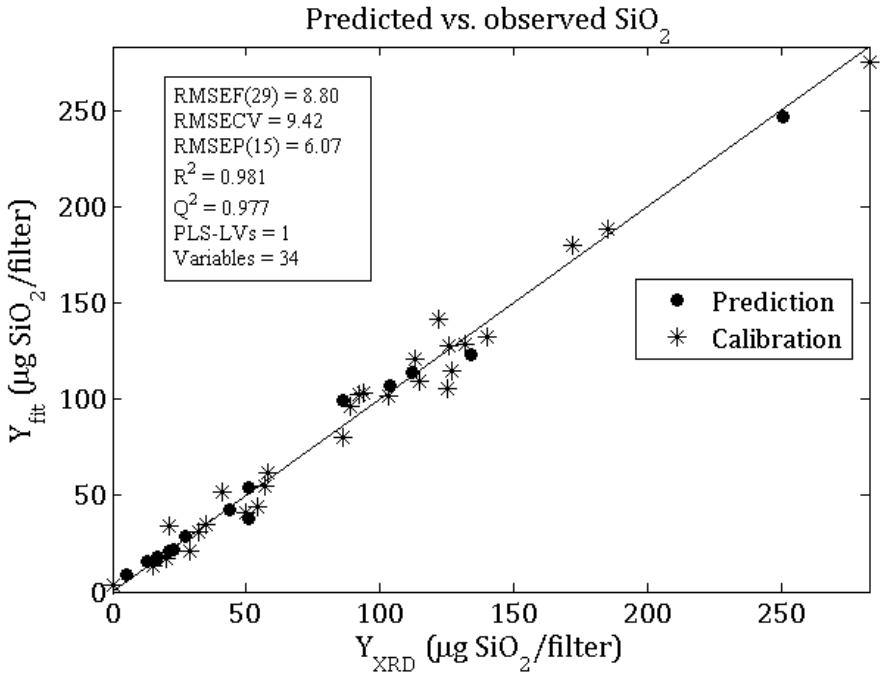


Fig. 5.2 Predicted versus observed α-quartz composition for the 34-variable PLS model

Table 5.1 Summary of select PLS regressions (lines 1-10) and manual OLS regression (line 11). The twenty wavenumbers chosen for model #7 were not the same as model #10. In the former case, all 20 variables resided within the α-quartz doublet. The asterisk (*) on the 415-2,201 range (Model's 8-10) indicate that 58 wavenumbers from the α-quartz doublet region were excluded from the regression and BMCUVE procedures

Model	Range (cm^{-1})	Variables (#)	LVs (#)	R^2	Q^2	RMSEC (µg SiO_2)	RMSECV (µg SiO_2)	RMSEP (µg SiO_2)
1	751-857	76	2	0.980	0.972	9.34	10.44	6.41
2		34	1	0.981	0.977	8.80	9.42	6.07
3		8	1	0.980	0.971	9.24	10.62	6.03
4		2	1	0.975	0.969	10.22	11.05	6.18
5	415-2201	1262	3	0.972	0.925	11.38	17.35	9.97
6		139	2	0.981	0.972	9.29	10.64	9.69
7		20	1	0.981	0.977	9.09	9.66	6.19
8	415-2201*	1204	3	0.972	0.918	11.78	18.04	10.44
9		223	2	0.981	0.972	9.24	10.71	9.28
10		20	2	0.985	0.980	8.15	8.95	9.42
11	767-816	1	-	0.935	0.922	15.62	17.02	11.27

Table 1 shows that BMCUVE successfully suppresses redundant and/or artifact-related absorbance by removing less important wavenumber; ultimately, this improves the performance of PLS regression relative to the cases where a wider range is used. Indeed, when all 76 wavenumber were employed for the doublet case, 2 latent variables were selected to achieve the minimized RMSECV. The removal of 42 wavenumber led to the elimination of an additional latent variable showing an improved Q^2, RMSECV, and RMSEP. Only one true "effect," which is captured by a corresponding PLS component in the 2, 8, and 34 wavenumber models, accurately explains the relationship between the IR spectra (X) and the mass of silica (Y). The near coincidence of the X-loadings (p_1) and PLS loadings weights (w_1) (when plotted against wavenumber, not shown) [26] simplifies the interpretation of the latent "effect," i.e., the latent variable is simply mapping the spectral variation due to changes in the mass loading of silica on each filter sample.

A spectrum from the granite mine data illustrates the wavenumber elimination path for regressions involving the α-quartz doublet (Fig. 5.3). The absorbance and first-derivative spectrum are plotted with the selected wavenumbers clearly identified. The selection of wavenumbers from both components of the doublet demonstrates that the most successful PLS regressions required attributes from both the α-quartz lattice vibration at 780 cm^{-1} and the transverse optical (TO) Si-O-Si symmetric stretch at 800 cm^{-1} [22, 40-42].

Two latent variables were required in the 76-wavenumber regression. These were needed to accommodate contributions from scattering artifacts and residual PVC bands. Wavenumber elimination in first-derivative spectra began by eliminating wavenumbers with an absorbance of approximately zero followed by the weak PVC band at ~834 cm^{-1}. This order of elimination endorses the use of a reliability measure for wavenumber elimination. Overall, the most reliable variables had a large mean response and small resampling variance (IQR) which was equated to a large reliability index and thus a higher likelihood of retention. On the contrary, variations in PVC bands were random (large IQR) and uncorrelated with the silica concentration vector (small mean response), *i.e.*, regression coefficients on PVC wavenumbers showed low reliability in quartz prediction leading to early rejection. Wavenumbers near a band's steepest points show the largest mean response, as would be

expected for a first-derivative spectrum, as shown in figure 5.3; hence, this favored their retention in the PLS models as dictated by BMCUVE.

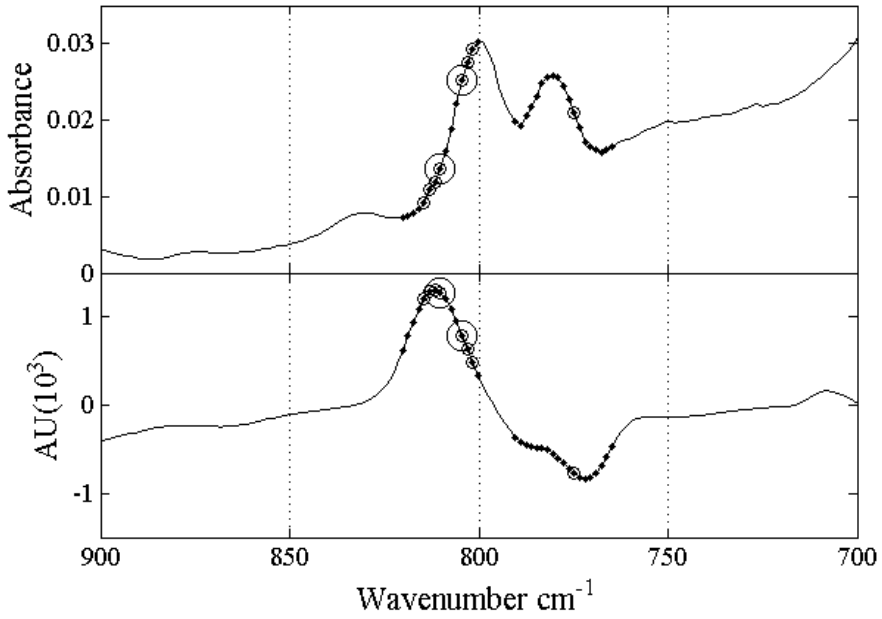


Fig. 5.3 Absorbance (*top*) and first-derivative (*bottom*) spectrum of the region containing the α-quartz doublet for the 20th granite mine sample. The original regression employed 76 wavenumbers from 751 to 857 cm^{-1}. BMCUVE selected the 34 (*dots*), 8 (*small circles*), and 2 (*large circles*)

Three models were selected when the 415-2,201cm^{-1} spectral range (1,262 channels) was screened using BMCUVE. The 1,262 wavenumber regression required 3 latent variables with an unacceptably high cross-validation and prediction error. Although model performance improved after only a single pass of BMCUVE, removing a total of 1,123 wavenumbers achieved substantially better regression performance. This was complemented by a reduction in the optimal number of PLS components. For fewer than 30 wavenumbers, only the spectral regions that included the α-quartz doublet were selected. This further highlights the value of the α-quartz region to silica prediction. The 139-wavenumber PLS regression was recorded and plotted to illustrate that other wavenumbers outside the doublet range were useful for quantitative analysis (Fig 5.4). Although somewhat obscured by residual PVC bands, features that roughly represent known silica vibrations are evident.

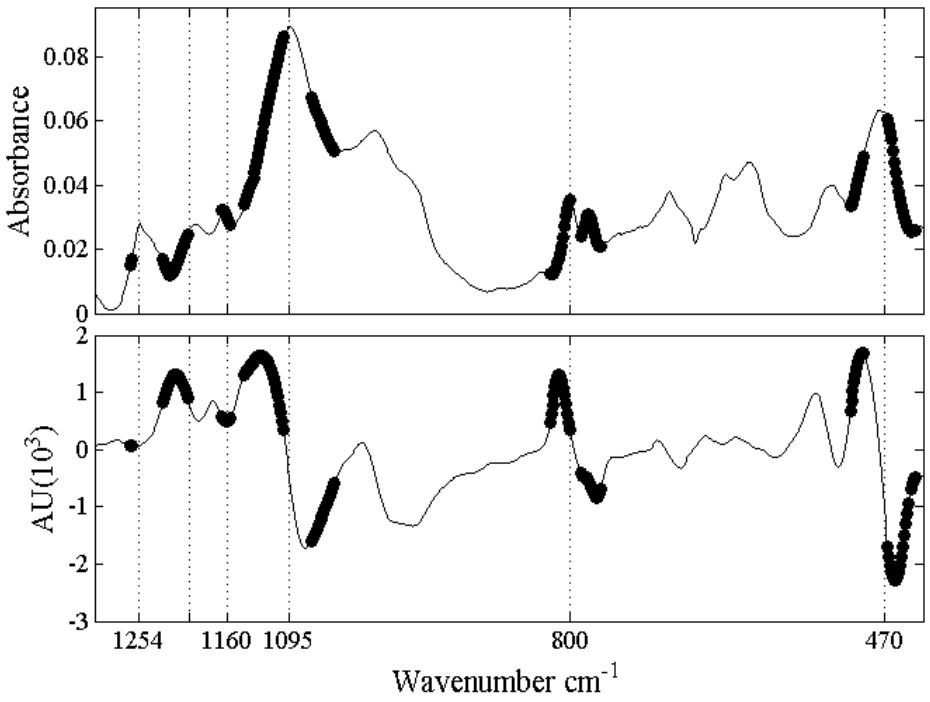


Fig. 5.4 Wavenumbers (*black dots*) retained upon removing 1,123 redundant predictors using BMCUVE for the 415 to 2,201 cm^{-1} range. *Vertical lines* mark the observed TO_1 mode (470 cm^{-1}), quartz doublet (780 and 800 cm^{-1}), TO_3 mode (~1,080 cm^{-1}) and tentative LO_4 (~1,160 cm^{-1}), TO_4 (1,200 cm^{-1}; unlabeled) and LO_3 mode (1,254 cm^{-1}).

Adopting some shorthand from Innocenzi and group symmetry nomenclature from Scott and Porto, bands in the spectrum shown in Fig 5.4 are assigned as the Si-O-Si rocking vibration (TO_1, *E*) at 470 cm^{-1}, the antisymmetric A_2 mode at 780cm^{-1}, the Si-O-Si symmetric stretch (TO_2 , *E*) at 800 cm^{-1}, the Si-O-Si antisymmetric stretch at ~1,080 cm^{-1} (TO_3, *E),* and possibly the LO_4 (*E*) mode at ~1,160 cm^{-1} [22, 42-45]. Four wavenumbers selected between the 1,200 cm^{-1} and 1,254 cm^{-1} band are possibly modeling effects from the TO_4 (*E*) mode and/or the LO_3 (*E*) modes, respectively. Extensive PVC interference together with the weak IR activity of the TO_4/LO_4 doublet in α-quartz greatly complicates determinations in this region [22, 41, 45].

Figure 5.4 illustrates that wavenumbers near the largest amplitude features in the derivative spectra were retained, with the exception of those near 1,160 cm^{-1}. All variables beyond 1,263 cm^{-1} were removed by BMCUVE thus nullifying the need to show the spectrum out to 2,201 cm-1. Figure 5.4 illustrates that α-quartz transmission spectra exhibit a broad envelope spanning the region from 1,000 to 1,300 cm^{-1}. Assignment of specific bands

is complicated by the polarization and orientation dependence of bands of α-quartz in this region [44, 46, 47]. In fact, large absorbance deviations (probably caused by scattering) near the tails of the 1,000-1,300 cm^{-1} envelope might support the hypothesis that this band is shifting and broadening as the sampling volume changes with α-quartz concentration. Therefore, for example, the assignment of the TO_3 mode is tentative for transmission IR spectra as this band is often unresolved from an A_2 mode unique to α-quartz [22]. Furthermore, our LO_3 assignment should not be confused with those observed in sol-gel and vitreous-SiO_2 thin films [41, 42, 48]. In those studies, the LO_3 mode often appears as an unmistakable broad shoulder between 1,200 and 1,260 cm^{-1} which shows an absorbance increase and band shift in proportion to incidence angle (Berreman effect) and thermal treatment [49, 50].

Multivariate analysis was repeated using half the spectral range while intentionally excluding the α-quartz doublet to investigate whether absorbance residing entirely outside the doublet was capable of predicting silica (Table 1, rows 8-10). The prediction error minimum was achieved using 223 wavenumbers, some of which included features from known quartz modes at 516.9 and 695.5 cm^{-1}. Five variables corresponding to baseline (>1,430 cm^{-1}) were also included in this model, suggesting a slightly suboptimal result (see Appendix E, Fig E.15). In fact, a 20-wavenumber model was tabulated because it showed the lowest RMSECV error and utilized absorption features exclusively from the TO_1 (469 cm^{-1}) mode and A_2-TO_3 envelope (~1,080 cm^{-1}). Given a larger calibration and validation sample set, it is anticipated that an optimal model resides somewhere between 20 and 223 predictors. Notably, the exclusion of the α-quartz doublet led to higher prediction error on average.

The results of the manual OLS regression are shown in Table 1 (#11; see Appendix E, Fig E.16). Kaolinite interference was not visibly evident in the silver and granite mine spectra. Not surprisingly, substantial calcite content was observed in limestone mine spectra (Fig E.14), which obscured the ability to clearly identify the presence of kaolinite using the 915 cm^{-1} band. However, the absence of easily recognizable OH stretching vibrations of kaolinite (> 3,500 cm^{-1}) in those spectra suggest that kaolin was not present [51], therefore nullifying the need for confounder correction for these samples.

Table 1 indicates a substantial increase in training and cross-validation error for the OLS model. This was complemented by a decline in the R^2 and Q^2 statistics relative to the

PLS results. Prediction testing yielded error statistics greater than all PLS models (although PLS model #8 showed comparable performance).

Discussion

Manual calibration and PLS regression

The RMSEC and RMSECV statistics from the OLS regression were either higher than or similar to the largest predictor PLS models (Table 1, #5 and #8). In other words, PLS regression predicted α-quartz comparably to the manual calibration when a broad range of wavenumbers was indiscriminately used, i.e., when wavenumbers were not screened using BMCUVE. A comparative advantage of the PLS method is revealed here, namely that the success of calibration is no longer contingent on the intervention of the practitioner, i.e. since manual calibrations may require some judgment and care when estimating the area of the quartz doublet, especially in the presence of a curved baseline. Many of the dilemmas common to manual calibration are traded for the simpler choice of a derivative filter in PLS regression.

Wavenumber selection used in tandem with PLS regression greatly improves prediction relative to OLS regression. With the exception of a single wavenumber model (not tabulated), PLS regression achieved a better prediction then the OLS regression for every model, across all performance measures for the doublet range (#1-4, columns 5-9). Additionally, PLS models developed using wavenumbers from 415 to 2,201 cm^{-1} (excluding the α-quartz doublet; #8-10) performed substantially better than the manual method, achieving the best calibration (RMSEC, RMSECV) overall. The best predictive models (RMSEP) clearly required the use of the doublet region (#1-4, #7). Ultimately, PLS approaches estimated silica dust more effectively and, when used with feature selection, acted in an exploratory capacity, i.e., BMCUVE identified hitherto unexploited silica vibrations relevant to the prediction of airborne α-quartz.

Methodological aspects of PLS regression and wavenumber selection

Figure 5.2 illustrates that training samples selected by SPXY filled the calibration space and maximized scatter about the ideal reference line. Simply using the remaining 15 samples for prediction testing failed to adequately account for dispersion within the modeling space thereby making RMSEP artificially small. In spite of this, SPXY had an unforeseen

benefit of stabilizing backward elimination by improving the odds of selecting the same wavenumbers for repeated application of the full routine.

The results of BMCUVE are not precisely reproducible due to the resampling variance inherent to Monte Carlo estimators [52]. Rerunning the entire wavenumber selection routine led to the selection of slightly different channel-wavenumbers for fixed initial conditions. Wavenumbers that were selected often resided within the same IR vibrations from trial-to-trial. In other words, the same fundamental normal modes were deemed important to the PLS modeling problem even if the exact channels were not selected. Trial-to-trial resampling variability dictates the outcome of latent variable estimation in MCCV as well; namely, repeated runs of MCCV may result in the selection of a different number of latent variables, all other things being equal. Only an infinite amount of resampling completely eliminates this effect, although reasonable performance is often achieved for a moderately sized resampling number [30]. Future studies will investigate ways to limit trial-to-trial variance as well as experiment with objective model selection criterion (e.g., Akaike information criterion) [53].

The application of routines such as BMCUVE are typically justified in the the PLS literature in terms of their ability to create more parsimonious predictive models [36, 54]. Certainly, fewer channel-variables aid in interpreting the behavior of a model's predictive performance (e.g., Fig. 5.4), improve precision (e.g., Table 1), and possibly minimize uncertainty due to the estimation of PLS parameters [55]. From a practical standpoint, parsimony for its own sake is a relatively unimportant element to a useful calibration, i.e., moderate redundancy is completely acceptable when the latent structure is well captured by the PLS components.

The 223-wavenumber regression (Table 1, row 8) aptly illustrates this point. Relative to the 20-predictor model, an insignificant change in precision is observed with spectral features spanning ~400-1,800 cm^{-1}. More importantly, two components are required to predict silica in both regressions. Although redundancy is clearly present in the larger model, the identical latent effects are captured sufficiently over the larger spectral range. The routine, on-site quantification of airborne silica might find this redundancy useful when identifying outlying samples. In fact, intentionally including wavenumbers in the regression

that correspond to known interferences (e.g., kaolinite, ~915 cm^{-1}) may facilitate the determination of silica in coal dust samples using models developed in this study.

Infrared spectra and PLS regression

The presence of uncompensated PVC bands did not hamper quantitative analysis. The retention of absorption features near α-quartz vibrations after MCUVE confirms this hypothesis. For example, modeling error was surprisingly low for the best PLS regressions (e.g., Table 1, #10) when considering the close proximity of the A_2 and TO_3 modes of quartz to a PVC band (1,086 cm^{-1} and 1,102 cm^{-1}; Fig. 5.1). Furthermore, wavenumbers from the quartz *E* mode at 695.5 cm^{-1} were discarded late in the MCUVE routine in spite of residual PVC background (C-Cl stretch at ~690 cm^{-1}) [56-58], obscuring any clear identification.

Reliable regression coefficients were identified near the TO_4, LO_4, and LO_3 quartz modes for the 139-predictor PLS model. These modes were unearthed using BMCUVE over the half-spectral range, as illustrated in Fig 5.4. Three factors complicate these band assignments, particularly for TO_4 and LO_4. They relate to interplay between PVC interference, resolution degradation due to derivative filtering, and the theoretical probability of observing TO_4-LO_4 splitting in α-quartz spectra. First, PVC has a CH_2-rocking and CH-wagging mode near the theoretical locations of the TO_4 and LO_3 modes, confounding the ability to determine visually which modes are displayed in Fig 5.4 [56, 59]. Without visual verification, it is possible that BMCUVE accidentally retained redundant PVC background in the 139-wavenumber regression.

If PVC bands were inadvertently selected for the 139-predictor regression this would indicate an early and major failure of BMCUVE. Because approximately 89% of the redundant variables were eliminated prior to the 139-predictor regression, it is unlikely that the normal operating behavior of BMCUVE was compromised. As alluded to in the discussion of Fig 5.3, wavenumber elimination appeared to behave hierarchically—wavenumbers comprising near-zero absorbance were first removed, followed by PVC background absorption bands, then by silica modes obscured by PVC, and so on. Considering this elimination order, the wavenumbers from ~1,200 to 1,230 cm^{-1} used in the PLS regression may measure a concentration-dependent variation in the TO_4 mode (although slightly perturbed by a constant PVC band contribution). Additionally, convolution of TO_4 with LO_3 features (1,252 cm^{-1}) may have resulted from the Savitzky-Golay derivative

transformation. Therefore, PLS may have measured the combined changes in LO_3 and TO_4 absorbance superimposed on a constant PVC background ($LO_3 + TO_4$ + PVC constant).

To circumvent a detailed discussion of TO-LO splitting and its impact on IR activity therein, Fig 5.5 shows a reference spectrum (1-cm^{-1} resolution) of Min-U-Sil 5 silica dust (US Silica, Berkeley Springs, WV) pressed into a KBr disk. The second-derivative spectrum reveals the presence of the suspected TO and LO modes at 1,081 and 1,166 cm^{-1} but the LO_3 mode was not resolvable in this spectrum. Calculated [44, 56, 60] and experimental spectra [61] further support these assignments. The TO_3 vibration appears quite broad in part due to the contribution from the A_2 mode at ~1,060 cm^{-1}, which is barely resolved even in the second derivative spectrum measured at 1-cm^{-1} resolution. Broadening is attributed to the random orientations of the crystalline particles within the KBr substrate, yielding mixed absorption and reflection contributions to the band. Analogous or worse artifacts are anticipated for crystallites deposited onto PVC substrates for mine samples.

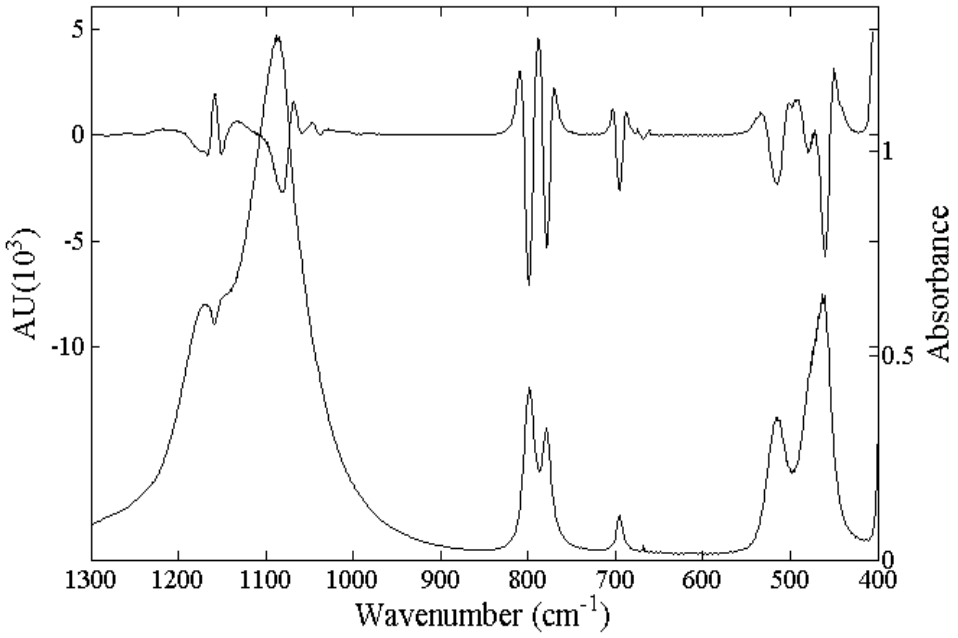


Fig. 5.5 Second-derivative (*left ordinate*) and absorbance spectra (*right ordinate*) for Min-U-Sil 5 analytical standard for silica

Conclusions

In US mines, respirable air samples are subjected to analysis for silica, and the nonexplicitly stated exposure limit for silica is an 8-hr time-weighted average (TWA) concentration of 100 $\mu g/m^3$. Precise estimates of airborne silica, especially near this limit, are

crucial for the eventual realization of end-of-shift exposure assessment. In order to prove the efficacy of a field-portable FT-IR method for the measurement of silica directly on filter samples, the interplay of preprocessing, wavenumber selection, and validation of PLS regression is paramount.

Careful preprocessing was shown to be critical to building acceptable models for the IR spectra of noncoal mine samples. PVC background correction and first-derivative transformations removed most scattering and filter interference. SPXY chose samples near the ends of the calibration range (0 and ~293 μg-SiO_2/filter). This ensured that high-concentration samples were well represented in the calibration samples. This was particularly important in this study where few dust samples were available beyond 150 μg-SiO_2/filter. An appropriate preprocessing routine, such as those using scatter correction and derivative transformation, should always precede FT-IR calibration with PLS regression, particularly for direct-on-filter α-quartz collection.

Even with some uncertainty in band assignments, wavenumbers known to constitute quartz vibrations were always incorporated into high-performing regressions by BMCUVE. Often, fewer than 50 wavenumbers realized the best predictive performance irrespective of the number of initial wavenumbers supplied to the selection routine. The BMCUVE algorithm led to the effective rejection of PVC spectral residues and other redundancies. Overall, IR absorption at the periphery of the quartz doublet led to an accurate prediction of the mass of α-quartz at the slight cost of precision and parsimony. Regardless, practical considerations may dictate that a less parsimonious model will better handle confounding absorbance, particularly for the positive interferant kaolinite.

Measuring the silica content of dust-laden filter samples using field-portable FT-IR spectroscopy appears viable using a PLS regression. In most instances, PLS regression clearly outperforms the ordinary least squares approach (MSHA P-7 analogous), which suggests potential improvement in the sensitivity and accuracy of a PLS-based analytical method for quantifying silica at end-of-shift. Notably, prediction improves when wavenumbers corresponding to known quartz vibrations are selected using BMCUVE. Given the success and minimal technical intervention of the user in the proposed PLS modeling approach, it is recommended that models 1-4 (Table 1) be utilized for end-of-shift airborne α-quartz assessment in noncoal mines.

References

[1] Mine Safety and Health Adminstration (2013) Infrared Determination of Quartz in Respirable Coal Mine Dust - Method No. MSHA P7. Pittsburgh Safety and Health Technology Center, Pittsburgh

[2] World Health Organization (1997) Silica Volume 68. In: Monographs on the Evaluation of Carcinogenic Risks to Humans. Lyon, France

[3] Steenland K, Mannetje A, Boffetta P, Stayner L, Attfield M, Chen J, *et al.* (2001) Cancer Causes Control 12: 773-784

[4] Weeks JL, Rose C (2006) Am J Ind Med 49: 523-534.

[5] Hochgatterer K, Moshammer H, Haluza D (2013) Lung 191: 257-263

[6] Calvert GM, Rice FL, Boiano JM, Sheehy JW, Sanderson WT (2003) Occup Environ Med 60: 122-129

[7] Mannetje A, Steenland K, Attfield M, Boffetta P, Checkoway H, DeKlerk N, *et al.* (2002) Occup Environ Med 59: 723-728

[8] Mazurek JM, Attfield MD (2008) Am J Ind Med 51: 568-578

[9] Leung CC, Yu ITS, Chen W (2011) Lancet 379: 2008-2018

[10] National Institute of Occupational Safety and Health (2003) Silica, Crystalline by IR (KBr pellet)- Method 7602 In: NIOSH Manual of Analytical Methods (NMAM) , 4th edn. Center for Disease Control and Prevention, Atlanta

[11] National Institute of Occupational Safety and Health (2003) Silica, Crystalline, by XRD (filter redeposition)- Method 7500 In: NIOSH Manual of Analytical Methods (NMAM) , 4th edn. Center for Disease Control and Prevention, Atlanta

[12] Madsen FA, Rose MC, Cee R (1995) Appl Occup Environ Hyg 10: 991-1002

[13] Miller AL, Drake PL, Murphy NC, Noll JD, Volkwein JC (2012) J Environ Monit 14: 48-55

[14] Miller AL, Drake PL, Murphy NC, Cauda EG, LeBouf RF, Markevicius G (2013) Aerosol Sci Technol 47: 724-733

[15] Kauffer E, Masson A, Moulut JC, Lecaque T, Protois JC (2005) Ann Occup Hyg 49: 661-671

[16] Eller PM, Feng HA, Song RS, Key-Schwartz RJ, Esche CA, Groff JH (1999) Am Ind Hyg Assoc J 60: 533-539

[17] Schwerha DJ, Orr CS, Chen BT, Soderholm SC (2002) Anal Chim Acta 457: 257-264

[18] Health and Safety Executive (2005) Crystalline silica in respirable airborne dusts Direct-on-filter analyses by infrared spectroscopy and X-ray diffraction In: Methods for the Determination of Hazerdous Substances. HSE Books, Sudbury

[19] Chen CH, Tsaia PJ, Lai CY, Peng YL, Soo JC, Chen CY, *et al.* (2010) J Hazard Mater 176: 389-394

[20] Nayak P, Singh BK (2007) Bull Mater Sci 30: 235-238

[21] Painter PC, Coleman MM, Jenkins RG, Whang PW, Walker PL (1978) Fuel 57: 337-344

[22] Scott JF, Porto SPS (1967) Phys Rev 161: 903-910

[23] Lee T, Chisholm WP, Kashon M, Key-Schwartz RJ, Harper M (2013) J Occup Environ Hyg 10: 425-434

[24] Abdi H (2010) Wiley Interdiscip Rev: Comput Stat 2: 97-106

[25] Næs T, Isaksson T, Fearn T, Davies T (2002) A User-friendly Guide to Multivariate Calibration and Classification. NIR Publications, Chichester

[26] Wold S, Sjöström M, Eriksson L (2001) Chemom Intell Lab Syst 58: 109-130

[27] Kalivas JH, Gemperline PJ (2006) In: Gemperline PJ (ed) Practical Guide to Chemometrics, 2nd edn. CRC/Taylor & Francis, Boca Raton

[28] Bye E (1992) Chemom Intell Lab Syst 14: 413-417

[29] Ritz M, Vaculikova L, Plevová E, Matýsek D, Mališ J (2012) Acta Geodyn Geomater 9: 511-520

[30] Cai W, Li Y, Shao X (2008) Chemom Intell Lab Syst 90: 188-194

[31] Isaksson T, Næs T (1988) Appl Spectrosc 42: 1273-1284

[32] Galvão RKH, Araujo MCU, José GE, Pontes MJC, Silva EC, Saldanha TCB (2005) Talanta 67: 736-740

[33] Abdi H, Williams LJ (2010) Wiley Interdiscip Rev: Comput Stat 2: 433-459

[34] Xu, HS, Liang YZ, (2001) Chemom Intell Lab Syst 56: 1-11

[35] Höskuldsson A (2001) Chemom Intell Lab Syst 55: 23-38

[36] Balabin RM, Smirnov SV (2011) Anal Chim Acta 692: 63-72

[37] Shao J (1993) J Am Stat Assoc 88: 486-494

[38] Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C (1996) Anal Chem 68: 3851-3858

[39] Ainsworth S (2005) J ASTM Int 2: 1–14

[40] Saikia B, Parthasarathy G, Sarmah NC (2008) Bull Mater Sci 31:775-779

[41] Innocenzi P (2003) J Non-Cryst Solids 316: 309-319

[42] Osswald J, Fehr KT (2006) J Mater Sci 41: 1335-1339

[43] Hirata T (1999) Solid State Commun 111: 421-426

[44] Piro OE, Castellano EE, González SR (1988) Phys Rev B: Condens Matter Mater Phys 38: 8437-8443

[45] Kirk CT (1988) Phys Rev B: Condens Matter Mater Phys 38: 1255-1273

[46] Spitzer WG, Kleinman DA (1961) Phys Rev 121: 1324-1335

[47] Ocaña M,Fornes V, Garcia-Ramos JV, Serna CJ (1987) Phys Chem Miner 14: 527-532

[48] Almeida RM, Pantano CG (1990) J Appl Phys 68: 4225-4232

[49] Berreman DW (1963) Phys Rev 130: 2193-2198

[50] Gallardo J, Durán A, Di Martino D, Almeida RM (2002) J Non-Cryst Solids 298: 219-225

[51] Prost R, Dameme A, Huard E, Driard J, Leydecker JP (1989) Clays Clay Miner 37: 464-468

[52] Efron B, Tibshirani RJ, (1993) An Introduction to the Bootstrap. Chapman & Hall, Boca Raton

[53] Burnham KP, Anderson DR (2004) Sociol Methods Res 33: 261-304

[54] Mehmood T, Liland KH, Snipen L, Sæbø S (2012) Chemom Intell Lab Syst 118: 62-69

[55] Martens H, Høy M, Westad F, Folkenberg D, Martens M (2001) Chemom Intell Lab Syst 58: 151-170

[56] Krimm S (1968) Pure Appl Chem 16: 369-388

[57] Stromberg RR, Straus S, Achhammer BG (1958) J Res Natl Bur Stand 60: 147-152

[58] Tabb DL, Koenig JL (1975) Macromolecules 8: 929-934

[59] Ramesh S, Leen KH, Kumutha K, Arof AK (2007) Spectrochim Acta, Part A 66: 1237-1242

[60] Sato RK, McMillan PF (1987) J Phys Chem 91: 3494-3498

[61] Francis S, Stephens WE, Richardson N (2009) Environ Health 8: S4 1-4

[62] Weakley AT, Griffiths PR, Aston DE (2012) Appl Spectrosc 66:519–529

Chapter 6. Model selection in partial least-squares regression using backward Monte Carlo unimportant variable elimination

Andrew T. Weakley, D.E. Aston, submitted to Chemometrics and Intelligent Laboratory Systems, April 28, 2014 (under consideration)

Abstract

Eliminating spectral channels containing artifacts and/or unwanted background often improves the precision and stability of partial least-squares (PLS) regressions. In this study, Monte Carlo unimportant variable elimination was formulated in a backward elimination manner to achieve such ends. Two articulations of this BMCUVE routine were tested on small-, moderate-, and large-sample prediction problems using mid-infrared, near-infrared, and Raman spectral channels as predictor-variables, respectively. The performance of the BMCUVE routine on each dataset was compared against the least-squares successive projections algorithm and moving window PLS methods. BMCUVE demonstrated equivalent or superior levels of predictive performance on seven comprehensive model selection criteria. These were organized hierarchically favoring prediction interval coverage probabilities at the 80%, 90%, and 95% significance levels (1-3) over a minimized standard error of validation (4), a maximized cross-validated coefficient of determination (5), a minimized Bayesian information criterion (6), and parsimony condition (7). Imposing a hierarchical protocol onto the model selection procedure helped identify predictive models showing equivalence across independent measures of standard error, lesser complexity, and improved precision over full and analyst-informed PLS regressions. These seven criteria appear applicable to estimation problems beyond PLS regression and may mitigate the impact of sample-set partitioning on the outcome of model selection.

Keywords: Model selection, prediction interval coverage probability, vibrational spectroscopy, unimportant variable elimination, Bayesian information criterion

Introduction

When spectroscopic features are used as predictors ($[X]$) in a multivariate calibration, partially- or non-selective absorbance or scattering attributes may influence the quantitative determination of a target analyte [1]. Common attributes impacting otherwise sound

experimental designs include substrate and/or ambient background, baseline artifacts, or other matrix interactions. Partial least-squares (PLS) estimators, such as the popular nonlinear iterative partial least-squares (NIPALS) [2, 3], accommodate these systematic perturbations at the cost of a more complex model. Simply stated, a PLS regression will require supplementary components to map analyte response ($\boldsymbol{y}$) onto [X] risking an inaccurate prediction of unknown future samples and obscuring a straightforward interpretation of the regression's latent structure [4-7].

Applying signal correction during the estimation of PLS components attempts to relegate inconsequential variation to a subspace uncorrelated to $\boldsymbol{y}$ [8, 9]. Unwanted variation is parsed from meaningful $\boldsymbol{y}$-correlated information while the quantity of spectroscopic predictors remains unchanged. Although clearly advantageous, conventional PLS algorithms (*e.g.*, NIPALS) are still preferred for routine multivariate calibration. Renewed interest in so-called feature selection techniques [9-14] aims to improve regression by identifying and using only germane spectroscopic predictors. Such techniques facilitate the removal of excessive redundancy, enhance precision, can reduce complexity, and improve interpretability.

A class of numerical approaches known as randomized wrapper methods (RWM) [11, 15] repeatedly passes predictors (or predictor weights) through some importance-filtering scheme until only a handful of quintessential variables are identified for a final model [16-18]. Typically, model inputs (or weights) are randomized in a manner that foregoes exhaustively searching all possible combinations of predictors (with p predictors, $2^p - 1$ combinations). Genetic algorithms often fall under the RWM rubric [17, 18]. However, randomization dramatically reduces runtime at the risk of missing the best subset of modeling variables.

An individual or limited set of criteria is usually applied to the selection of an optimal PLS model in terms of both predictors and latent variables. Typical choices might include a minimized root-mean squared error of validation (RMSEV) and/or of cross-validation (RMSECV), although other criteria have been proposed [17, 19, 20]. However common, such practices ignore, or at least leave unsubstantiated, the impact of parameter stability on the prediction of unknown, future samples [1, 21, 22]. Including more rigorous model

selection criteria appears prudent to address issues related to prediction stability and model uncertainty.

This study establishes the RWM known as Monte Carlo unimportant variable elimination (MCUVE) [16] in a backward elimination protocol. The major improvement over the basic MCUVE routine includes optimization in two-dimensions, *i.e.*, the successive assessment of predictor importance and estimation of model complexity. Optimization in this manner assumes that systematic variation is suppressed as a function of variable reduction resulting in a simpler, more stable PLS model. More importantly, we demonstrate that prizing a measure of prediction stability, as opposed to optimizing validation or goodness-of-fit statistics, better informs the practicing analyst as to the long-run, predictive value of a PLS model.

Methods

Steps 1-3: Base Monte Carlo Unimportant Variable Elimination Algorithm

Figure 6.1 illustrates the four steps comprising BMCUVE. The base, or single pass, MCUVE algorithm captures steps 1-3 and is executed in a manner analogous to Cai, Li, and Shao's [16] seminal work with major differences thusly described. A single pass of the full routine (*k*) involves completing steps 1-4, storing the appropriate estimates for validation and model selection (step 5), and iterating until the procedure reaches a stopping condition.

Prior to executing the routine, the calibration-training matrix $[X]_c$ and the number of components $LV(\#)$ must be developed and estimated, respectively. In this study, available samples were partitioned into training, validation, and, for larger sample problems ($N > 50$), prediction sets using the Kennard-Stone algorithm [23]. For small sample problems, only training and validation sets were developed. Monte Carlo cross-validation (MCCV) was used with the so-called "Wold R" criterion (= 0.9) to attain an initial estimate of $LV(\#)$ [24-27]. Information describing the number and type of predictors in $[X]_c$ and value of $LV(\#)$ was superscripted "1" and stored outside the routine for validation (Figure 6.1; "Store for validation"). Note that for *k > 1,* an updated estimate of $LV(\#)$ is determined using the faster 5-fold cross-validation.

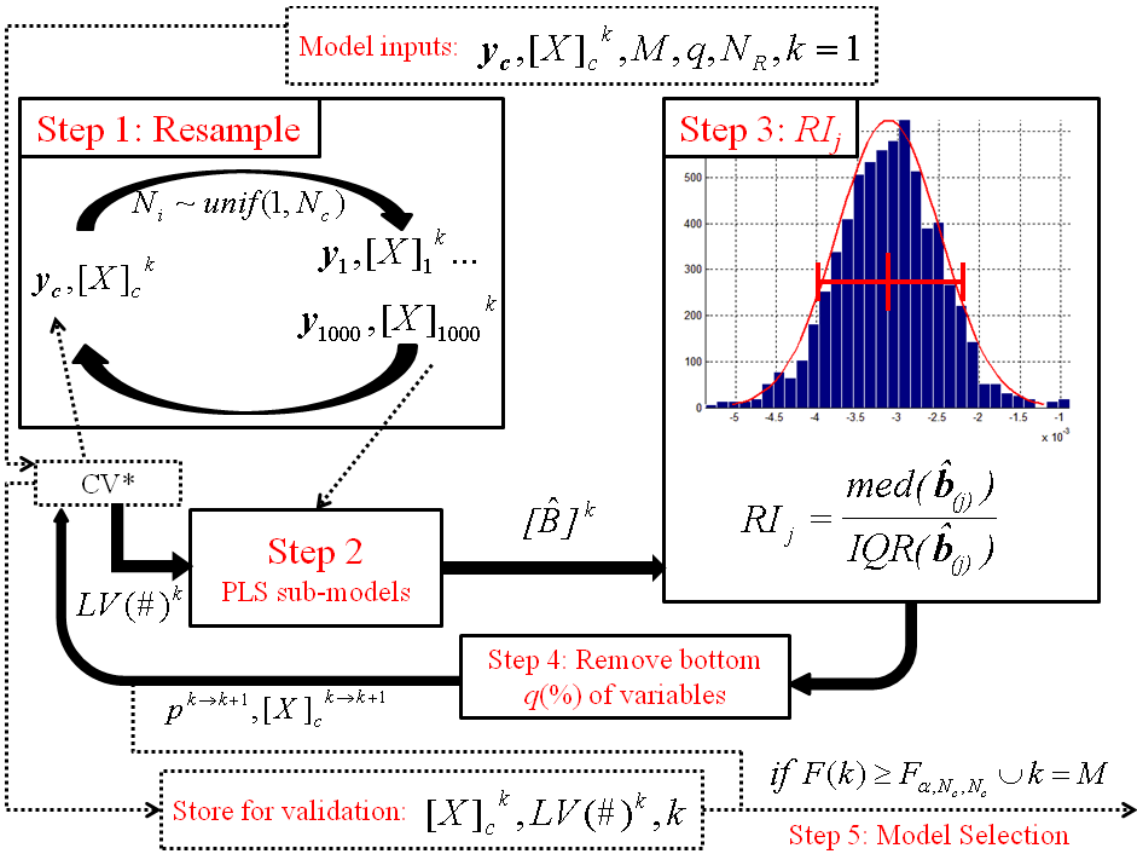


Figure 6.1: Flow diagram of the BMCUVE routine. Steps 1 through 4 indicate: (1) Monte Carlo resampling to build temporary training sets, (2) PLS sub-model and regression coefficient development, (3) ranking and sorting any regression coefficient using its empirical reliability index (RI_j), and (4) removing the q(%) least reliable coefficients according to $|RI_j|$. The design matrix of vibrational spectra, $[X]$, and PLS component number, $LV(\#)$, are superscripted (k) to indicate re-estimation *after* each pass of the routine. Validation and model selection follow elimination (Step 5). Asterisk (*) on CV indicates that method of $LV(\#)$ estimation changes from Monte Carlo to 5-fold cross-validation for $k > 1$.

Drawing a fixed proportion of the N_c available calibration samples initiates Monte Carlo sampling (Figure 6.1; Step 1). For this study, 75% of the calibration samples are randomly chosen (without replacement) and stored temporarily as $\boldsymbol{y}_i, [X]_i^{\,k}$. The subscript "$i$" denotes that the composition, *not* quantity, of each temporary training set changes per round of resampling. In this study, resampling is repeated 1000 times (=N_R) to create the corresponding 1000 temporary training sets ($\boldsymbol{y}_1 \ldots \boldsymbol{y}_{1000}$, $[X]_1^{\,k} \ldots [X]_{1000}^{\,k}$). A large resampling number ensures that any variance due to Monte Carlo sampling is minimized [28, 29].

The temporary training sets are passed to Step 2 where 1000 PLS sub-models are built for a fixed $LV(\#)^k$. Each individual PLS sub-model can be represented by

$$\boldsymbol{y}_i = [X]_i^{\,k}\widehat{\boldsymbol{b}}_i^{\,k} + \boldsymbol{\varepsilon},\ \ i = 1, \ldots, N_R \qquad \text{Equation 6.1}$$

where $\boldsymbol{y}_i$ is the temporary analyte response vector, $\widehat{\boldsymbol{b}}_i^{\,k}$ is a vector of estimated regression coefficients, and $[X]_i^{\,k}$ is a matrix of N_i randomly selected spectra (rows) on p^k predictor-variables (columns) at the k^{th} iteration. Regression coefficients estimated for each PLS sub-model are stored row-wise in a matrix $[\hat{B}]^k$ containing p^k rows and 1000 columns.

The matrix of regression coefficients is passed to step 3. Any row, $\widehat{\boldsymbol{b}}_{(j)}^{\,k}$, in $[\hat{B}]^k$ provides an estimate of the j^{th} regression coefficient's sampling distribution (histogram in Figure 6.1, Step 3). Test statistics (*e.g.*, mean, variance) succinctly describing the character of this distribution are then calculated. The reliability of each coefficient, and by extension predictor importance, is evaluated using a robust form of the reliability index [30]

$$RI_j = \frac{med(\widehat{\boldsymbol{b}}_{(j)})}{IQR(\widehat{\boldsymbol{b}}_{(j)})},\ j = 1, \ldots, p. \qquad \text{Equation 6.2}$$

Here, RI_j is for the j^{th} regression coefficient, $med(\widehat{\boldsymbol{b}}_{(j)})$ is the sample-median of $\widehat{\boldsymbol{b}}_{(j)}$ and $IQR(\widehat{\boldsymbol{b}}_{(j)})$ contains the associated interquartile range. Predictors with large regression coefficients (viz., median) and small trial-to-trial variation (viz., IQR) are *relatively* more stable than coefficients with small regression coefficients and large dispersion (small $\pm RI_j$). Ultimately, a cutoff threshold for RI_j is required to determine whether a regression coefficient, and therefore predictor variable in $[X]_c^{\,k}$, has crucial or non-essential predictive value.

Step 4: Backward Elimination

Step 4 requires a reliability cutoff followed by appropriate iteration stopping rules. In this study, a cutoff level $q(\%)$ was defined at 10% for removing the 10% least reliable predictor-variables (±). The remaining predictors were passed back to step 1. Although ultimately arbitrary, this choice of cutoff proved particularly valuable when visualizing BMCUVE's logarithmic elimination path, particularly when a large number of predictor evaluations.

Steps 1- 4 are repeated until one of two stopping criteria is reached. The simplest criterion prematurely terminates elimination when M iterations are reached (Figure 6.1,

$k = M$). If desired, the following formula can be used to estimate the number of iterations required

$$M = \left. log_{10}\left(\frac{p_f}{p_i}\right) \middle/ log_{10}\left(1 - \frac{q(\%)}{100}\right)\right. . \qquad \text{Equation 6.3}$$

This shows that the number of passes (M) are required for a desired, final number of predictor-variables (p_f) given p_i initial predictors. Equation 6.3 must be rounded to an integer prior to execution and reduces to $M = \frac{-log_{10}(p_i)}{log_{10}\left(1-{}^{q(\%)}/_{100}\right)}$ when elimination proceeds to a single predictor.

Rounding influences the risk of BMCUVE performing redundant iterations. This is particularly true when the difference between successive predictor-subsets is small, *e.g.*, removing 90% of 3 variables (Figure 6.1; Step 4) results in the routine retaining 2.7 variables for the next pass and rounding to 3. To counter, an auxiliary elimination rule automatically reduces repetitive subsets by an additional unreliable predictor (*e.g.*, a subset of 3 predictors is reduced to 2), which obviously influences the accuracy of Equation 6.3.

The second stopping criterion compares the relative error of the current (k^{th}) cross-validated model to the smallest observed model error up to and including the k^{th} model. A two-sample variance test employs the sample statistic

$$F(k) = \frac{MSECV_k}{MSECV_{min}} \qquad \text{Equation 6.4}$$

to identify the termination point. Here, $F(k)$ is the sample *F*-statistic comparing the mean-squared error of cross-validation for the k^{th} PLS model ($MSECV_k$) to the PLS model with a minimum MSECV observed up to and including the k^{th} model ($MSECV_{min} \in 1 \ldots k$). A one-sided hypothesis is tested using Schnedecor's *F*-distribution with N_c-1 and N_c-1 degrees of freedom, $F_{\alpha,Nc-1,Nc-1}$ [31]. The current model's performance is significantly worse if $F(k) > F_{\alpha,Nc-1,Nc-1}$. In other words, a significant *F*-statistic offers evidence that predictor elimination has proceeded too far and has begun removing reliable information from the model. At just this point, the algorithm is terminated with only the $k - 1$ previous models retained for validation. In this study, the significance level (α) was set to 0.0001 allowing backward elimination to proceed almost entirely unconstrained until only a single predictor remained.

Step 5: Validation and Model Selection

Seven criteria, ranked in order of importance, facilitated selecting a best predictive model. These criteria include prediction interval coverage probabilities ($PICP_{1-\alpha}$) at or above three nominal $(1-\alpha)^{\text{th}}$ levels (1-3) [22], a small RMSEV (4), a large cross-validated coefficient of determination (Q^2) (5), a small Bayesian information criterion (6) [32], and the smallest subset of LVs and predictors meeting the above criteria (7). In other words, performance measures using independent validation samples were prioritized (1-4) followed by cross-validated (5) and calibration-only (6) criteria. Once a model was chosen, a true measure of predictive performance utilized the third independent prediction set to estimate the root-mean-squared error of prediction (RMSEP).

Validation can, at most, consider M different PLS models chosen during elimination (Equation 6.3). In the interest of rigor and stability, $PICP_{1-\alpha}$ measures succinctly qualify model uncertainty and its theoretical impact on predicting unknown samples. The $PICP_{1-\alpha}$ expression takes the form

$$PICP_{1-\alpha} = \sum_{l=1}^{N_v} I(|\, y_l - \hat{y}_l| < t_{1-\frac{\alpha}{2},N_c-df}\hat{\sigma}_l)/N_v \qquad \text{Equation 6.5}$$

where $PICP_{1-\alpha}$ is the prediction interval coverage probability at the $(1-\alpha)^{\text{th}}$ level, $y_l - \hat{y}_l$ is the prediction residual for the l^{th} independent validation sample, $t_{1-\frac{\alpha}{2},N_c-df}$ is the Student's *t*-distribution with the appropriate degrees of freedom (df), N_v is the number of validation samples, and $\hat{\sigma}_l$ is the estimated standard deviation of the PLS prediction error [21, 22]. An indicator variable, $I(*)$, equals 1 if the condition in parentheses is true and zero otherwise. In other words, $PICP_{1-\alpha}$ summarizes the proportion of validation samples falling within a theoretical prediction interval at a given significance level. As a model selection criterion, a $PICP_{1-\alpha}$ statistic closer to the $(1-\alpha)^{\text{th}}$ level offers evidence that a model has better predictive ability than an alternative for a given $\hat{\sigma}_l$. Ultimately, $PICP_{1-\alpha}$ empirically validates the stability of $\hat{\sigma}_l$ where

$$\hat{\sigma}_l = \sqrt{\frac{MSEC}{N_c} + \boldsymbol{x}_l{}^T var(\widehat{\boldsymbol{b}})\boldsymbol{x}_l + MSEC} \qquad \text{Equation 6.6}$$

is estimated from calibration data for a centered (and optionally scaled) latent variable model. Here, the mean-squared error of calibration, or $MSEC$ $(= \frac{\left\|\hat{\boldsymbol{y}}_c{}^k - \boldsymbol{y}_c{}^k\right\|^2}{N_c-df})$ estimates the standard

error of calibration, $\boldsymbol{x}_l$ is the spectrum of the l^{th} validation sample, and $var(\widehat{\boldsymbol{b}})$ is the covariance of the regression coefficients.

Coverage probabilities, were calculated at the 80%, 90%, and 95% levels for each model. The required $var(\widehat{\boldsymbol{b}})$ was estimated using a bootstrapping-by-residuals method [28]. The generalized degrees of freedom (GDF) approach provided an estimate for df in Equation 6.5 as well as for the $MSEC$ [33].

The Bayesian information criterion (BIC) [32, 34] was also calculated to aid model selection. The BIC measures the relative performance of a range of alternative models using information contained only in the calibration data penalized by the number of regression parameters (df). A computationally convenient form

$$BIC = MSEF\left(N_c + log_{10}(N_c)\frac{df}{1-\frac{df}{N_c}}\right) \qquad \text{Equation 6.7}$$

requires estimating the mean-squared error of fit ($MSEF = \frac{\left\|\hat{\boldsymbol{y}}_c{}^{k} - \boldsymbol{y}_c{}^{k}\right\|^2}{N_c}$), the number of calibration samples (N_c), and df. When used alone for model selection, a minimized BIC indicates a relatively better model than alternatives.

The high computational cost of calculating the GDF and $PICP_{1-\alpha}$ for (possibly) M models prompted using the pseudo degrees of freedom (PDF) approximation [35] to initially narrow down viable models according to the seven aforementioned criteria. Model selection (Step 5) often concerned between three and five viable candidates that scored well on the seven criteria. These candidates were next subjected to a rigorous screening via $PICP_{1-\alpha}$ measures using the GDF and a RMSECV and $LV(\#)$ precisely confirmed using MCCV.

Filtering with Bootstrap Confidence Intervals (BCIs)

A single-pass filtering of predictors prior to BMCUVE was also explored in the interest of reducing runtime. To this end, nonparametric 95% bias-corrected and accelerated BCIs were developed for each regression coefficient in the full ($[X]_c{}^1$) PLS model using 5000 bootstrap replicates [28, 36]. As a filter, confidence intervals facilitate two one-sided hypothesis tests. Simply, if the BCI around the j^{th} regression coefficient contained zero, this suggested zero as a plausible value for the true (but unknown) coefficient. Thus, a predictor was removed from $[X]_c{}^1$ if its coefficient's BCI contained zero.

Algorithm Test Data

Mid-infrared (IR) transmission, near infrared reflectance (NIR), and micro-Raman spectra have been successfully employed to predict airborne silica (μg-SiO_2) in non-coal mine dusts, moisture content in corn (%H_2O) [37], and hard-segment fraction from thermoplastic polyurethane (TPU) blends [38], respectively. These three data sets were chosen to test BMCUVE and the model selection criteria given a diversity of possible background, baseline, and artifact interferences as well as the size of each calibration problem. Specifically, the small IR data set contained 44 total samples, the NIR data set contained 80 samples, and larger Raman set contained 220 total samples. In the interest of manuscript length, all information related to data preprocessing is confined to supporting information.

Six separate calibrations were performed on each of the three data sets including the BCI filtered and regular BMCUVE routines. The BMCUVE routines were judged against two capable variable selection algorithms including the successive projections algorithm (SPA) and interactive moving window PLS (MWPLS) method [37, 39]. Given that the MWPLS method requires some user judgment, no greater than five of the best MWPLS models were weighted and averaged using a non-negative least-squares routine to ensure the best possible outcome [40]. The moving window size was fixed at 21 variables for all datasets.

The fifth calibration scenario tested whether expert knowledge about the calibration problem yielded better results than BMCUVE. Predictors comprising relevant bands from the mid-IR (~751-857 cm^{-1}) and Raman calibrations (~1354-1797 rel. cm^{-1}) were manually selected and exclusively used for PLS modeling. Manual wavelength selection for corn- NIR prediction was not attempted given the admitted lack of knowledge by these authors. A final full calibration tested the performance of indiscriminately using all available predictor-variables.

The NIPALS, MCUVE, PLS cross-validation, and MWPLS algorithms were available in the open-source, libsPLS (v. 1.6, Changsha Nice City, China) Matlab package. The inverse *t*-distribution, and bootstrap functions were available in either the base Matlab 2011a (The Mathworks, Nattick, MA) software or Statistics Package. Programs were written

in the Matlab language to calculate GDF, SPA, $PICP_{1-\alpha}$, and BIC as well as to fully develop each step of the BMCUVE procedures.

Results

Table 6.1 compares BMCUVE selection to the SPA, MWPLS, manual, and full PLS procedures. The regular BMCUVE solution required the fewest number of components showing a small RMSEP (= 0.00172). The BCI-BMCUVE appears virtually as capable in predicting hard segment content in TPU blends. Manually selecting Raman features for calibration ("Manual PLS") appears to fare the worst on all measures of performance. Ultimately, using computational or a semi-guided procedure (MWPLS) for predictor selection is warranted for these TPU blend spectra.

Table 6.1: Model selection results for the calibration, validation, and prediction of TPU hard segment fraction using Raman spectra. The selection method, predictor and PLS-component number (Var(#),LV(#)), RMSE statistics, coverage probabilities, and BIC are labeled accordingly.

Method	Range (rel cm^{-1})	Var(#)	LV(#)	Q^2	RMSEC (10^3)	RMSECV (10^3)	RMSEV (10^3)	RMSEP (10^3)	PICP80	PICP90	PICP95	BIC (10^4)
BCI-BMCUVE	Figure 2,3 (•)	50	6	**0.989**	1.76	**1.86**	1.88	1.82	**0.82**	0.89	**0.95**	3.73
BMCUVE	Figure 2,3 (○)	50	**5**	0.989	1.75	**1.86**	1.81	1.72	**0.82**	**0.92**	0.94	3.68
SPA	Figure 2,3 (Δ)	**25**	-	0.975	1.84	2.74	**1.68**	1.86	**0.89**	**0.98**	**0.98**	4.62
MWPLS	[777.6 - 831.9] [1707.8 - 1761.2] [2817.2 - 2859.2] [3006.3 - 3047.1]	105	8	0.979	2.09	2.52	2.2	1.93	**0.80**	**0.91**	**0.98**	5.69
Manual PLS	[1354.4 - 1797.4]	180	6	0.972	2.7	3.07	2.47	2.14	**0.86**	**0.98**	**0.98**	9.28
Full PLS	[-54.5 - 3682.0]	1580	8	0.989	**1.62**	1.87	1.74	**1.44**	**0.83**	0.89	**0.97**	**3.63**

A full PLS regression showed the smallest RMSEC, RMSEP, and BIC using 8 components. Notable differences across standard errors (RMSE) are visible. Coincidence of a minimized BIC and RMSEC is not surprising given that both criteria use calibration data and the same estimated degrees of freedom. Overall, the BMCUVE and SPA methods required the fewest predictors. As a more appropriate measure of parsimony, PLS models generally required fewer degrees of freedom than the ordinary least-squares model chosen by SPA (*e.g.*, $GDF_{BMCUVE} = 9.8$, $DF_{SPA} = 25$). Even so, performing model selection had notable benefits including the removal of extraneous spectroscopic features/factors corresponding to scattering artifacts, baseline perturbations, and/or inconsequential blend-specific interactions [38].

$PICP_{1-\propto}$ statistics that met their $(1-\alpha)^{\text{th}}$ significance levels tended to show balanced calibration, validation, and prediction standard errors (*e.g.*, RMSEC ≅ RMSEV). This promising result indicates that even approximate $PICP_{1-\propto}$ measures (Equation 6.6) can adequately validate prediction uncertainty ($\hat{\sigma}_l$) at a given level of precision. In other words, good coverage on $PICP_{1-\propto}$ statistics indicates that a model will *consistently* predict unknown, future samples according to a given error distribution, in this case $e_l \sim t(N_c - df, \hat{\sigma}_l)$, but offers no insight into the *precision* of a predictive model. Simply comparing the estimated standard errors for the MWPLS and BCI-BMCUVE models confirms this sentiment: BCI-BMCUVE selection shows slightly improved precision relative to MWPLS ($RMSEC_{BMCUVE} < RMSEC_{MWPLS}$) but both methods cover validation samples adequately. Hence, a proper optimization still requires input from more customary validation statistics (*e.g.*, RMSEV) to select a best model.

Figures 6.2 and 6.3 present pertinent ranges from an example TPU Raman spectrum with wavenumbers chosen by the two BMCUVE routines and SPA. Clearly, each routine identifies channels on normal modes related to polymer hard segment content [41]. For example, each protocol selected spectral features from hard segment aromatic C-H stretches (~1605 rel. cm^{-1}) as well as the CH_2 deformations and wagging modes between 1180-1400 rel. cm^{-1}. Both BMCUVE routines selected many of the same variables (circled bullets) indicating that, at least for large sample problems, applying a BCI filter prior to BMCUVE achieves a similar outcome at a reduced computational cost (16.3 min vs. 23.25 min).

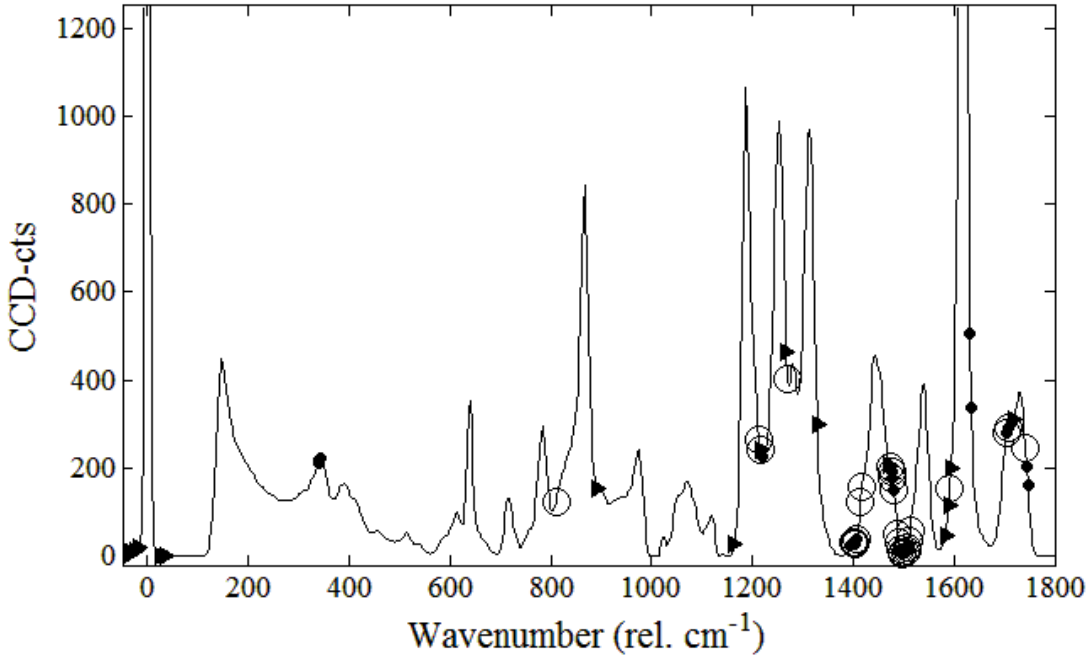


Figure 6.2: Wavenumbers selected using the regular BMCUVE (○), BCI + BMCUVE (•), and SPA (►) procedures for hard segment prediction. First-derivative spectra were actually used in each calibration thus explaining the location on this example TPU Raman spectrum. Only half of the spectral range utilized by each procedure is shown to improve perspective.

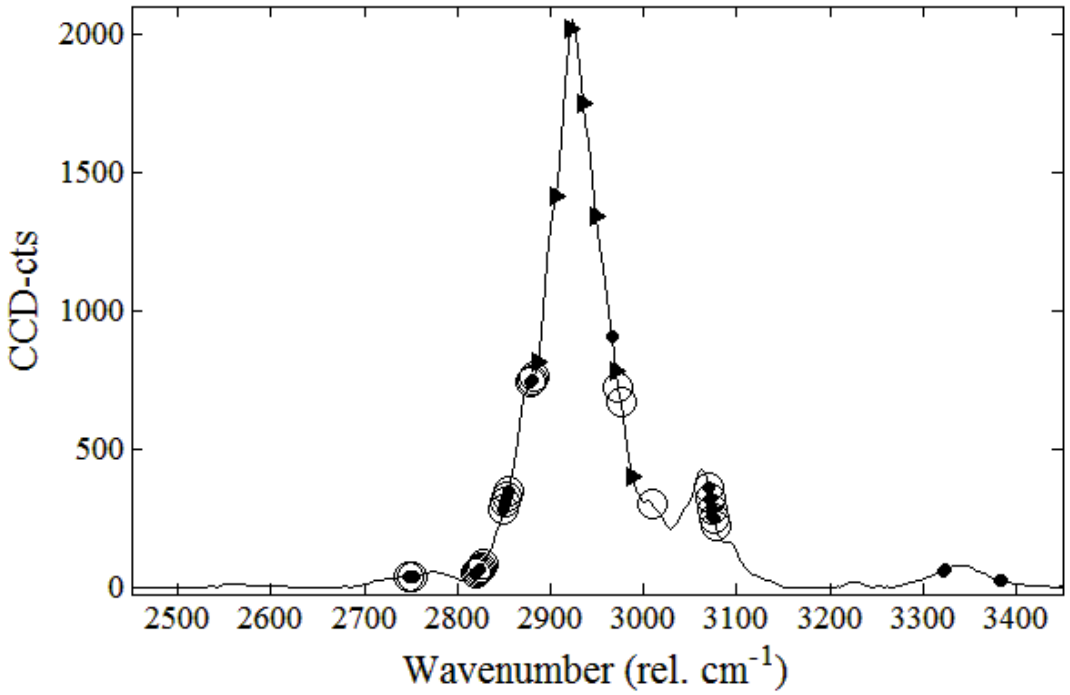


Figure 6.3: Pertinent, remaining range in the example TPU Raman spectrum (2400-3450 rel. cm-1). The regular BMCUVE (○), BCI + BMCUVE (•), and SPA (►) predictors are highlighted accordingly. The 1800-2400 rel. cm-1 and 3450 - 3600 rel. cm-1 ranges were excluded due to an absence of both Raman fundamentals and selected features.

Roughly one-quarter of the predictors selected by the SPA resided on the Rayleigh (laser) line (±50 rel. cm^{-1}). This indicates that the SPA found coherent scattering variations highly reliable relative to many of the Raman modes. Regression performance was not hampered by using these channels (see Table 6.1), but the predominance of these modes in the model may have led to large over-coverage on $PICP_{90}$ and high cost (DF_{SPA} = 25). Although masked by the SPA's selected points, the BCI-BMCUVE routine was not immune to the large influence of coherent scattering effects selecting a single point at roughly 27 cm^{-1}.

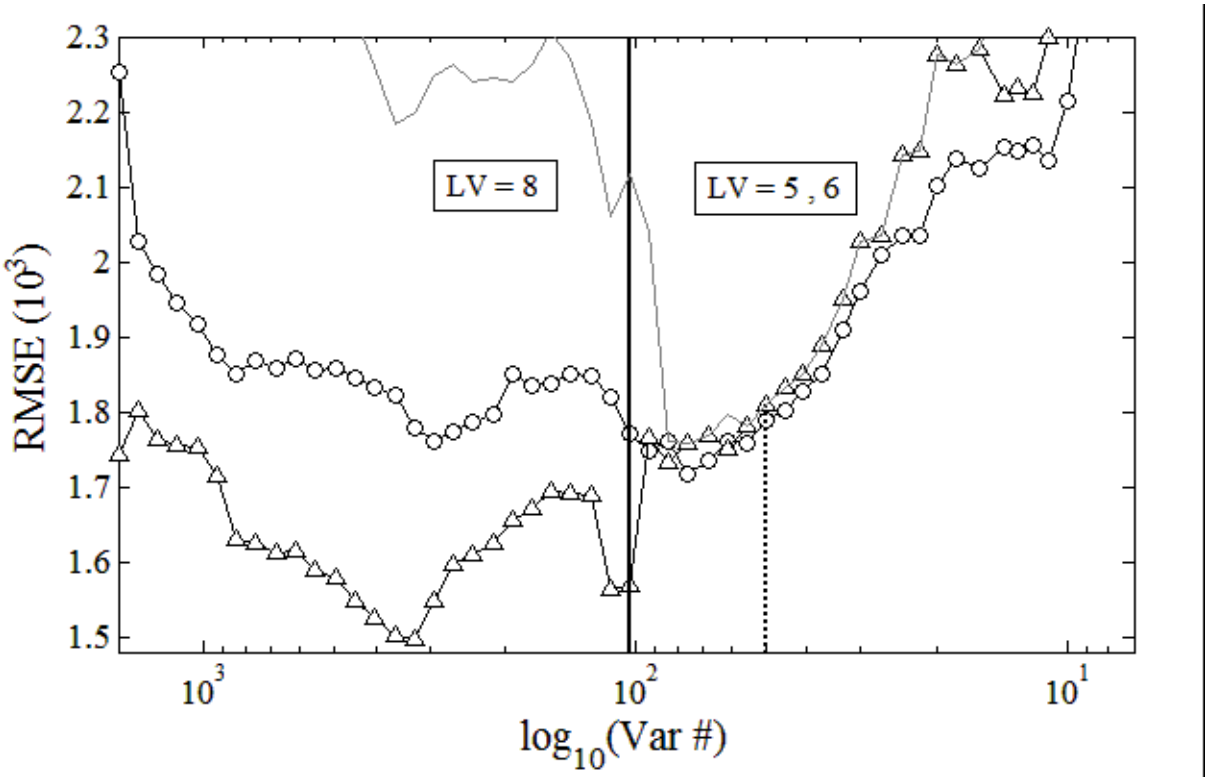


Figure 6.4: Semi-log plot of the predictor elimination path for the regular BMCUVE algorithm applied to the TPU hard-segment data set. Notably, the 50-predictor model (dashed line) was observed when the cross-validated error (○) roughly balanced the validation error (Δ). The solid black line demarcated a transition from an 8 to a 5-6 LV model. The grey series corresponds to RMSEV statistics derived by fixing the number of LVs to 5.

Visualizing error reduction as a function of predictor elimination assists in determining candidates for an optimum PLS model. Figure 6.4 illustrates an example of such a procedure for regular BMCUVE. As variable elimination proceeds, a gradual drop in the RMSECV error (○) is observed until a global minimum is reached at 76 predictors and 6 LVs. A local minimum, nearly identical in magnitude, is reach near 293 predictors for an 8

LV model. The RMSEV (Δ) reaches a minimum at 325 variables. Customary model selection might assign a global optimum at or between 293-325 predictors using 8 LVs. However, perusing $PICP_{1-\propto}$ measures illustrated that a better model—that is, one with theoretically greater long-run predictive value—is located on 50 predictors using 5 LVs (dashes). Notably, at 50 predictors we see that the cross-validated and validation errors are nearly equal.

Prioritizing $PICP_{1-\propto}$ statistics over a minimized RMSEV might appear disconcerting given the large gap between the 325 predictor RMSE minimum and 50 predictor model. An auxiliary RMSEV error profile (grey) demonstrates the impact of fixing the number of LVs to 5 and recalculating standard errors. Here, the 50 predictor model is located near the validation minima spanning 55-93 predictors. This profile demonstrates that our multifaceted optimization procedure concerns not only parsimony/minimization over a two-dimensional search domain (LVs and predictors) but, more importantly, stability in the long-run prediction of analytical quantity. According to our seven-point optimization criteria, the 50 predictor model exhibits far better properties than either validation minima at 76 or 325 predictors.

Figure 6.5 demonstrates a major strength of prioritizing $PICP_{1-\alpha}$ statistics for model selection: complete baseline and redundancy removal. The 50 and 325 predictors used in the 5-LV $PICP_{1-\alpha}$ and 8-LV RMSEV-minimized models (Figure 6.4) are plotted on an example TPU Raman spectrum. Superimposing these features provides evidence that at least two of the three LVs selected for the 325-predictor solution were contributions from both Rayleigh scattering (±50 rel. cm^{-1}) and residual baseline (~2000-2600 rel. cm^{-1}). The additional LV required for the BCI-BMCUVE (Table 6.1, Figures 6.2 and 6.3) model is now also easily justified as a Rayleigh scattering factor.

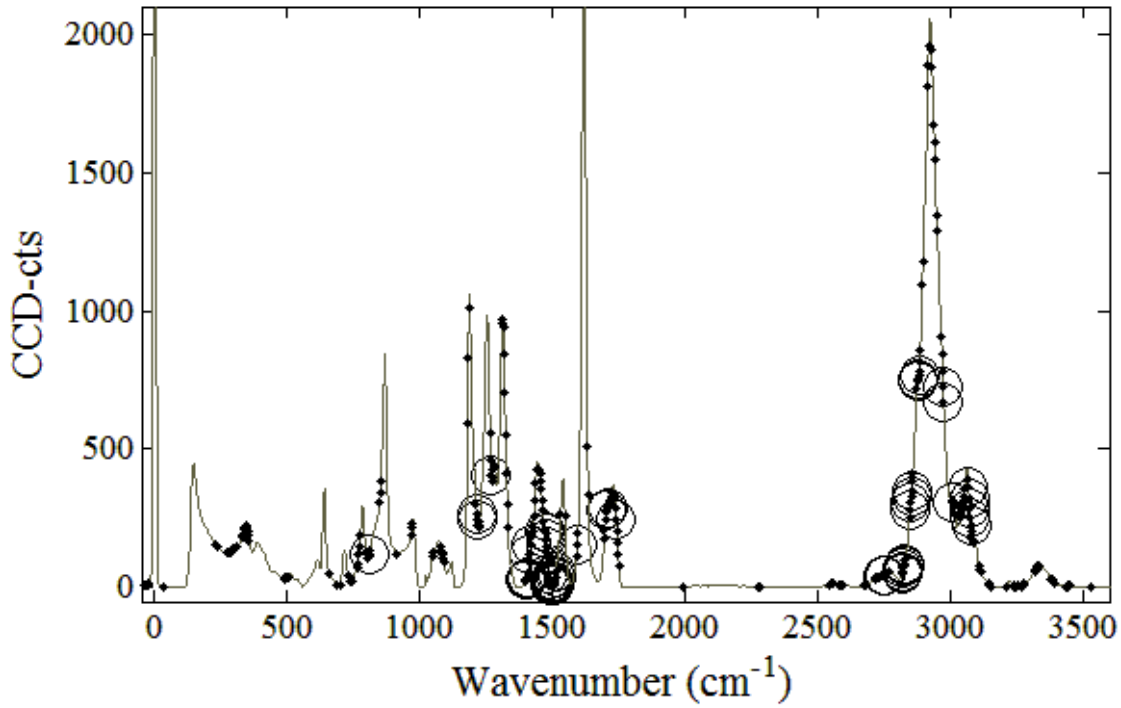


Figure 6.5: Predictors selected using the $PICP_{1-\alpha}$ statistics (50 predictors, 5 LVs) and RMSEV minimum (325 predictors, 8 LVs).

Table 6.2 summarizes model selection for predicting moisture content in corn using NIR spectra. Both BMCUVE routines achieve the best overall regression performance across RMSE statistics, Q^2, while using a minimal number of LVs. According to the seven-point criteria, BCI-BMCUVE performs slightly better than the regular BMCUVE by meeting coverage on all $PICP_{1-\alpha}$ measures while showing a lower RMSEP at the cost of two additional LVs. The model chosen by SPA again indicates that stable, lower-precision (RMSEP = 0.023) models are attainable using coverage probabilities. Specifically, SPA met and/or was closer to the nominal $PICP_{1-\alpha}$ targets than either BMCUVE solution. Good coverage for the MWPLS model, which exhibited a 10-fold reduction in precision, is also indicated.

Table 6.2: Model selection results for the calibration, validation, and prediction of moisture content (%H_2O) in corn using NIR spectra.

Method	Range (nm)	Var(#)	LV(#)	Q^2	RMSEC (%H_2O)	RMSECV (%H2O)	RMSEV (%H2O)	RMSEP (%H2O)	PICP80	PICP90	PICP95	BIC (10^2)
BCI-BMCUVE	Figure 6 (•)	98	9	0.997	0.014	0.020	**0.015**	**0.014**	**0.83**	**0.92**	**1.00**	0.96
BMCUVE	Figure 6 (○)	93	**7**	**0.998**	0.015	**0.018**	**0.015**	0.015	0.79	**0.96**	**1.00**	1.04
SPA	Figure 6 (Δ)	**17**	-	0.995	**0.012**	0.026	0.017	0.023	**0.83**	**0.92**	**0.96**	**0.75**
MWPLS	[1354 - 1394] [1848 - 1888] [2000 - 2040]	63	10	0.847	0.083	0.150	0.117	0.109	**0.83**	**0.92**	**0.96**	35.58
Full PLS	[1120 - 2478]	680	12	0.993	0.014	0.033	0.023	0.020	**0.92**	**1.00**	**1.00**	3.07

Figure 6.6 illustrates those predictors chosen for determining moisture content in corn using the BMCUVE routines and SPA. Notably, the predictors selected by SPA predicted corn analogously to Galvão and coworkers' [37] study in terms of RMSEP. Although unavailable, the wavelengths selected by SPA would probably correspond to the many of the same NIR wavelengths. An exact reproduction was not observed given methodological differences related to sample-set partitioning as well as the model selection criteria used; namely, $PICP_{1-\alpha}$ measures were used to identify a model optimum in this study.

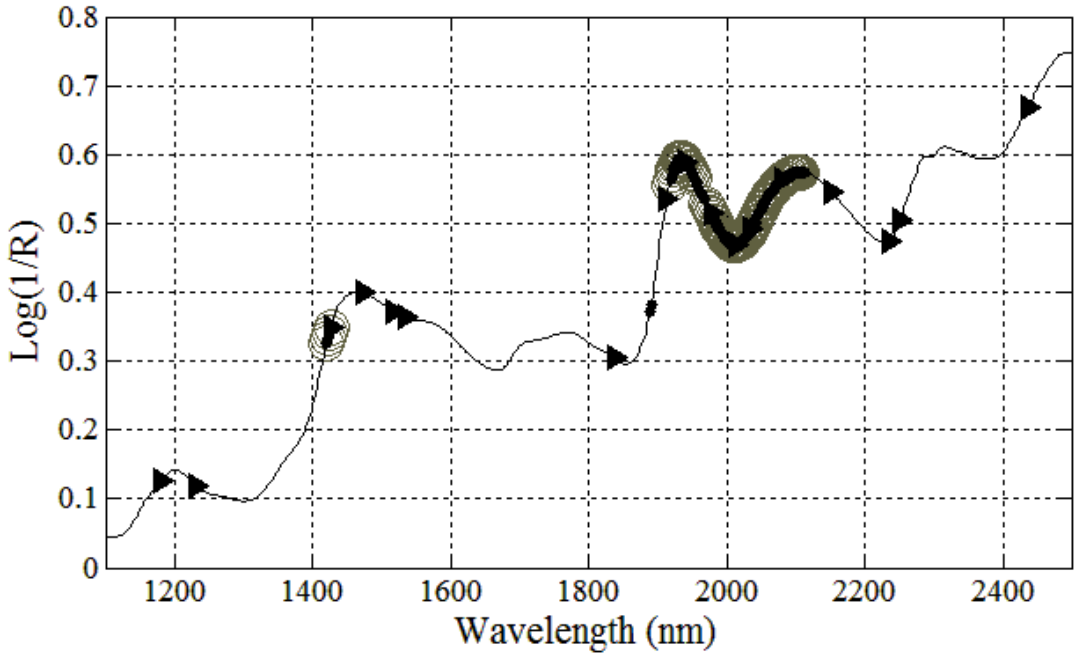


Figure 6.6: Average NIR infrared spectrum (N = 80). The regular BMCUVE (○), BCI filtered BMCUVE (•), and SPA (►) predictors are highlighted accordingly.

Table 6.3 presents the model selection results for airborne silica calibrations using transmission FT-IR. Again, a nearly identical performance of the BMCUVE methods is instantly apparent. The near equivalence across RMSE statistics for each BMCUVE routine is also observed. Absorbance features from so-called α-quartz doublet (~751-857 cm^{-1}) were selected by all methods. Notably, regulatory enforcement of airborne silica in the US mandates the use of this doublet [42]. Although difficult to confirm given the small sample size, doublet features appear a necessary and sufficient precondition for a PLS model to achieve adequate coverage on $PICP_{1-\alpha}$ measures.

Table 6.3: Model Selection results for the calibration, validation, and prediction of airborne silica (µg SiO2) content in filter-adsorbed M/NM mine samples.

Method	Range (cm^{-1})	Var(#)	LV(#)	Q^2	RMSEC (µg-SiO_2)	RMSECV (µg-SiO_2)	RMSEV (µg-SiO_2)	RMSEP (µg-SiO_2)	PICP80	PICP90	PICP95	BIC
BCI-BMCUVE	[447.6 , 449.0] [771.9-783.3] [801.7 - 815.9]	26	**1**	0.979	8.21	8.72	8.82	-	**0.80**	0.87	0.93	2029.2
BMCUVE	[449,450.4] [771.9 - 781.9] [803.1 - 817.3] [1215.3]	22	**1**	**0.979**	**8.16**	**8.64**	8.83	-	**0.80**	0.87	0.93	**1998.1**
SPA	[791.8] [872.5]	**2**	-	0.969	9.39	10.49	8.68	-	**0.87**	0.87	0.87	2640.3
MWPLS	[776.2-818.7]	42	**1**	0.974	8.73	9.64	**7.63**	-	**0.80**	**0.93**	**1.00**	2017.2
Manual PLS	[750.7 - 856.9]	76	2	0.974	8.81	9.60	7.94	-	**0.80**	**0.93**	0.93	2361.5
Full PLS	[415.0 - 3985.8]	2522	3	0.951	11.54	13.32	12.14	-	**0.87**	0.87	**1.00**	4305.2

A small number of samples (N = 44) prevented an ability to partition the sample data into three sets for calibration, validation, and prediction. The absence of RMSEP statistics renders it difficult to assess the true performance for each method. Regardless, if the TPU and corn calibrations act as any guide, the BMCUVE routines, MWPLS, and manual PLS performed identically (within error) according to the seven model selection criteria. Notably, the number of validation samples are also few (N_v = 15). This influences the dependability of $PICP_{1-\propto}$ measures for model selection in that, for example, only two validation samples separate acceptable coverage at the 80% (12/15) and 90% (14/15) significance levels.

Discussion

Cai, Li, and Shao's MCUVE protocol [16]constituted forward variable selection with variables incrementally added to $[X]_c$ until prediction no longer improved. Furthermore, the number of PLS components were estimated prior to running MCUVE and remained fixed during the duration of forward selection. This essentially one-dimensional optimization procedure relied on a reliability cutoff level determined *a posteriori* from validation testing.

Conversely, Figure 6.1 illustrates that BMCUVE employs backward variable elimination and requires a predefined cutoff level. BMCUVE also integrates a refined estimate of LV number into each pass of the algorithm under the assumption that removing artifacts and background reduces model complexity. This assumption appeared valid and spectroscopically justifiable for the hard segment polymer and airborne silica prediction, in particular. However, this now two-dimensional optimization problem required extensive validation via the seven model selection criteria.

Table 6.4: Performance of the 50 predictor TPU calibration as a function of LV(#) chosen using the regular BMCUVE procedure. Nearly every model met their nominal coverage targets limiting the selectivity of $PICP_{1-\alpha}$ statistics.

Var(#)	LV(#)	Q^2	RMSEC (10^3)	RMSECV (10^3)	RMSEV (10^3)	PICP 80	PICP 90	PICP 95	BIC (10^4)	GDF
50	1	0.804	7.52	7.74	7.11	**0.82**	**0.94**	**0.98**	6.45	3.9
	2	0.885	5.84	5.93	5.35	**0.82**	**0.94**	**0.98**	3.88	3.5
	3	0.986	1.94	2.06	1.91	**0.86**	**0.92**	**0.98**	0.44	5.6
	4	0.989	1.78	1.85	1.89	0.79	**0.92**	0.94	**0.37**	5.7
	5	0.989	1.75	1.86	1.81	**0.82**	**0.92**	0.94	0.37	9.8
	6	0.988	1.75	1.93	1.80	**0.83**	**0.91**	**0.95**	0.38	14.7
	7	0.987	1.76	2.01	1.77	**0.88**	**0.92**	**0.97**	0.40	18.4
	8	0.986	1.77	2.09	1.82	**0.86**	0.88	**0.95**	0.42	23.9
	9	0.985	1.77	2.14	1.85	**0.83**	**0.91**	**0.97**	0.43	27.3

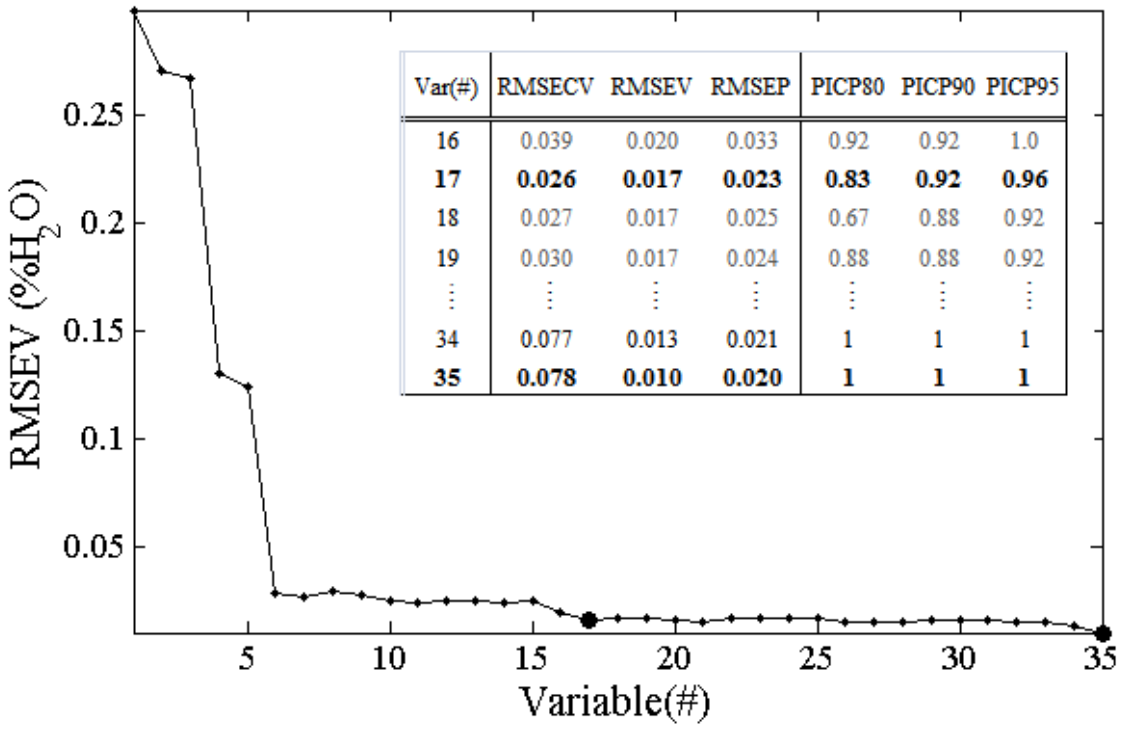
Prioritizing $PICP_{1-\alpha}$ statistics and RMSEV was vital to the selection of stable models with high levels of precision. This proved true irrespective of target analyte or the means of spectroscopic measurement. Although these four criteria generally sufficed, Table IV examines an instance where model selection demands reliance on lesser criteria. Specifically, $PICP_{1-\alpha}$ measures occupy a place of less *selective* importance than otherwise, *i.e.*, each model usually covers the validation samples at all three significance levels. Choosing candidates near $PICP_{1-\alpha}$ targets with low RMSEV suggested that a 4-6 LV model is relatively better than the alternatives. Ultimately, all criteria were required to inform the final choice of a five LV regression. Specifically, the 5-LV model exceeded or nearly met all coverage targets (1-3), showed a negligibly larger RMSEV relative to the 6-LV case (4),

exhibited a large Q^2 statistic (5), nearly minimized the BIC (6), and was more parsimonious GDF (= 9.8) than the 6-LV case (7).

Table IV illustrates that the BIC is more informative when model selection is not straightforward. Because the BIC is penalized by the cost of regression (GDF) and dependent on the magnitude of the naïve standard error ($MSEF$), five additional degrees of freedom pushed the BIC higher for the 6-LV model relative to its 5-LV counterpart. However, the reduced cost of the 4-LV model (GDF = 5.7) pushed the BIC slightly lower than the 5-LV case, in spite of a larger RMSEC ($MSEF = RMSEC^{2\frac{N_c - GDF}{N_c}}$). In the final analysis, the BIC distinguished a subset of viable candidates when the superior, validation set based criteria proved non-selective.

The current study faithfully replicated the SPA procedure outlined by Galvão and coworkers' [37] with the crucial exception of sample-set partitioning and the criteria utilized for model selection. Using the same NIR data, preliminary validation was performed to select a viable regression from $N_c \; x \; p_i$ models according to a minimized RMSEV. Once calculated, a close analog to the reliability index (Equation 6.2) was used to rank and sort predictors according to importance. Predictive models were next constructed by cumulatively adding predictors until a scree plot (*i.e.*, predictor # vs. RMSEV) was generated. Model selection then applied a one-sided variance test using Schnedecor's *F*-distribution ($F_{\alpha,Nv,Nv}$) to identify where parsimony and precision were balanced [43].

Figure 6.7 highlights that our *F*-test/scree procedure does not indicate a 17-predictor model as per Galvão and coworkers SPA study [37]. In fact, the *F*-test indicates that eliminating variables prior to 35 predictors is unwarranted. Indeed, the RMSEV is substantially lower for the 35-predictor regression than competing models (RMSEV = 0.01). Notably, the RMSECV is uncharacteristically large (RMSECV/RMSEV = 7.8) possibly indicating a regression over-fit. Incompatibility between the RMSECV and RMSEV might suggest an unstable predictive model and therefore yield a poor prediction of unknown samples.



Var(#)	RMSECV	RMSEV	RMSEP	PICP80	PICP90	PICP95
16	0.039	0.020	0.033	0.92	0.92	1.0
17	**0.026**	**0.017**	**0.023**	**0.83**	**0.92**	**0.96**
18	0.027	0.017	0.025	0.67	0.88	0.92
19	0.030	0.017	0.024	0.88	0.88	0.92
⋮	⋮	⋮	⋮	⋮	⋮	⋮
34	0.077	0.013	0.021	1	1	1
35	**0.078**	**0.010**	**0.020**	**1**	**1**	**1**

Figure 6.7: Scree plot illustrating reduction in RMSEV as a function of predictor number. The *F*-test and $PICP_{1-\alpha}$ models are chosen at 35 and 17 predictors, respectively.

Perhaps this example, once again, demonstrates the true efficacy of the $PICP_{1-\alpha}$ approach: the regression best covering validation samples corresponds to the 17-predictor model. Integrating $PICP_{1-\alpha}$ measures into model selection allows a nearly exact reproduction of Galvão and coworkers' [37] results in terms of both prediction error and model size. Similar to the results on Tables I-III, standard error measures now exhibit a greater degree of balance than the 35-predictor model (*e.g.*, RMSECV/RMSEV = 1.6, RMSEP/RMSEV = 0.7). Overall, the use of $PICP_{1-\alpha}$ statistics might mitigate the influence of methodological differences on the selection of stable PLS models and therefore improve the reproducibility of an experiment.

Contrary to least-squares regressions, PLS regression coefficients ($\widehat{\boldsymbol{b}}_j$) are not directly used in the determination of an analytical quantity but calculated as a matter of convenience or utility [2]. They are actually *dependent* quantities having been developed from the loading weights and X-scores used during the NIPALS procedure. Figure 6.8 illustrates regression

coefficient dependence irrespective of LV number, *i.e.*, adjacent coefficients are generally similar in magnitude and interval width.

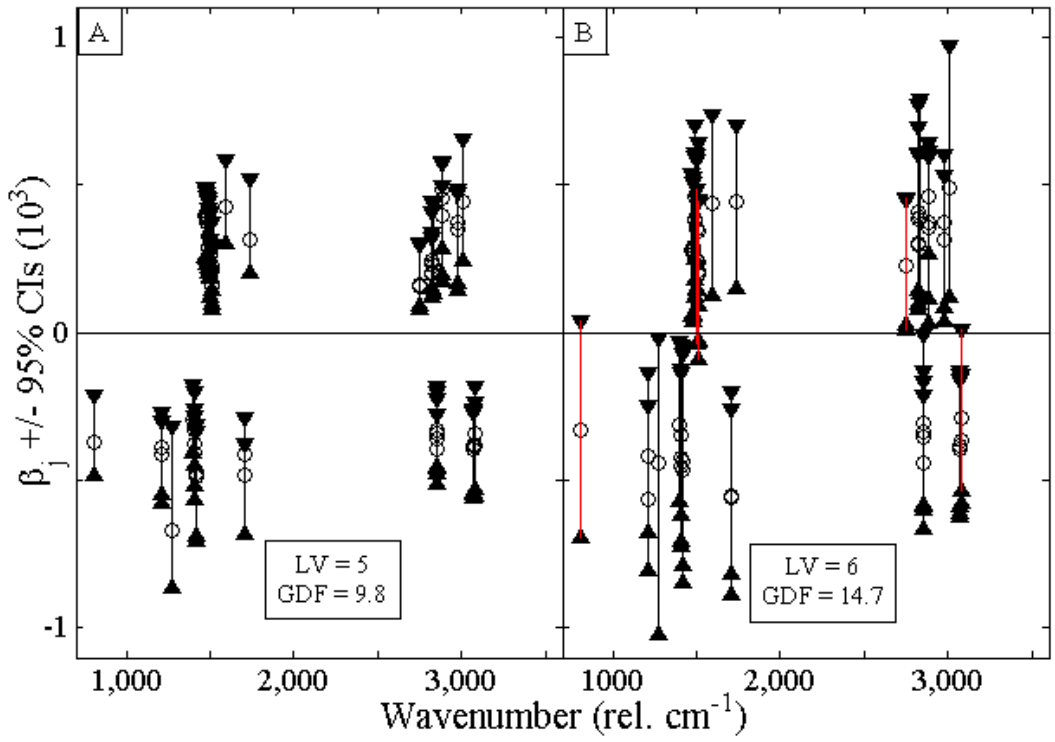


Figure 6.8: Juxtaposed regression coefficient (○) and 95% confidence interval estimates (▲,|,▼) for the 5 and 6 LV, 50-predictor TPU hard segment regression (A, B). An incorrect estimate of PLS components expands the width of each confidence interval (B) consequently leading to zero-crossings (red).

It is well-known that a correct estimate of LV number is critical to a PLS regression problem. Comparing Figures 6.8A and 6.8B shows that this translates directly into regression coefficient stability. Specifically, confidence intervals developed from the 5-LV model span exclusively nonzero values suggesting neither the presence of specious nor noisy parameters. The 6-LV model exhibits substantially wider intervals, consuming approximately five additional degrees of freedom (GDF = 14.7), consequently indicating that zero is a probable value for some coefficients (*e.g.*, 810.2 rel. cm^{-1}). Phrased differently, the higher cost of regression translates directly into a greater sensitivity of regression coefficients to random perturbations [33].

Conclusion

This study concerned selecting spectroscopic features germane for single-analyte PLS regressions using a revised Monte Carlo unimportant variable elimination routine. The performance of two variants of BMCUVE proved comparable or superior to preexisting methods of predictor selection (SPA and MWPLS) in terms of both precision and three prediction interval-based measures (*i.e.*, $PICP_{1-\propto}$). Independent of spectroscopic probe, stable and high-precision regressions always used features justifiable in terms of a target analyte's normal vibrations (*e.g.*, Figures 6.2 and 6.6). As hypothesized, reassessing model complexity (number of latent variables) at each iteration of BMCUVE appeared prudent, *i.e.*, a reduction in the number of PLS components confirmed that artifacts and interferences were major source-contaminants of a PLS model's latent structure. Ultimately, the subsequent removal of such features led to more stable (Figure 6.8) and less costly regressions.

Expanding the scope and diversity of model selection criteria was critical in selecting stable and accurate predictive models. Assigning top priority to $PICP_{1-\propto}$ measures helped facilitate the identification of regressions that exhibited balance across distinct standard error measures (RMSEC ≅ RMSEV). This signified that a chosen model will predict an unknown, future sample within a theoretical interval governed by $\hat{\sigma}_l$ (Equation 6.6). $PICP_{1-\propto}$ criteria provide no information about the magnitude of prediction uncertainty ($\hat{\sigma}_l$). This required the use of the RMSEV, the BIC, and Q^2 to necessarily identify a best linear model.

Comprehensive model selection criteria appear to aid in reproducing a predictive modeling experiment in terms of RMSEP. Specifically, Galveo and coworker's SPA study was reproduced (within error) in spite of using a different sample set partitioning method prior to feature selection (Table 6.2; Figure 6.6) [37]. Furthermore, the SPA uses an ordinary least-squares estimator. This indicates that the proposed criteria can be easily generalized to modeling problems unrelated to PLS.

Selecting a minimal number of spectroscopic features for a regression might appear desirable in itself. Improving precision and interpretability are well-known benefits of feature selection as long as removing artifacts and background elucidates the estimation of a PLS model's true latent structure. In other words, parsimony in PLS regression should only concern the estimated cost of the model (*e.g.*, GDF) insofar as unwanted, systematic variation that predominated $[X]_c$ was consequently removed with the discarded predictors

(Table 6.4; Figure 6.8). Accuracy and parametric stability takes precedence over reducing the size of the design matrix in absolute terms.

Overall, these seven comprehensive criteria substantially aid the selection of the most stable, precise, and accurate PLS regressions given a range of viable candidates provide by BMCUVE. Furthermore, austere commitment to these criteria tended to favor models free from spectroscopic artifacts, superfluous baseline, and excessive redundancy.

References

[1] H. Martens, M. Høy, F. Westad, D. Folkenberg, M. Martens, Analysis of designed experiments by stabilised PLS Regression and jack-knifing, Chemom. Intell. Lab. Syst. 58 (2001) 151-170.

[2] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109-130.

[3] T. Næs, T. Isaksson, T. Fearn, T. Davies, An User-friendly Guide to Multivariate Calibration and Classification, Chichester, West Sussex, NIR Publications, 2002.

[4] T. Mehmood, H. Martens, S. Saebo, J. Warringer, L. Snipen, A Partial Least Squares based algorithm for parsimonious variable selection, Algorithms Mol. Biol. 6 (2011) 27.

[5] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer, 2008.

[6] J. Fernandez Pierna, O. Abbas, V. Baeten, P. Dardenne, A Backward Variable Selection method for PLS regression (BVSPLS), Anal. Chim. Acta 642 (2009) 89-93.

[7] J.H. Kalivas, P. Gemperline, Calibration, in: P. Gemperline (Eds), Practical Guide to Chemometrics, CRC/Taylor & Francis, Boca Raton, 2006, pp. 144-147

[8] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J. Chemom. 16 (2002) 119-128.

[9] S. Wold, H. Antti, F. Lindgren, J. Öhman, Orthogonal signal correction of near-infrared spectra, Chemom. Intell. Lab. Syst. 44 (1998) 175-185.

[10] R. M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta 692 (2011) 63-72.

[11] T. Mehmood, K. H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62-69.

[12] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (2001) 23-38.

[13] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc., Series B Stat Methodol 72 (2010) 3-25.

[14] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14-32.

[15] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507-2517.

[16] W. Cai, Y. Li, X. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188-194.

[17] R. Leardi, Application of genetic algorithm–PLS for feature selection in spectral data sets, J. Chemom. 14 (2000) 643-655.

[18] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486-497.

[19] K. P. Burnham, D. R. Anderson, Multimodel Inference: Understanding AIC and BIC in Model Selection, Socio. Meth. Res. 33 (2004) 261-304.

[20] N. M. Faber, R. Rajkó, How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative, Anal. Chim. Acta 595 (2007) 98-106.

[21] L. Zhang, Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): A practitioner's perspective, Chemom. Intell. Lab. Syst. 97 (2009) 152-158.

[22] K. Zhao, D. Valle, S. Popescu, X. Zhang, B. Mallick, Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection, Remote Sens. Environ. 132 (2013) 102-119.

[23] R. W. Kennard, L. A. Stone, Computer Aided Design of Experiments, Technometrics 11 (1969) 137-148.

[24] B. Li, J. Morris, E. B. Martin, Model selection for partial least squares regression, Chemom. Intell. Lab. Syst. 64 (2002) 79-89.

[25] J. Shao, Linear Model Selection by Cross-validation, J. Am. Statist. Assoc. 88 (1993) 486-494.

[26] S. Wold, Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models, Technometrics 20 (1978) 397-405.

[27] W. J. Krzanowski, Cross-validatory choice in principal component analysis; some sampling results, J. Stat. Comput. Sim. 18 (1983) 299-314.

[28] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, 1993.

[29] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemom. Intell. Lab. Syst. 56 (2001) 1-11.

[30] V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, C. Sterna, Elimination of Uninformative Variables for Multivariate Calibration, Anal. Chem. 68 (1996) 3851-3858.

[31] J. Devore, Probability and Statistics for Engineering and the Sciences, Cengage Learning, 2007.

[32] G. Schwarz, Estimating the Dimension of a Model, Ann. Statist. 6 (1978) 461-464.

[33] J. Ye, On Measuring and Correcting the Effects of Data Mining and Model Selection, J. Am. Statist. Assoc. 93 (1998) 120-131.

[34] N. Krämer, M. Sugiyama, The Degrees of Freedom of Partial Least Squares Regression, J. Am. Statist. Assoc. 106 (2011) 697-705.

[35] H. van der Voet, Pseudo-degrees of freedom for complex predictive models: the example of partial least squares, J. Chemom. 13 (1999) 195-208.

[36] X.M. Sun, X.P. Yu, Y. Liu, L. Xu, D.L. Di, Combining bootstrap and uninformative variable elimination: Chemometric identification of metabonomic biomarkers by nonparametric analysis of discriminant partial least squares, Chemom. Intell. Lab. Syst. 115 (2012) 37-43.

[37] R. K. H. Galvão, M. C. U. Araújo, W. D. Fragoso, E. C. Silva, G. E. José, S. F. C. Soares, *et al.*, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemom, Intell. Labs. Syst. 92 (2008) 83-91.

[38] A. T. Weakley, P. C. Warwick, T. E. Bitterwolf, D. E. Aston, Multivariate analysis of micro-Raman spectra of thermoplastic polyurethane blends using principal component analysis and principal component regression, *Appl. Spectrosc.* 66 (2012) 1269-78.

[39] J.H. Jiang, R. J. Berry, H. W. Siesler, Y. Ozaki, Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data, *Anal. Chem.* 74 (2002) 3555-3565.

[40] R. Bro, S. De Jong, A fast non-negativity-constrained least squares algorithm, *J. of Chemom.* 11 (1997) 393-401.

[41] C. Wilhelm, J.L. Gardette, Infrared analysis of the photochemical behaviour of segmented polyurethanes: 1. Aliphatic poly(ester-urethane), Polymer 38 (1997) 4019-4031.

[42] Mine Safety and Health Adminstration (2013) Infrared Determination of Quartz in Respirable Coal Mine Dust - Method No. MSHA P7. Pittsburgh Safety and Health Technology Center, Pittsburgh.

[43] D. M. Haaland, E. V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, *Anal. Chem.* 60 (1988) 1193-1202.

Chapter 7. Conclusions

Chemometrics provide powerful tools to interrogate high-dimensional data in a manner that facilitates the convergence of statistical and scientific paradigms. The preceding chapters were organized implicitly and rationally with this in mind; namely, to demonstrate that a clear and sequential improvement in the experimental design, measurement, preprocessing, interpretation, and statistical validation of multivariate predictive models is both achievable and aids in elucidating scientifically relevant relationships contained on spectroscopic signatures. A scientific investigation is always limited by the innate accuracy and precision of a measuring instrument and probe as well as the skill exercised to produce an analytical measurement. While chemometrics will never resolve such fundamental limitations, novel preprocessing algorithms for baseline and signal correction (elucidated in chapters 2 and 3), latent variable models (chapter 4 and 5), and advanced validation protocols (chapter 6) significantly aid the straightforward evaluation of real complex chemical systems.

Chapter 4 illustrated that the purely mathematical (*i.e.*, unsupervised) PCA could aid in exploring the specific hydrogen bonding behavior of TPU blends. Major sources of non-essential variation in $[X]$ were filtered by the PCA thereby clarifying the estimation and interpretation of the principal components in terms of physically meaningful factors. Visualizing between-sample behavior on PC score and reconstruction plots justified the real deviations in spectroscopic behavior between each blend (see Figure 4.3-4.7) in terms of both hydrogen bonding interactions and viscoelastic behavior. A PCR successfully predicted the bulk hard segment content in each TPU blend using independent validation samples.

Chapter 5 applied the experience gleaned from PCA and PCR to inform the direction of a small-sample, PLS regression tasked with determining the quantity of airborne silica on polymeric sampling-filters using FT-IR spectra. Major improvements over current assessment methods of airborne silica included: an estimation of absorbed silica *within* the original sampling-filter, greater prediction accuracy and precision than an ordinary least-squares alternative, and minimal user intervention in the estimation procedure was demonstrated. The coincidence of statistical and scientific interference was again confirmed when the BMCUVE feature selection routine was used to identify absorbance features in the FT-IR spectra germane to accurately predicting silica. BMCUVE further suppressed or

outright eliminated the influence of substrate sampling filter and extraneous mineral species absorbance within the IR spectra. Notably, the precision improvements observed in select PLS regressions were directly interpretable in terms of the TO, LO, and lattice normal molecular vibrations of α-quartz (see Figures 5.3, 5.4) indicating *a fortiori* that field portable methods are entirely feasible at non-coal mining operations.

Chapter 6 collated the lessons learned in the preceding chapters to develop a more theoretically-orientated investigation into how the process of PLS model selection is improved by prioritizing metrics of long-run stability ($PICP_{1-\alpha}$), validation sample-statistics, and information criteria. Applying seven comprehensive criteria to the small (silica FT-IR), medium (corn NIR), and large sample (TPU Raman) calibrations proved essential in identifying predictive models exhibiting high precision, balance across measures of standard error (e.g., RMSEV), and requiring fewer PLS components. Again, the most statistically accurate and stable PLS models were justifiable in terms of analyte selective absorbance or scattering.

Any multivariate calibration requires extensive and repeated evaluation of a proposed model until scientific legitimacy substantiates statistical likelihood. Even an ideal mixture of spectroscopically-resolved chemical responses will often require repeating an entire preprocessing, variable scaling, calibration, and validation procedure until a final model can be justified appropriately. Applying ever more advanced chemometric analyses cannot inherently circumvent, prevent, or deny the incidence of spurious inferences; if anything, the inappropriate application of chemometrics only multiplies the likelihood of mistaking chance correlation for real scientific causality.

Practical domains outlined in the introduction (1-4) were articulated and arranged to protect against an overly optimistic interpretation of latent variable models. The novel contributions to applied chemometrics presented thereafter were tailored specifically to these domains where numerical algorithms facilitated and streamlined the application of chemometrics to a diverse array of scientific and engineering problems. Given the ubiquity and diminishing cost of high-throughput instrumentation, a practical and theoretical engagement with chemometrics and numerical analysis will continue to lead to substantial improvements in the accuracy and generalizability of baseline correction algorithms

(Chapters 2 and 3), further substantiate the use of PCA and multivariate calibrations to probe complex chemical matrices (Chapter 4), significantly extend accurate and rapid multivariate calibration approaches to occupational safety applications (Chapter 5), and further develop comprehensive performance criteria aiming to inform the choice of only the most stable predictive models (Chapter 6).

Appendix A

Copyright letters from *Applied Spectroscopy* and *Analytical and Bioanalytical Chemistry*

Subject: SAS - Manuscripts for PhD dissertation
From: aweakley@vandals.uidaho.edu
Date: 4/10/2014 10:05 PM
To: exdir@s-a-s.org

You have received a message via the SAS Website (http://s-a-s.org/)

From: aweakley@vandals.uidaho.edu

Subject: SAS - Manuscripts for PhD dissertation

Ms. Saylor:

My name is Andrew Weakley and I have three papers authored in Applied Spectroscopy. I am also about 1 month away from completing my dissertation and would like to have permission to use these manuscripts as the first major part (3-4 chapters) of my dissertation. Would that be possible? If so, I would like to have some official copyright release form to include as an appendix in my dissertation. The published works were:

1. Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, "Automatic baseline correction of vibrational circular dichroism spectra." Appl. Spectrosc. 2013, 67(10): 1117-1126.

2. Andrew T. Weakley, P.C. Temple Warwick, Thomas E. Bitterwulf, d. Eric Aston, "Multivariate analysis of micro-Raman spectra of thermoplastic polyurethane blends using principal component analysis and principal component regression." Appl. Spectrosc. 2012, 66(11): 1269-1278.

3. Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, "Automatic baseline subtraction of vibrational spectra using minima identification and discrimination via adaptive, least-squares thresholding." Appl. Spectrosc. 2012, 66(5): 519-529.

Thank you for your time,

Permission granted for the use requested.
Granted this 11th day of April 2014.
Full citation required.
Bonnie Saylor, Executive Director, SAS

Andy

Confirmation of your Copyright Transfer

Dear Author,

Please note: This e-mail is a confirmation of your copyright transfer and was sent to you only for your own records.

The copyright to this article, including any graphic elements therein (e.g. illustrations, charts, moving images), is hereby assigned for good and valuable consideration to Springer-Verlag Berlin Heidelberg effective if and when the article is accepted for publication and to the extent assignable if assignability is restricted for by applicable law or regulations (e.g. for U.S. government or crown employees). Author warrants (i) that he/she is the sole owner or has been authorized by any additional copyright owner to assign the right, (ii) that the article does not infringe any third party rights and no license from or payments to a third party is required to publish the article and (iii) that the article has not been previously published or licensed.

The copyright assignment includes without limitation the exclusive, assignable and sublicensable right, unlimited in time and territory, to reproduce, publish, distribute, transmit, make available and store the article, including abstracts thereof, in all forms of media of expression now known or developed in the future, including pre- and reprints, translations, photographic reproductions and microform. Springer may use the article in whole or in part in electronic form, such as use in databases or data networks for display, print or download to stationary or portable devices. This includes interactive and multimedia use and the right to alter the article to the extent necessary for such use.

Authors may self-archive the Author's accepted manuscript of their articles on their own websites. Authors may also deposit this version of the article in any repository, provided it is only made publicly available 12 months after official publication or later. He/she may not use the publisher's version (the final article), which is posted on SpringerLink and other Springer websites, for the purpose of self-archiving or deposit. Furthermore, the Author may only post his/her version provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final publication is available at Springer viahttp://dx.doi.org/[insert DOI]".

Prior versions of the article published on non-commercial pre-print servers like arXiv.org can

Article Details

Journal title

Analytical and Bioanalytical Chemistry

Article title

Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least-squares regression

DOI

10.1007/s00216-014-7856-y

Corresponding Author

Andrew Weakley

Copyright transferred to

Springer-Verlag Berlin Heidelberg

Transferred on

Fri Apr 25 21:43:54 CEST 2014

Appendix B

Supporting Information for Automatic baseline subtraction of surfaced enhanced infrared absorbance (SEIRA) and Raman spectra using minima identification and discrimination via adaptive, least-squares thresholding

Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, Applied Spectroscopy, 2012, **66**(5): 519-529.

Abstract

Supporting information includes the results of baseline correction for all spectra not directly presented in the manuscript, the selected results from the repeated sampling studies, and the Matlab code for the vibrational spectrum simulator. All real and simulated spectra were baseline corrected using the low-high orientation unless otherwise stated.

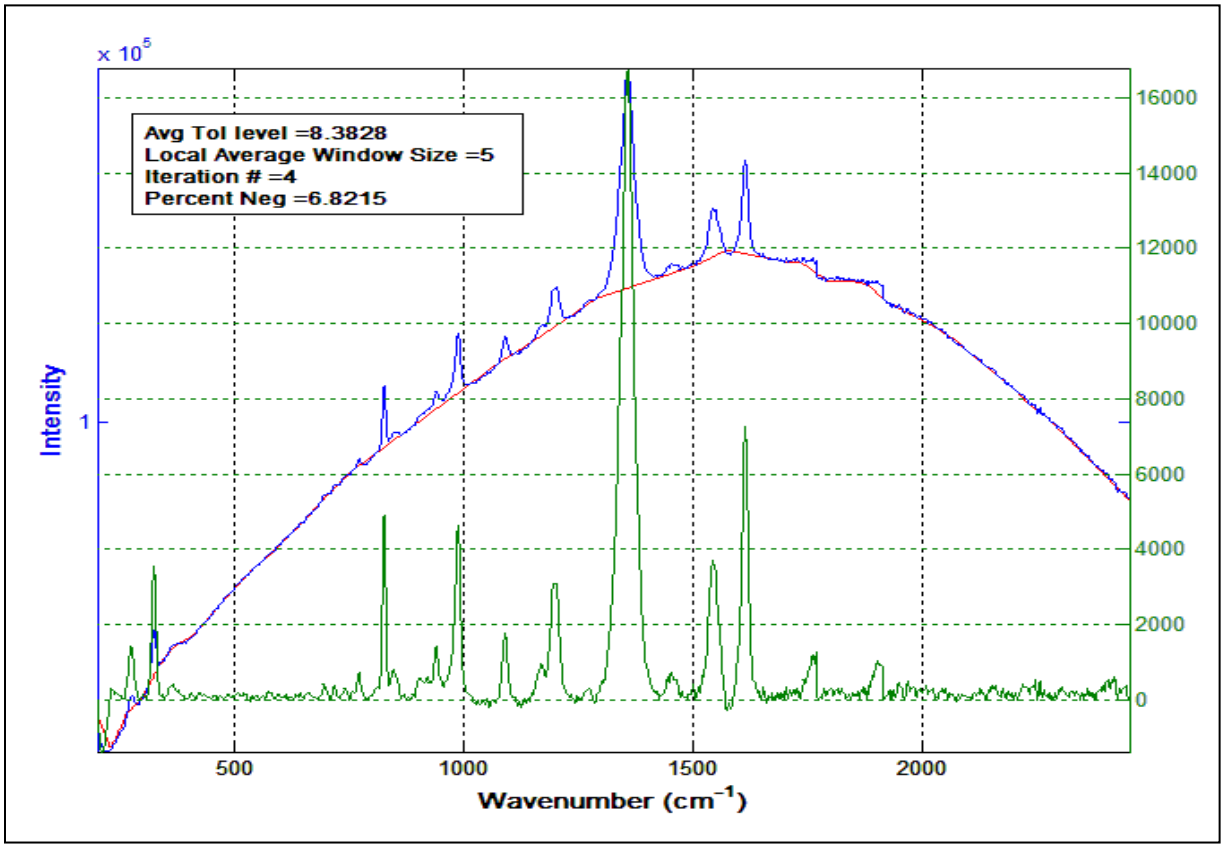


Figure B.1: Raman spectrum of HNBB acquired using a 785nm source. The first estimated baseline and original spectrum are plotted on the left hand axis with the corrected spectrum on the right hand axis. Spectral artifacts located at ~1750 and 1850 cm^{-1} were not completely removed by the algorithm. Only supplementary methods such as smoothing can eliminate these spectral artifacts.

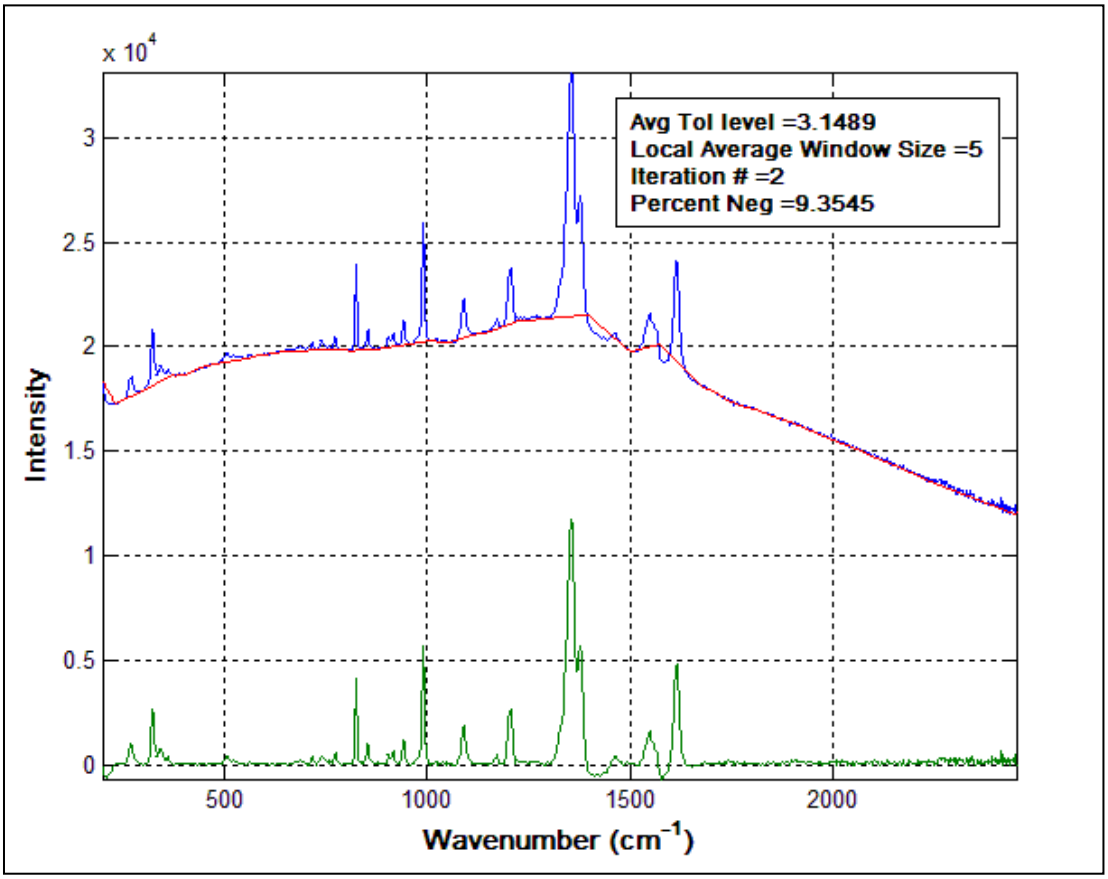


Figure B.2: Raman spectrum of original, estimate baseline, and corrected spectrum of HNBB. Two regions were incorrectly subtracted at ~1400 and ~1600 cm^{-1}. Although not shown, baseline correction was nearly perfect for an n_{tol} value of 3.3.

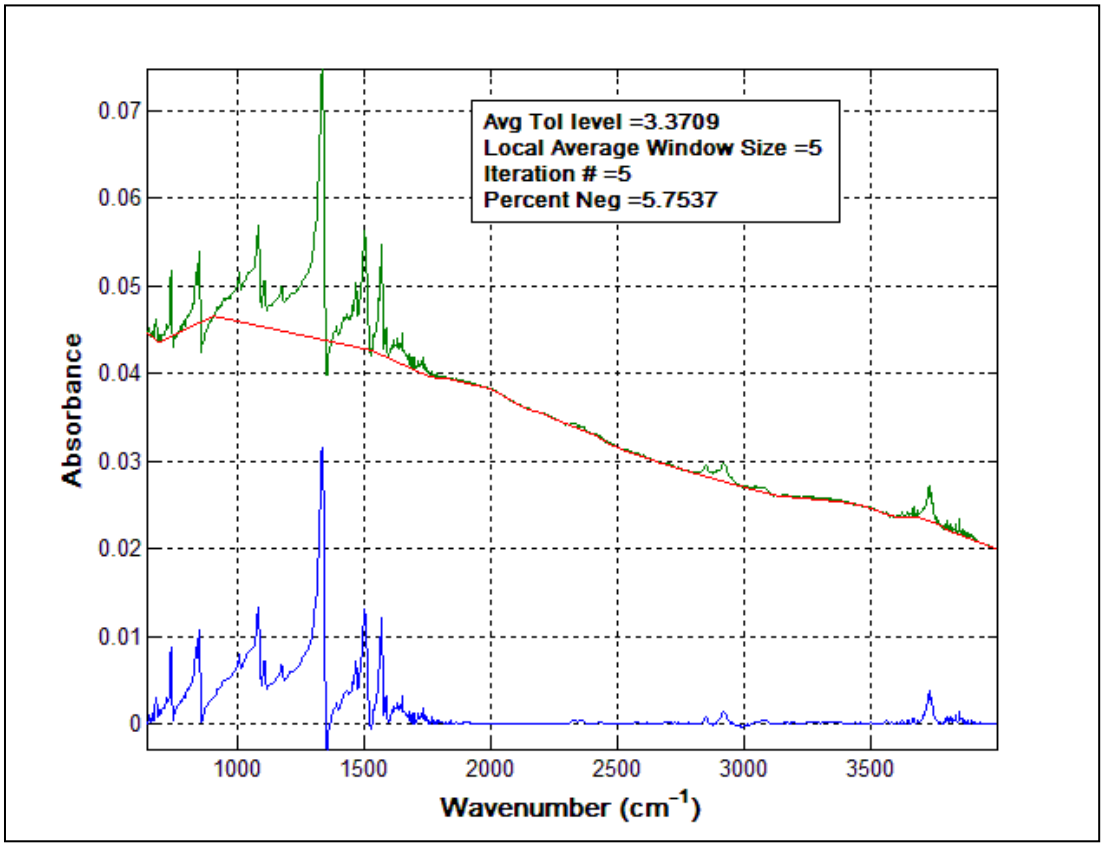


Figure B.3: SEIRA spectrum of the original, estimated baseline, and corrected spectrum of PNTP. A vertical offset of 0.6AU was subtracted from the original and estimated baseline to allow all objects to be plotted together.

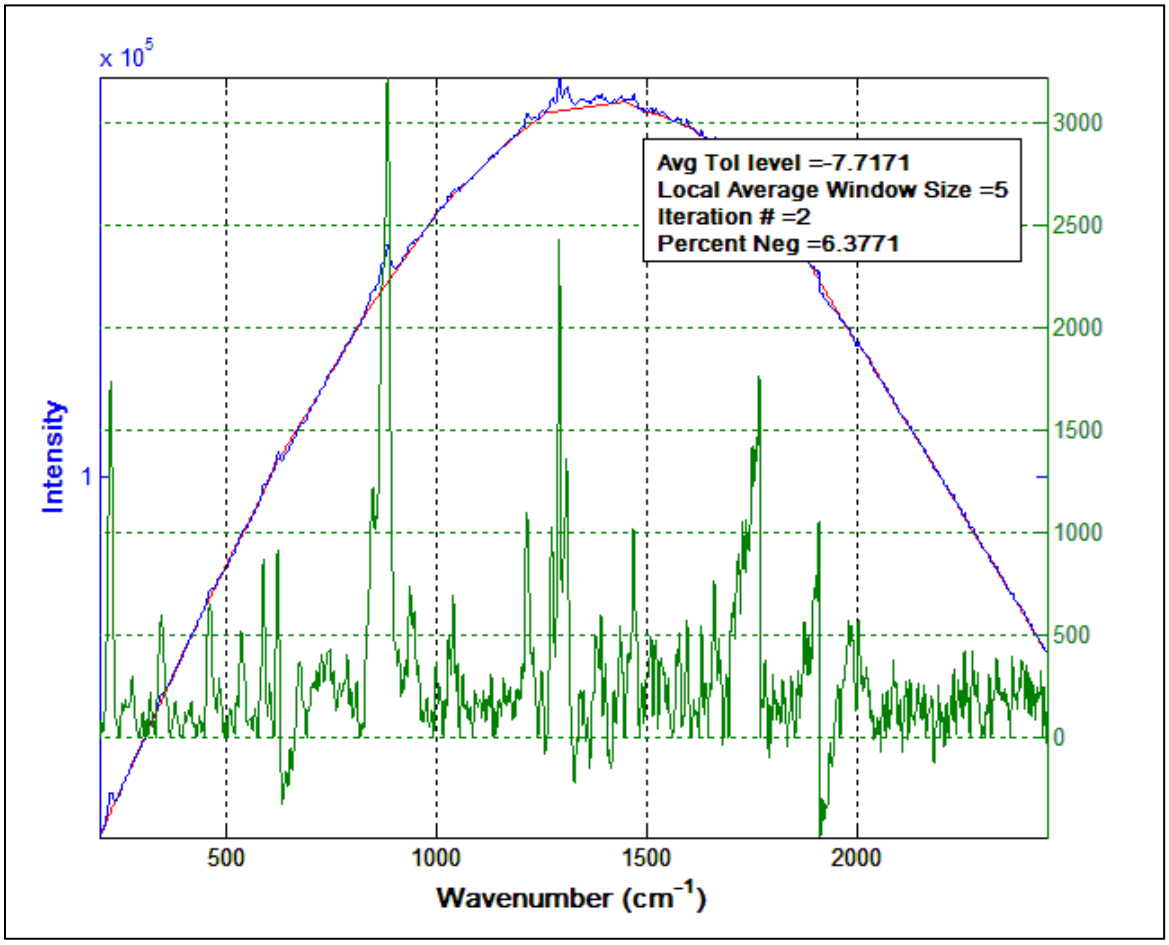


Figure B.4: Raman spectrum of the original, estimate baseline, and corrected spectrum of SEMTEX acquired with a 785nm source. Although not easy to discern from the original spectrum the peaks located at ~1750 and ~1850 cm^{-1} are spectral artifacts that were included in the final corrected spectrum.

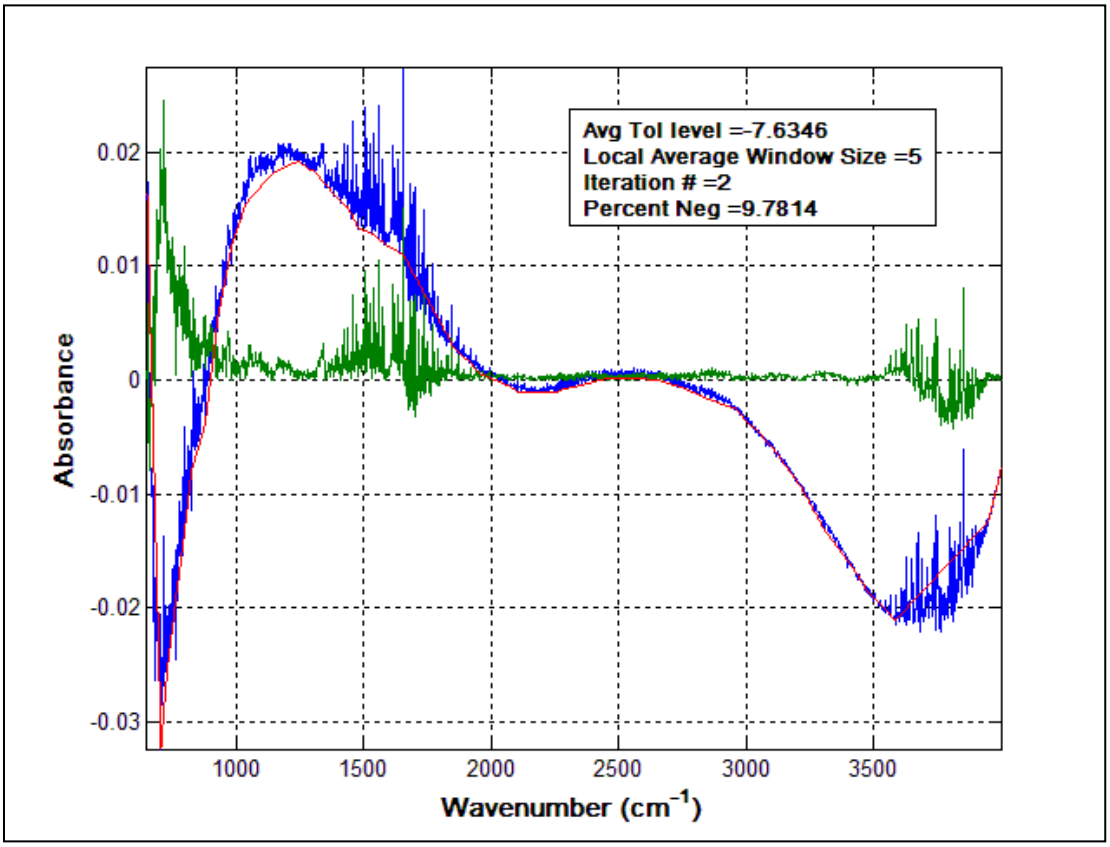


Figure B.5: SEIRA spectrum of the original, estimated baseline, and corrected spectrum of PNTP. Along with the complex background, positively streaking bands from 1400-2000 cm^{-1} and 3500-3900 cm^{-1} indicate the presence of water evaporation in the sample cell. Baseline correction proceeded in the high-low configuration.

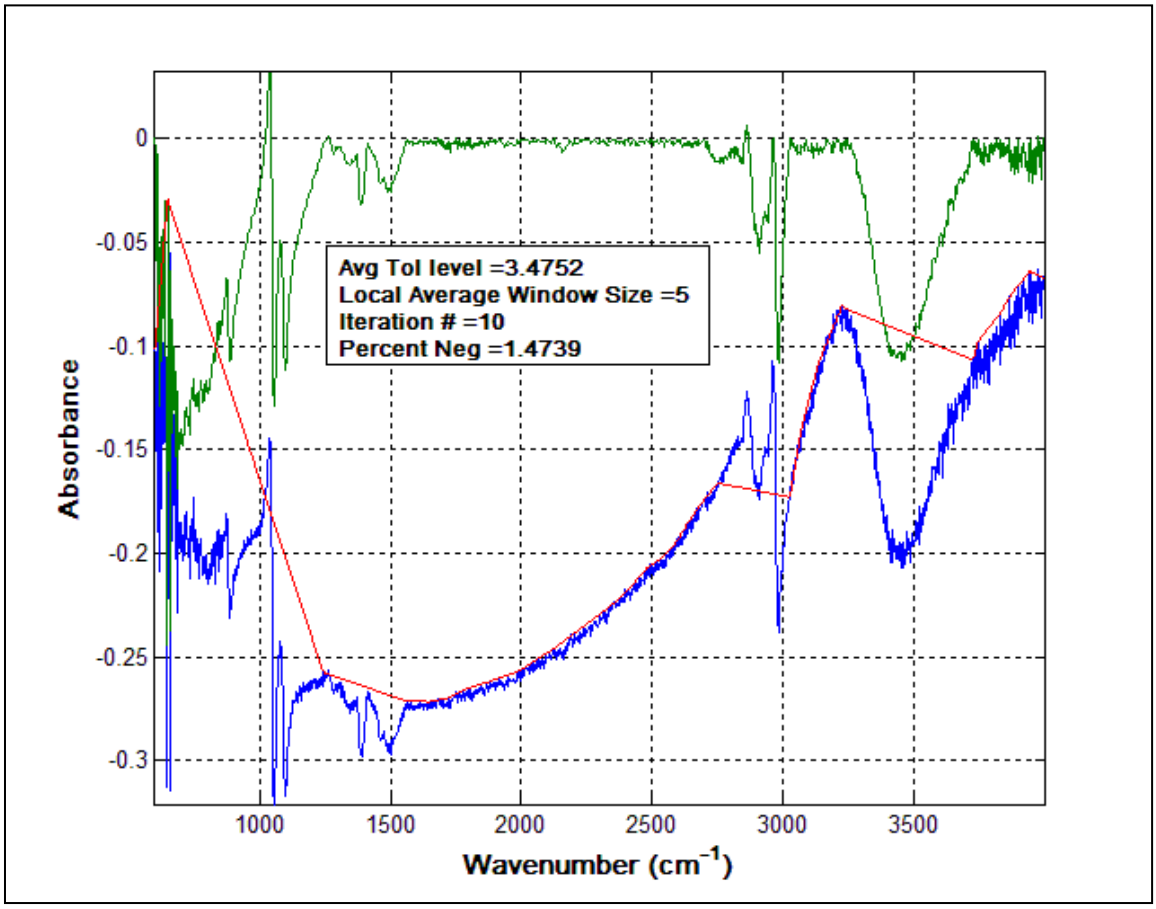


Figure B.6: SEIRA spectrum of the original, estimated baseline, and corrected spectrum of ethanol. The endpoints of the broad OH stretching located at 3400 cm^{-1} were identified prior to executing the algorithm. Baseline correction was performed in the high-low configuration.

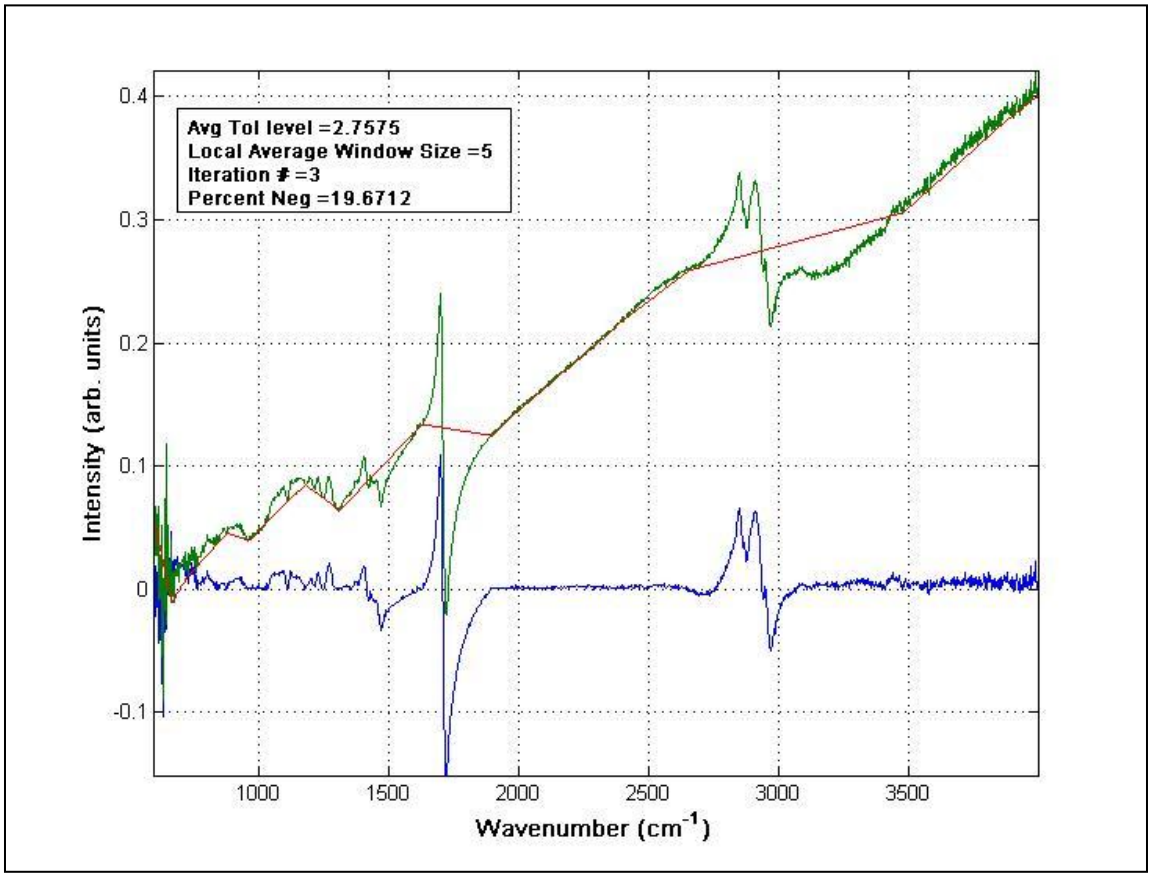


Figure B.7: SEIRA spectrum of heptanoic acid adsorbed on percolated silver nanoparticles.

Table B.1- Selected results from forty-eight repeated resampling scenarios using the fixed-window (n=5) algorithm. Note, this table has been expanded to include the number of bands (e.g., B_{raman}) used in each scenario. All statistics were calculated using a total of 500 spectra generated and baselines corrected per scenario.

Name	B_{raman}	B_{derv}	σ_{RMS}	n_{tolstp}	Raman	SIERA	Adapt	SNR_{AVG}	$RMSE_{AVG}$	$RMSE_{MED}$	t_{run}(s)
$FW_{500}^{(10,0)}$	10	0	1	1	1	0	1	289.46	13.83	12.13	238
$FW_{500}^{(10,0)}$	10	0	1	1	1	0	0	294.38	12.50	9.34	235
$FW_{500}^{(10,0)}$	10	0	10	0	0	0	1	30.85	25.21	23.56	262
$FW_{500}^{(10,0)}$	10	0	10	0	0	0	0	30.03	24.01	21.80	258
$FW_{500}^{(10,0)}$	10	0	1	1	0	0	1	292.52	16.14	14.01	260
$FW_{500}^{(10,0)}$	10	0	1	1	0	0	0	293.82	15.55	11.27	259
$FW_{500}^{(10,0)}$	10	0	10	1	0	0	1	29.92	27.17	26.25	252
$FW_{500}^{(10,0)}$	10	0	10	1	0	0	0	30.06	25.54	23.93	260

$FW_{500}^{(10,0)}$	10	0	1	0	1	0	1	290.39	14.77	12.14	265
$FW_{500}^{(10,0)}$	10	0	1	0	1	0	0	291.30	11.93	9.34	258
$FW_{500}^{(10,0)}$	10	0	10	0	1	0	1	30.16	24.76	23.74	270
$FW_{500}^{(10,0)}$	10	0	10	0	1	0	0	30.45	22.51	20.90	262
$FW_{500}^{(10,0)}$	10	0	1	0	0	0	1	293.41	16.62	13.74	262
$FW_{500}^{(10,0)}$	10	0	1	0	0	0	0	290.16	14.94	11.26	258
$FW_{500}^{(10,0)}$	10	0	10	0	0	0	1	30.20	27.05	25.69	263
$FW_{500}^{(10,0)}$	10	0	10	0	0	0	0	29.89	24.75	22.83	250
$FW_{500}^{(5,5)}$	5	5	1	1	1	0	1	214.40	21.99	20.06	264
$FW_{500}^{(5,5)}$	5	5	1	1	1	0	0	215.92	30.69	27.86	255
$FW_{500}^{(5,5)}$	5	5	10	1	1	0	1	22.77	23.21	21.83	271
$FW_{500}^{(5,5)}$	5	5	10	1	1	0	0	22.73	31.27	29.07	253
$FW_{500}^{(5,5)}$	5	5	1	1	0	0	1	214.89	14.94	13.48	258
$FW_{500}^{(5,5)}$	5	5	1	1	0	0	0	215.34	14.06	11.43	246
$FW_{500}^{(5,5)}$	5	5	10	1	0	0	1	22.34	21.22	20.45	276
$FW_{500}^{(5,5)}$	5	5	10	1	0	0	0	22.54	20.54	19.41	255
$FW_{500}^{(5,5)}$	5	5	1	0	1	0	1	221.77	21.98	20.29	277
$FW_{500}^{(5,5)}$	5	5	1	0	1	0	0	219.13	26.81	24.19	259
$FW_{500}^{(5,5)}$	5	5	10	0	1	0	1	22.96	23.63	22.04	272
$FW_{500}^{(5,5)}$	5	5	10	0	1	0	0	22.62	26.96	24.78	261
$FW_{500}^{(5,5)}$	5	5	1	0	0	0	1	218.80	15.02	14.14	262
$FW_{500}^{(5,5)}$	5	5	1	0	0	0	0	215.80	13.23	11.76	256
$FW_{500}^{(5,5)}$	5	5	10	0	0	0	1	23.15	21.03	20.60	263
$FW_{500}^{(5,5)}$	5	5	10	0	0	0	0	22.98	20.69	19.19	242
$FW_{500}^{(0,10)}$	0	10	1	1	1	1	1	95.10	15.87	13.28	38
$FW_{500}^{(0,10)}$	0	10	1	1	1	1	0	99.08	14.36	11.50	40
$FW_{500}^{(0,10)}$	0	10	10	1	1	1	1	11.33	20.75	20.21	33
$FW_{500}^{(0,10)}$	0	10	10	1	1	1	0	11.27	19.83	19.30	35
$FW_{500}^{(0,10)}$	0	10	1	1	0	1	1	98.60	14.82	12.79	28
$FW_{500}^{(0,10)}$	0	10	1	1	0	1	0	97.23	15.44	10.94	25
$FW_{500}^{(0,10)}$	0	10	10	1	0	1	1	11.53	21.04	20.54	27
$FW_{500}^{(0,10)}$	0	10	10	1	0	1	0	11.54	21.16	20.52	34

$FW_{500}^{(0,10)}$	0	10	1	0	1	1	1	99.40	14.63	12.63	31
$FW_{500}^{(0,10)}$	0	10	1	0	1	1	0	98.64	14.20	11.77	34
$FW_{500}^{(0,10)}$	0	10	10	0	1	1	1	11.32	20.24	20.25	36
$FW_{500}^{(0,10)}$	0	10	10	0	1	1	0	11.43	20.63	19.72	32
$FW_{500}^{(0,10)}$	0	10	1	0	0	1	1	98.05	14.85	12.58	32
$FW_{500}^{(0,10)}$	0	10	1	0	0	1	0	98.78	14.65	11.59	31
$FW_{500}^{(0,10)}$	0	10	10	0	0	1	1	11.31	20.19	20.23	28
$FW_{500}^{(0,10)}$	0	10	10	0	0	1	0	11.59	20.63	20.73	28

Table B.2- Selected results from forty-eight repeated resampling scenarios using the stepped-window algorithm. Note, this table has been expanded to include the number of bands (e.g., B_{raman}) used in each scenario. All statistics were calculated using a total of 500 spectra generated and baselines corrected per scenario.

Name	B_{raman}	B_{derv}	σ_{RMS}	n_{tolstp}	Raman	SIERA	Adapt	SNR_{AVG}	$RMSE_{AVG}$	$RMSE_{MED}$	t_{run} (s)
$SW_{500}^{(10,0)}$	10	0	1	1	1	0	1	295.06	15.21	13.13	701.00
$SW_{500}^{(10,0)}$	10	0	1	1	1	0	0	291.30	12.24	9.04	828.00
$SW_{500}^{(10,0)}$	10	0	10	1	1	0	1	30.32	25.04	24.34	753.00
$SW_{500}^{(10,0)}$	10	0	10	1	1	0	0	29.96	24.31	22.49	787.00
$SW_{500}^{(10,0)}$	10	0	1	1	0	0	1	289.22	20.05	17.16	737.00
$SW_{500}^{(10,0)}$	10	0	1	1	0	0	0	294.98	16.99	12.14	801.00
$SW_{500}^{(10,0)}$	10	0	10	1	0	0	1	29.88	28.32	26.63	804.00
$SW_{500}^{(10,0)}$	10	0	10	1	0	0	0	30.48	26.05	23.75	771.00
$SW_{500}^{(10,0)}$	10	0	1	0	1	0	1	291.01	15.29	13.37	763.00
$SW_{500}^{(10,0)}$	10	0	1	0	1	0	0	292.77	13.51	8.77	820.00
$SW_{500}^{(10,0)}$	10	0	10	0	1	0	1	30.36	25.56	24.50	778.00
$SW_{500}^{(10,0)}$	10	0	10	0	1	0	0	30.03	22.93	21.18	811.00
$SW_{500}^{(10,0)}$	10	0	1	0	0	0	1	293.73	18.81	16.79	767.00
$SW_{500}^{(10,0)}$	10	0	1	0	0	0	0	290.69	16.23	11.74	827.00
$SW_{500}^{(10,0)}$	10	0	10	0	0	0	1	30.48	27.67	26.56	779.00
$SW_{500}^{(10,0)}$	10	0	10	0	0	0	0	29.34	24.99	23.05	782.00
$SW_{500}^{(5,5)}$	5	5	1	1	1	0	1	216.90	23.05	20.57	782.00
$SW_{500}^{(5,5)}$	5	5	1	1	1	0	0	215.55	33.70	29.20	805.00
$SW_{500}^{(5,5)}$	5	5	10	1	1	0	1	22.55	24.17	22.81	809.00

$SW_{500}^{(5,5)}$	5	5	10	1	1	0	0	22.36	31.97	28.81	863.00
$SW_{500}^{(5,5)}$	5	5	1	1	0	0	1	212.00	15.28	13.87	809.00
$SW_{500}^{(5,5)}$	5	5	1	1	0	0	0	210.85	13.34	11.05	827.00
$SW_{500}^{(5,5)}$	5	5	10	1	0	0	1	22.76	21.82	20.34	797.00
$SW_{500}^{(5,5)}$	5	5	10	1	0	0	0	22.68	20.66	19.64	836.00
$SW_{500}^{(5,5)}$	5	5	1	0	1	0	1	216.91	22.52	19.81	804.00
$SW_{500}^{(5,5)}$	5	5	1	0	1	0	0	215.41	28.13	25.73	854.00
$SW_{500}^{(5,5)}$	5	5	10	0	1	0	1	22.60	23.75	22.66	802.00
$SW_{500}^{(5,5)}$	5	5	10	0	1	0	0	23.34	28.30	25.59	854.00
$SW_{500}^{(5,5)}$	5	5	1	0	0	0	1	219.98	14.59	13.35	823.00
$SW_{500}^{(5,5)}$	5	5	1	0	0	0	0	219.06	14.65	12.14	850.00
$SW_{500}^{(5,5)}$	5	5	10	0	0	0	1	22.51	20.97	19.69	821.00
$SW_{500}^{(5,5)}$	5	5	10	0	0	0	0	22.81	21.06	19.71	886.00
$SW_{500}^{(0,10)}$	0	10	1	1	1	1	1	97.81	27.82	25.73	94.00
$SW_{500}^{(0,10)}$	0	10	1	1	1	1	0	97.55	29.14	26.99	104.00
$SW_{500}^{(0,10)}$	0	10	10	1	1	1	1	11.22	27.47	26.41	95.00
$SW_{500}^{(0,10)}$	0	10	10	1	1	1	0	11.60	27.79	26.33	123.00
$SW_{500}^{(0,10)}$	0	10	1	1	0	1	1	99.04	18.04	14.88	94.00
$SW_{500}^{(0,10)}$	0	10	1	1	0	1	0	97.87	19.27	14.46	99.00
$SW_{500}^{(0,10)}$	0	10	10	1	0	1	1	11.42	23.70	21.89	99.00
$SW_{500}^{(0,10)}$	0	10	10	1	0	1	0	11.33	23.39	22.26	105.00
$SW_{500}^{(0,10)}$	0	10	1	0	1	1	1	96.78	29.21	26.17	99.00
$SW_{500}^{(0,10)}$	0	10	1	0	1	1	0	96.53	28.30	26.37	104.00
$SW_{500}^{(0,10)}$	0	10	10	0	1	1	1	11.27	27.09	25.86	90.00
$SW_{500}^{(0,10)}$	0	10	10	0	1	1	0	11.52	28.18	26.70	121.00
$SW_{500}^{(0,10)}$	0	10	1	0	0	1	1	98.53	17.85	15.40	82.00
$SW_{500}^{(0,10)}$	0	10	1	0	0	1	0	98.25	17.29	13.75	110.00
$SW_{500}^{(0,10)}$	0	10	10	0	0	1	1	11.25	23.04	21.68	96.00
$SW_{500}^{(0,10)}$	0	10	10	0	0	1	0	11.22	23.27	22.03	111.00

Matlab function for vibrational spectrum simulator

```
Function [Real_spec,nu,S,SNR]= raman_spectrum_sim(length, CS, ord, Fs, A,
B_num, B_num_der, std_noise, g_max, I_max, func1)

%{
PURPOSE:
This function generates a vibrational spectrum whose baseline
and band characteristics are random about user-specified, fixed values

INPUTS:
(1) length = of spectrum in data points
(2) CS = channel size
(3) ord = specifies number of sinusoids used to model the baseline
(4) Fs = the fixed frequency of the sinusoids
(5) A = fixed value of the sines
(6) B_num = number of Lorentzian bands
(7) B_num_der = number of derivative bands assuming a Lorentzian character
(8) std_noise = RMS noise
(9) g_max = two element vector specifying width of Raman and bipolar bands
(10) I_max = two element vector specifying amplitude of Raman/bipolar
bands
(11) func1 = Toggles sinusoid-only baseline

OUTPUTS:
(1) Real_spec = baseline + noise + pure bands (arb. units)
(2) nu = wavenumber vector (cm-1)
(3) S = pure bands + noise
(4) SNR = signal-to-noise ratio
%}
if isempty(length) ==1
    length=1200; %Number of points
end
if isempty(CS) ==1
    CS=2.5; %size of spectral channels
end
if isempty(ord)==1
    ord=5; %number of sinusoids used to generate the baseline
end
if isempty(Fs)==1
    Fs=2500; %controls 'sampling' rate (sampling rate of sinusoid
(cycle/nu))
end
if isempty(A)==1
    A=50; % fixed value of sine amplitude
end
if isempty(B_num) ==1
    B_num=10; %number of bands
end
if isempty(B_num_der)==1
    B_num_der=0; %number of bipolar (derivative bands)
end
if isempty(std_noise) ==1
    std_noise=1; %noise level in spectrum
```

```
end
if isempty(g_max)==1
    g_max=[40 200]; %maximum size of width parameter
elseif size(g_max) < 2
        fprintf('Define a size param as two element vector')
end
if isempty(I_max) ==1
    I_max=[350 3000]; %Max intensity of band
elseif size(I_max) < 2
    fprintf('Define a intensity param as two element vector')
end

%Step 1: Random baseline generation
pts=0:(length-1);
nu=CS*pts; %wavenumber axis


w1=randn(1,ord);
w2=randn(1,ord); %Weights for A
freq2=1/Fs; %true cycle/nu

for i=1:max(size(w1))
   freq11(i,:)=w1(i)*freq2*nu;
   A11(i,:)=w2(i)*A;
end
if func1== 0
    w3=randn(1,2); %Weights for function contribution to BL construction
    w3n=abs(w3)/sum(abs(w3),2);
elseif func1==1
    w3n(1)=1;
    w3n(2)=0;
end
BL1=w3n(1)*sum(A11,1)*sum(sin(2*pi*freq11),1)+w3n(2)*(1/2)*erfc(nu);

%Step 2: Add bands

if B_num > 0 || B_num_der > 0
    if B_num > 0
        for i=1:B_num
            t=abs(randn(100,1));
            t1=((t-min(t))/(max(t)-min(t)));
            [r,c,t1g]=find(t1 > 0.15 & t1 <0.9); %Ensures bands are not
too close to the ends of the spectrum
            xo=t1(r(end-1))*CS*size(nu,2); %Center of peak
            g=g_max(1)*t1(r(1)); %Width parameter
            I=I_max(1)*t1(r(end)); %Height of peak
            S1(i,:)=I*(1./(1+((nu-xo)/g).^2));
        end
    elseif B_num == 0
        i=1;
    end
    if B_num_der > 0
        for j=i:(i+B_num_der-1)
            t=abs(randn(100,1));
            t1=((t-min(t))/(max(t)-min(t)));
```

```
            [r,c,t1g]=find(t1 > 0.25 & t1 < 0.9); %Ensures bands are not
too close to the ends of the spectral window
            xo=t1(r(end-1))*CS*size(nu,2); %Center of peak
            g=g_max(2)*t1(r(1)); %Width parameter
            I=I_max(2)*t1(r(end)); %Height of peak
            S1(j,:)= (16*I*g^2)*(nu-xo)./(4*(nu-xo).^2+g^2).^2;
        end
    end

elseif B_num==1 == 0 && B_num_der == 0
S1=zeros(1,size(nu,2));
end

%Step 3: add noise

noise=std_noise*randn(1,size(nu,2)); %specifies noise

%Step 4: Sum all contributions

S=sum(S1,1)+noise+0.5*std_noise;
%Note, one half the noise level is added to the noisy, baseline-free
%spectrum to give a realistic RMSE value

Real_spec=S+BL1;
SNR=max(S)/std_noise;

end
```

Appendix C

Supporting Information for Automatic Baseline Correction of Vibrational Circular Dichroism Spectra

Andrew T. Weakley, Peter R. Griffiths, D. Eric Aston, Applied Spectroscopy, 2013, **67**(10): 1117-1126.

Abstract

Supporting information includes the results of baseline correction the S-Camphor spectra, flow chart of the Med($\mathbf{B_r}$) filtering procedure, and the 3-by-5 ANOVA table for the 32 segment baseline correction populations.

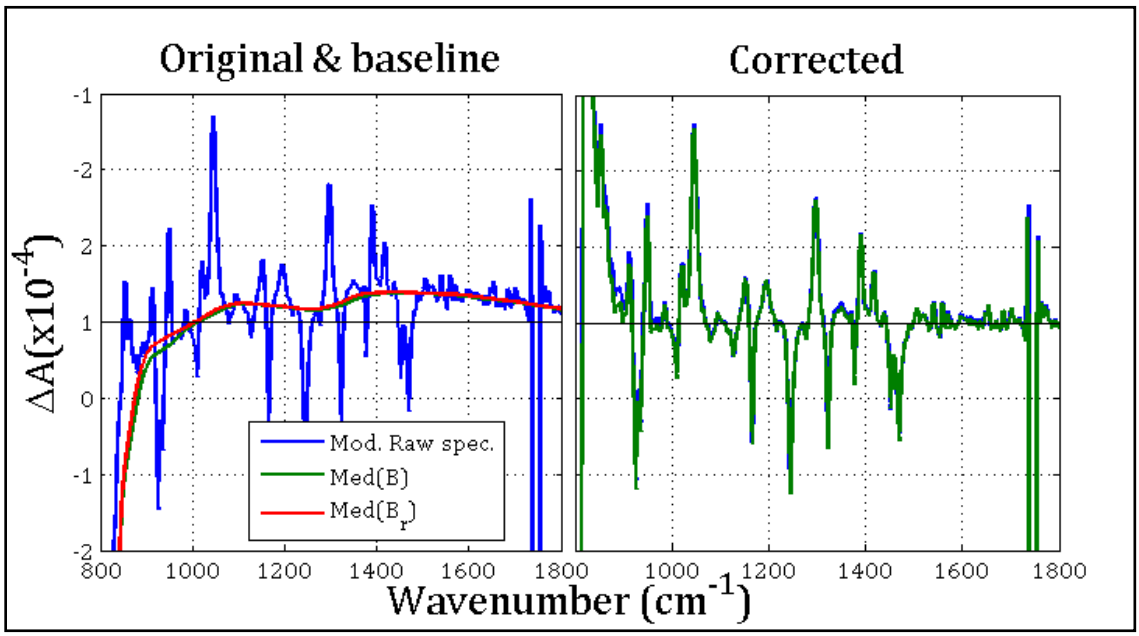


Figure C.8: "Better baseline" (S)-camphor VCD spectrum. Baseline correction used the default median filter and Med(B_r) filter.

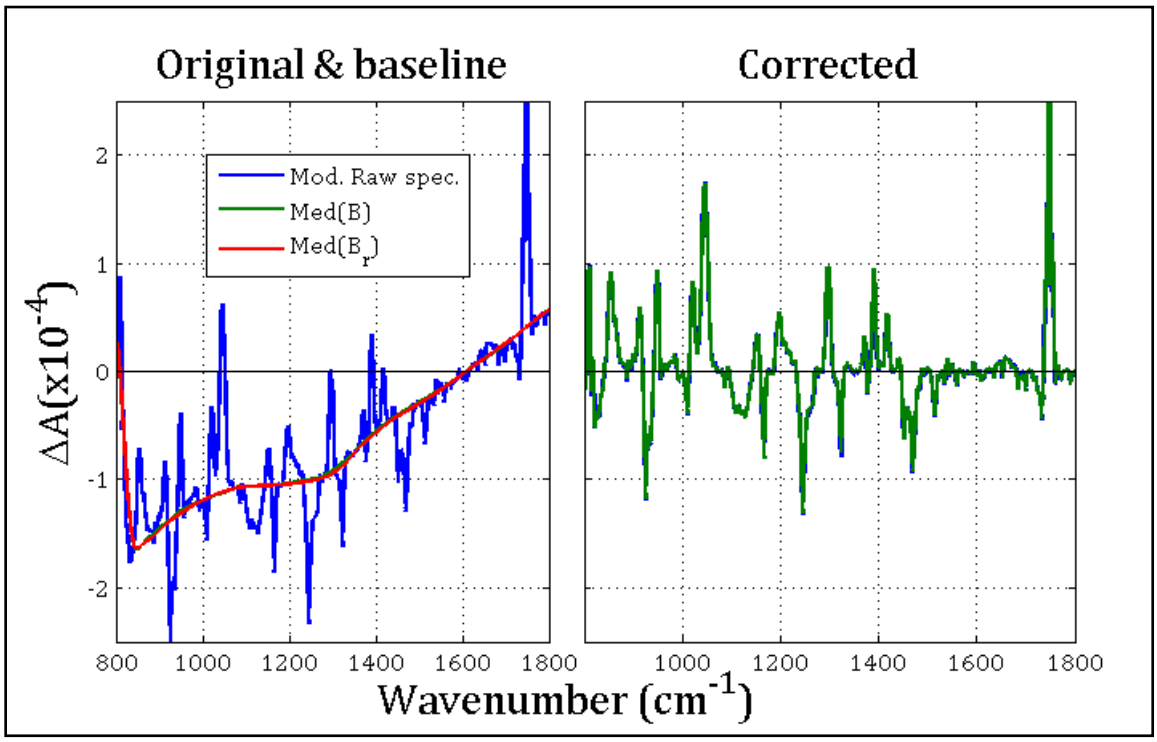


Figure C.9: "Worse baseline" (S)-camphor VCD spectrum. Baseline correction used the default median filter and Med(B_r) filter.

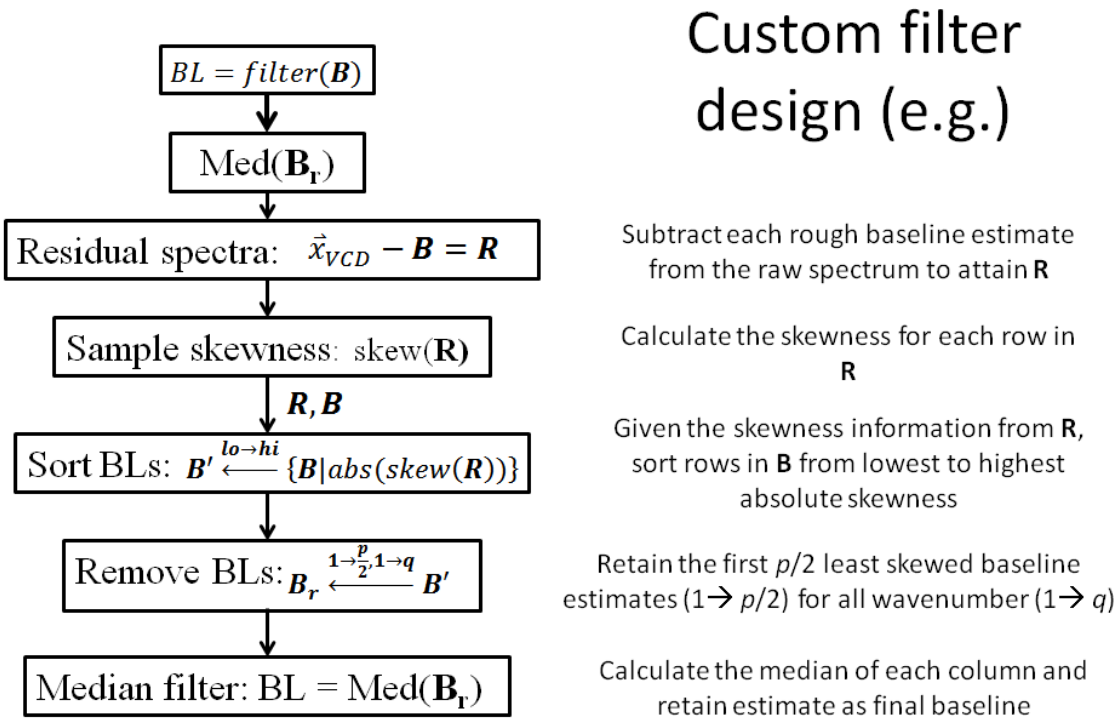


Figure C.10: Flow chart detailing the Med(B_r) filtering operation.

Table C.1: ANOVA table for the 32-segment population. Main effects and interactions are all significant at the p = 0.05 level.

Effects	df	Sum Sq.	Mean Sq.	F-value	Pr(>F)
BLC	2	14.1	7.04	24.801	0
Win-*n*	4	542.1	135.52	477.667	0
BLC:Win-*n*	8	4.5	0.56	1.968	0.0465*

Appendix D:

Supporting Information for Multivariate analysis of micro-Raman spectra of thermoplastic polyurethane blends using principal component analysis and principal component regression

Abstract

Supporting information includes the rationale for the selection of TPUs, ^{13}C NMR spectra used to estimate the fraction of hard and soft segment present in each blend, and a brief discussion of post-processing methods employed to correct for spectrometer drift. Matlab programs were not included because analysis is not particularly novel with respect of the extraction of principal components (i.e., using the SVD algorithm), methods of cross-validation, or presumably, spectrometer alignment correction.

Rationale of TPU selection

TPU samples were chosen for three reasons: (1) they represent common, commercially available TPUs, (2) have similar melting points, and (3) exhibit slight differences in their bulk hardness values.

With commercially available TPUs, product consistency is ensured. Unfortunately, this is at the expense of detailed chemical information about the TPU hard and soft segments. Regardless, the primary position of this study is to demonstrate the place of multivariate methods in polymer materials research. Second, minimal separation in melting points between the TPU specimens facilitated efficient mixing with limited thermal degradation. Initially, a TPU with a higher melting point was blended with a lower melting point TPU (ΔT_m= 60°C) using a laboratory scale mixing-molder at atmospheric pressure. This resulted in the significant thermal degradation of the lower melting point TPU as evidenced by a color change and a large fluorescence background in Raman scanning.

Third, as stated in the introduction, differences in TPU mechanical properties, in this case hardness, are linked directly to polymer morphology contingent on both thermal history and polymer chemistry. In the present study, molding and annealing conditions are identical for each TPU blend thus minimizing the influence of thermal effects on the phase separation of TPUs.

^{13}C NMR spectra of EST85 and EST92 in d-DMSO

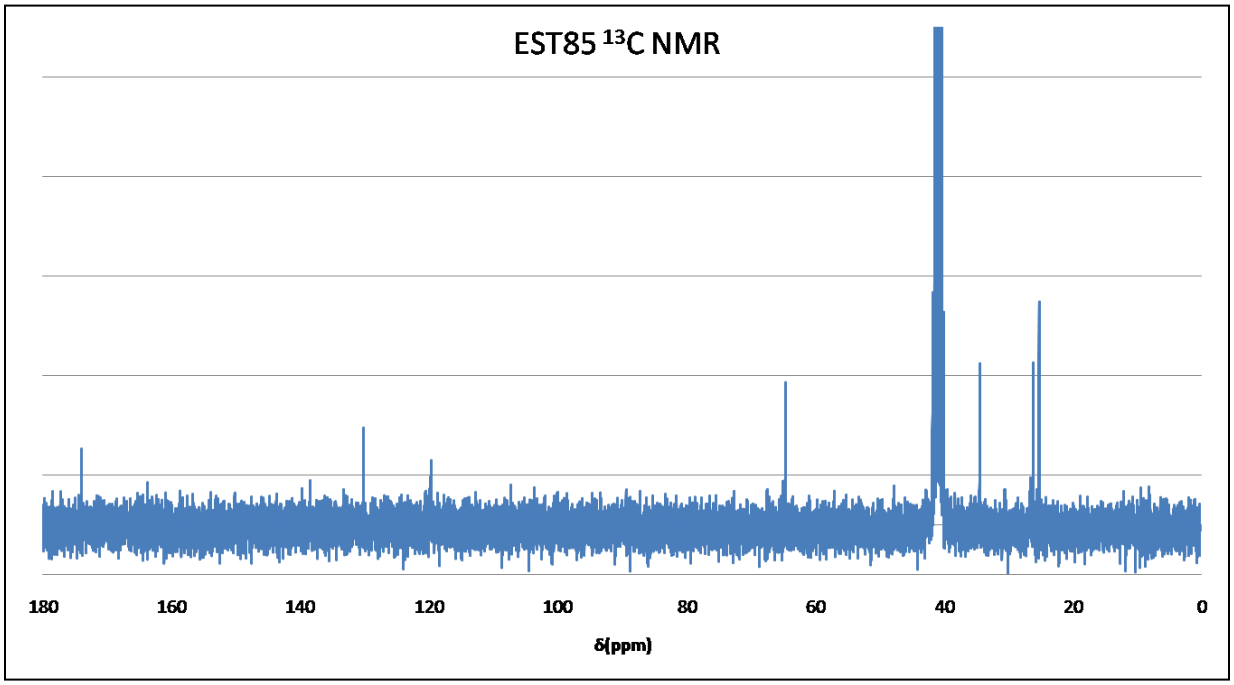


Figure D.18: NMR spectrum of EST85. Chemical shift identification is shown in Table D.1.

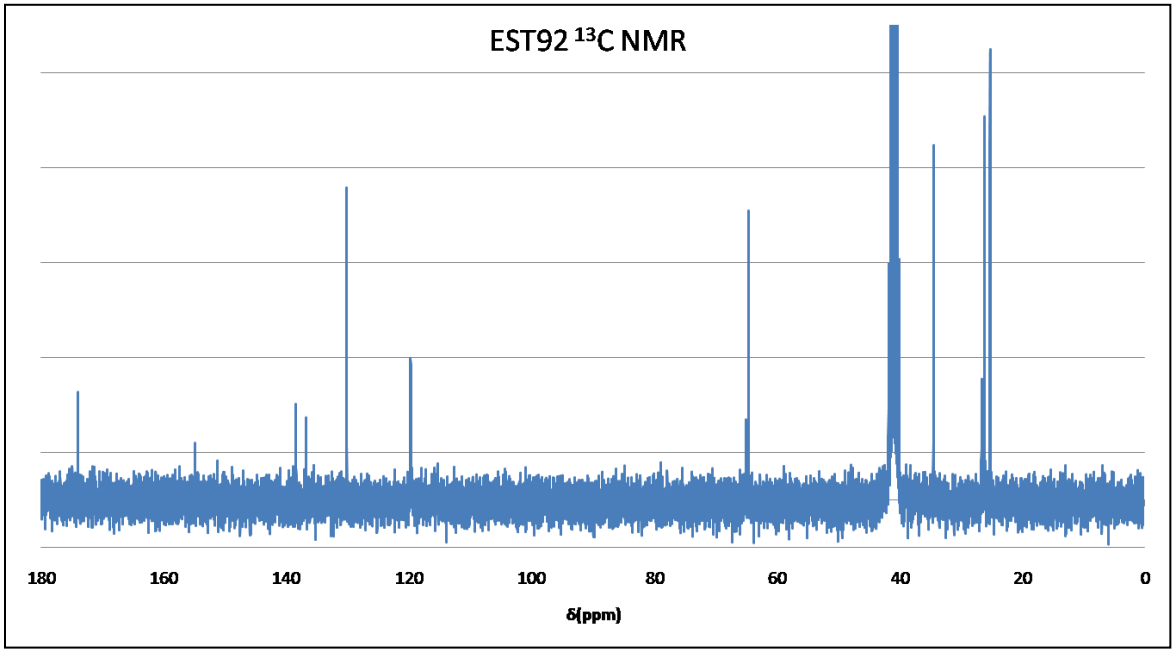


Figure D.19: ^{13}C NMR spectrum of EST92. Chemical shift identification is shown in Table D.2.

Table D.1: Chemical shift identification for EST85 in d-DMSO. Reference values are shown for comparison. The low signal-to-noise ratio prevented the identification of all useful chemical shifts.

Code*	HS_{ref} (ppm)	HS_{DMSO}(ppm)
A1	155	
A2	139.1	
A3	119.3	119.7
A4	130.2	130.1
A5	136.8	
A6	41.4	
A7	65	64.6
A8	26.3	26.6
Code	**SS_{ref} (ppm)**	**SS_{DMSO}(ppm)**
B1	174	174.0
B2	34.8	34.4
B3	25-26	25.2
B4	65	64.6
B5	25-26	26.1

*Coding scheme and reference values from: A. J. Brandolini, D. D. Hills. NMR spectra of polymers and polymer additives. New York, NY: Marcel Dekker. 2000. Pp. 470-471.

Table D.2: Chemical shift identification for EST92 in d-DMSO. Reference values are shown for comparison.

Code	HS_{ref} (ppm)	HS_{DMSO}(ppm)
A1	155.0	154.9
A2	139.1	138.4
A3	119.3	119.7
A4	130.2	130.1
A5	136.8	136.8
A6	41.4	41.6
A7	65.0	64.6
A8	26.3	26.6
Code	**SS_{ref} (ppm)**	**SS_{DMSO}(ppm)**
B1	174.0	174.0
B2	34.8	34.4
B3	25-26	25.2
B4	65.0	64.6
B5	25-26	26.1

Additional sources of error for PCA

Major PCA modeling error was introduced by voltage fluctuations in the laser power supply which produced a ±2 cm^{-1} shift or "drift" in the Rayleigh peak position. Realigning the spectrometer (i.e., defining a slightly different laser wavelength in the Witec Control Software) every 5 minutes during operation eliminated the influence of drift on PCA and PCR. The term "modeling error" is used to distinguish the fact that voltage supply fluctuations had the most effect on the PC-modeling in contrast to the linearly superimposed instrument noise sources found in Raman spectroscopy. Instrument noise, such as shot noise and detector background noise, did influence PC-modeling, albeit, to a lesser extent. Their influence was mitigated by sufficient integration time, adequate background removal, and noise reduction inherent to PCA.

Spectrometer drift was also numerically countered by employing a user-created program where the Rayleigh peak of each spectrum was fitted using a Gaussian function whereby the line-shape was estimated using a non-linear, Levenberg-Marquardt optimization routine. This form of post-processing, drift correction improved the PCA and PCR results.

Appendix E

Supporting information for Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least-squares regression

Andrew T. Weakley, Arthur Miller, Peter R. Griffiths, Sean J. Bayman, Anal Bioanal Chem, 2014, DOI: 10.1007/s00216-014-7856-y (available online)

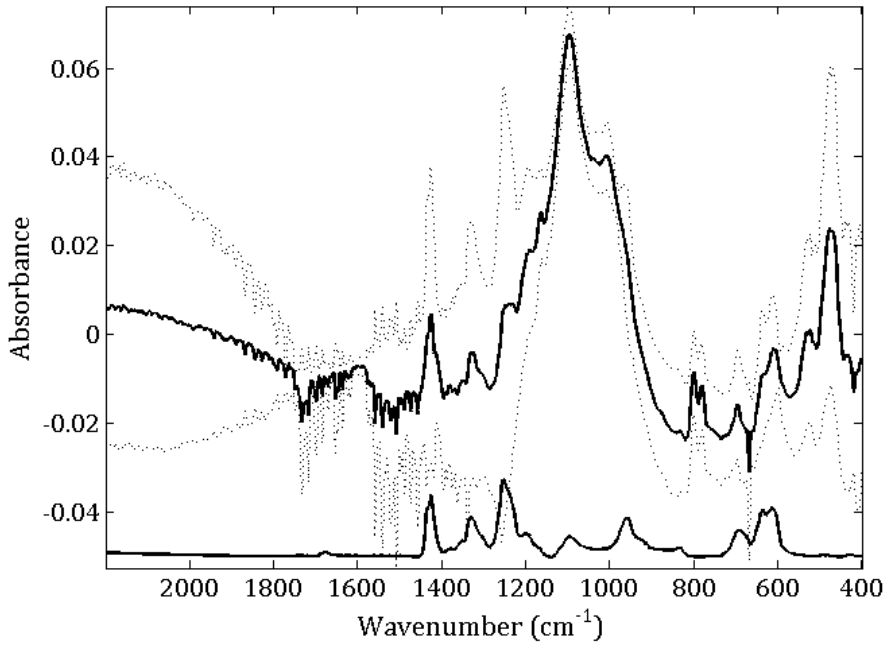


Fig E.13 Average spectrum calculated using 20 granite mine dust samples (top, solid) with accompanying +/- one standard deviation on each absorbance (dashed). A baseline corrected and scaled spectrum of PVC is also plotted with -0.05 absorbance offset. Spectra from this mine exhibited the fewest mineral and PVC confounders as indicated by a fairly flat baseline, clear α-quartz doublet, and absence of bands at wavenumbers greater than 1500 cm^{-1}. Note, imperfect background correction artificially caused negative absorbance features and offset in the average spectrum

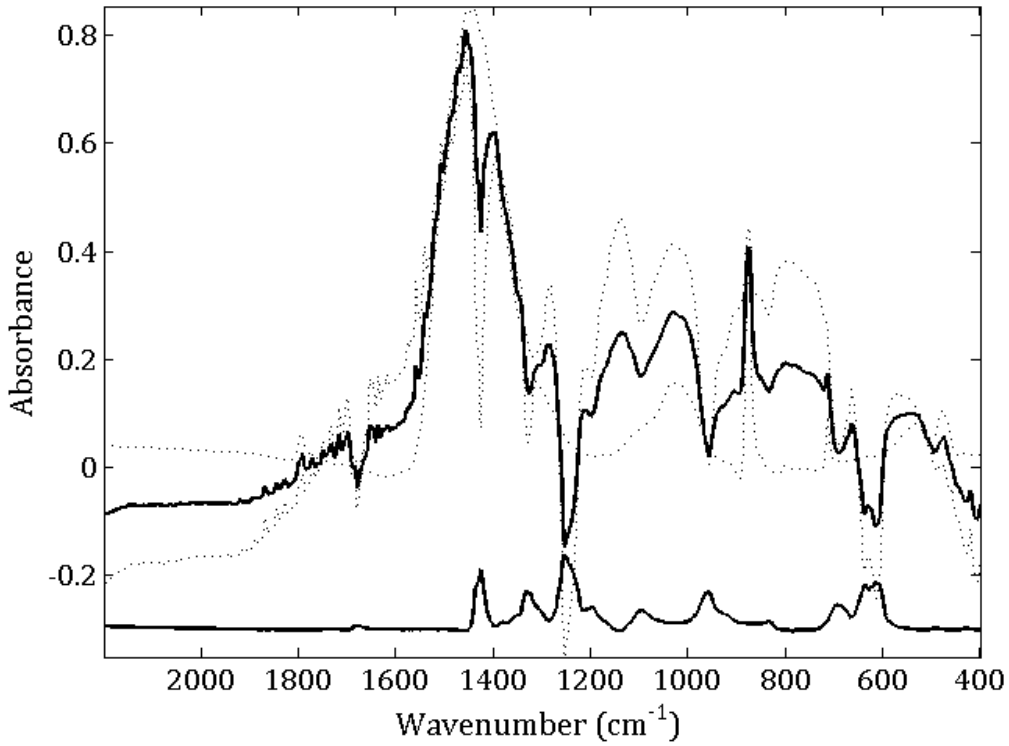


Fig E.14 Average spectrum calculated using 8 limestone mine dust samples (top, solid) with accompanying +/- one standard deviation on each absorbance (dashed). A baseline corrected and scaled spectrum of PVC is plotted with -0.3 absorbance offset. This spectrum again has a large negative contribution from PVC bands which is why the average absorbance spectrum has negative artifacts at wavenumbers commensurate with PVC bands. The large absorbance at 1440 cm^{-1} indicates the dominant presence of carbonate[34] and the apparent absence of the α-quartz doublet indicates a very low silica content

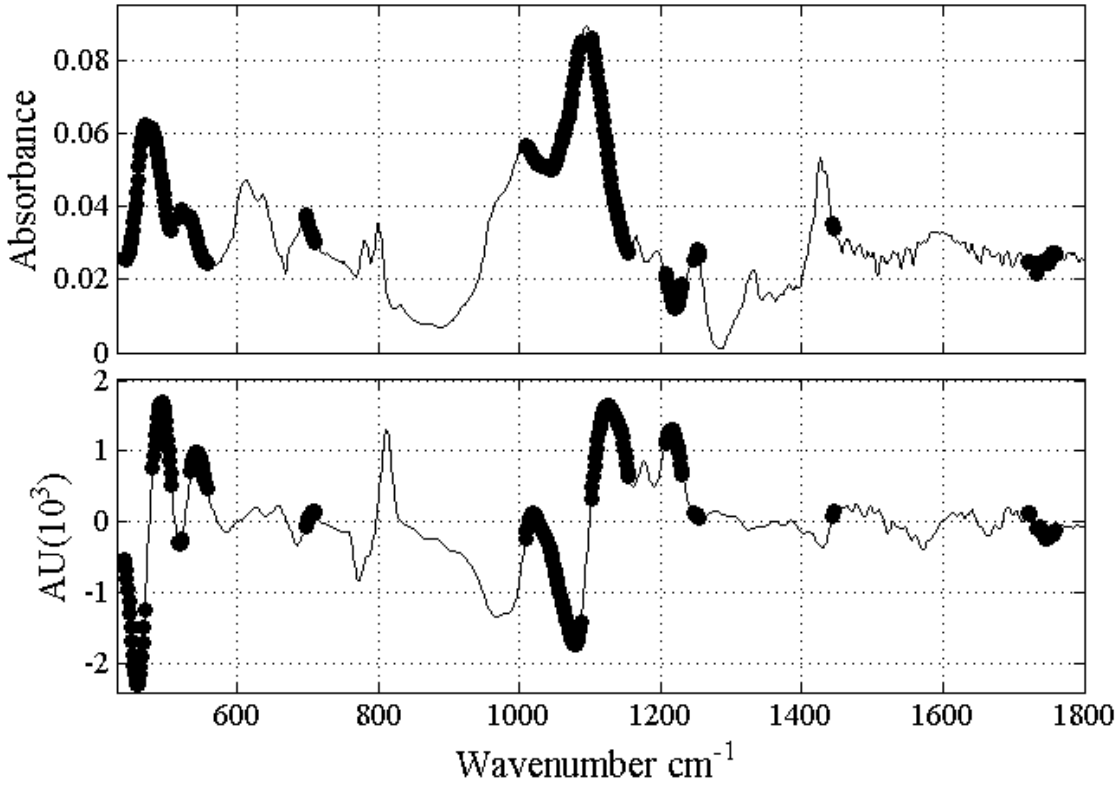


Fig E.15 The 223-wavenumber model was identified with the assistance of BMCUVE for the half-spectrum (sans "α-quartz doublet") PLS regression. This model indicated a prediction minimum (RMSEP) relative to alternative models. Notably, some features non-essential to predicting silica are evident (> 1400 cm^{-1})

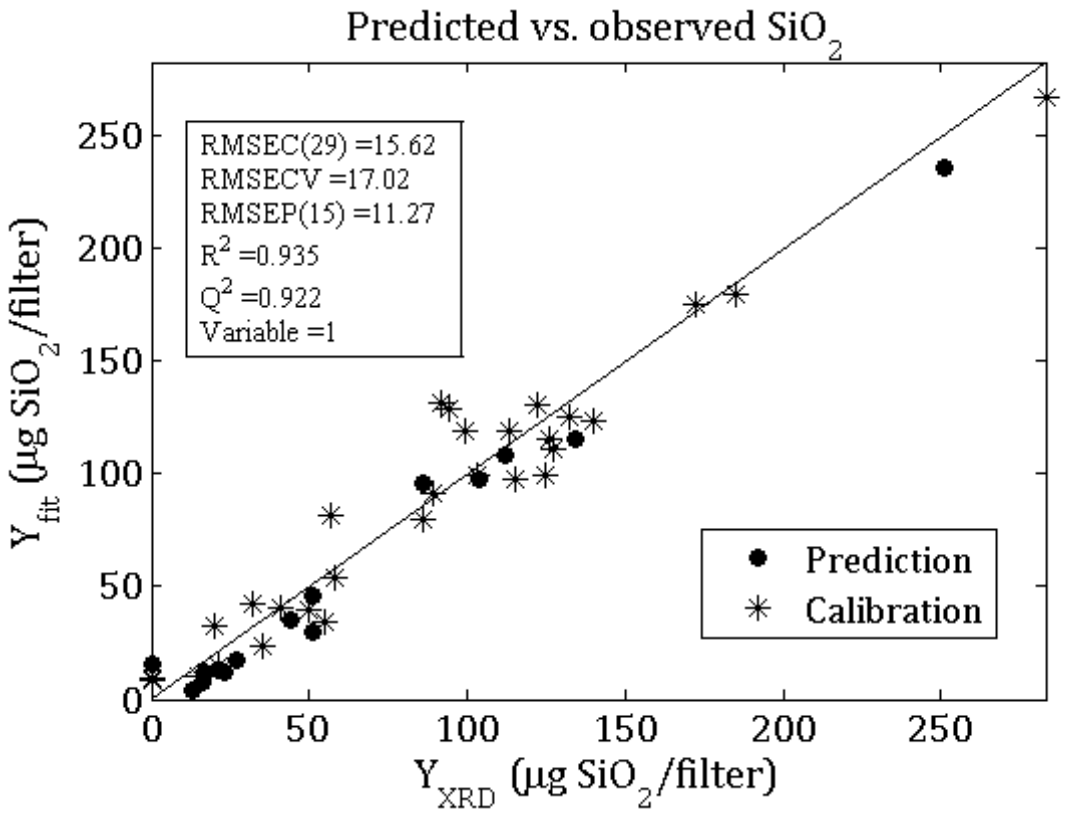


Fig E.16 Predicted versus measured silica mass using 29 training and 15 testing samples. Manual integration of the α-quartz doublet in each spectrum (using 37 variables) resulted in a single predictor variable which was regressed against the XRD-estimated mass of silica (Y_{XRD}) to foster calibration (Y_{fit})

Appendix F

Supporting information for Supporting Information for Model Selection in Partial Least-Squares Regression Using Backward Monte Carlo Unimportant Variable Elimination

Preprocessing

Baseline correction [1], spectral artifacts, and any residual background contributions were minimized using a 2nd-order, 21-point, 1st-derivative Savitzky-Golay transformation for the TPU Raman data. These 1st-derivative spectra were next screened using principal component analysis (PCA) score plots and leverage statistics. A 5-fold cross-validation was used to estimate the number of principal components included in the PC model. Specifically, components showing a ${Q_l}^2$statistic less than 0.95 were deemed useful to the PCA and retained for further $[X]$-block outlier assessment [2]. Spectra were projected onto the retained principal components to subjectively judge their influence on the relative orientation of the principal axes. Grossly influential spectra were removed. Leverage statistics were next calculated for each spectrum using only the variance contained in the significant PCs [3]. Observations exhibiting leverage greater than 2.5 times the average were removed unless otherwise stated.

Spectra were preprocessed by mean-centering. Although not optimal for every data set in an absolute sense, this simple scaling procedure was adequate to evaluate the relative performance of the BMCUVE routine against alternative selection approaches. Next, the total number of available observations were partitioned (50/30/20) into calibration, validation, and prediction sets according the Kennard-Stone algorithm [4]. Descriptive statistics were developed to ensure that the character of the calibration and prediction samples were comparable. Most critically, if the range of y-observations in each set differed substantially, samples were exchanged manually between the calibration and validation sets to ensure commensurability.

Prediction error estimation and degrees of freedom in PLS models

Equation F.1 conveniently provides an estimate of the regression coefficient's covariance (required for Equation 6.6) bootstrapping-by-residuals where the covariance of each of the *p* regression coefficients for the k^{th} model is estimated as

$$var(\widehat{\boldsymbol{b}}) = \frac{1}{B-1}\sum_{i=1}^{B}(\widehat{\boldsymbol{b}}_i - \overline{\boldsymbol{b}})(\widehat{\boldsymbol{b}}_i - \overline{\boldsymbol{b}})^T. \qquad \text{Equation F.1}$$

To develop this expression, *B* (=5000) independent bootstrap samples are drawn randomly (with replacement) from the available calibration data, *B* PLS models are built using the bootstrap replicates, and a matrix of regression coefficients, $[\hat{B}]$ (*p* x *B*), are

developed. A column from $[\widehat{B}]$ is thus the $p \times 1$ vector of bootstrapped regression coefficients, $\widehat{\boldsymbol{b}}_i$.

It has been demonstrated that PLS models consume more than one degree of freedom for each PLS component [5, 6]. A useful, rapid first-order estimate of the degree of freedom uses the concept of average leverage. This is known as the pseudo-degrees of freedom (pdf) approximation and is calculated readily using the sample data and cross-validation as

$$pdf^k = N_c\left(1 - \sqrt{\frac{MSEF^k}{MSECV^k}}\right) \qquad \text{Equation F.2}$$

with $MSEF^k$ and $MSECV^k$ representing the mean-square error of fit ($= \frac{\left\|\hat{\boldsymbol{y}}_c{}^k - \boldsymbol{y}_c{}^k\right\|^2}{N_c}$) and cross-validation for the k^{th} iteration of BMCUVE, respectively. Cross-validation is required prior to Figure 6.1, Step 2 to estimate $LV(\#)^k$ making it readily available for use in Equation F.2.

Runtime reduction using bootstrapped confidence intervals (BCIs) prior to BMCUVE

Table 6.5 concisely describes the efficacy of prefiltering unstable predictors using bootstrapped confidence intervals prior to BMCUVE. Relative to the standard routine, a modest filtering using 90% or 95% BCIs reduces runtime by roughly 27%. Overall, a steady reduction in runtime is observed for increasingly wider BCIs. Comparing the 90% BCI filter to regular BMCUVE shows that a 10-minute reduction in time is observed by reducing three iterations (M_{obs}). For an identical reduction in iteration number, only a 2-minute reduction in runtime is observed between the 90% and 99% BCI filtered routines. A large, nonlinear influence of initial subset size (p_{in}) on the computational cost of the routine is thus apparent.

Table F.1 Computational cost (min) of the BMCUVE routine as a function of BCI filter width ($BCI_{(1-\alpha)}$) and *q(%)* for the large-scale TPU blend elimination. The number of LVs (= 5) remained fixed for each pass for ease of comparison. M_{est} and M_{obs} denote the number of iterations predicted by Equation 3 and those actually observed. The regular BMCUVE is denoted as such.

$BCI_{100(1-\alpha)}$	*q(%)*	p_{in}	M_{est}	M_{obs}	Time (min)
Regular	10%	1580	70	58	37
90%		1132	67	55	27
95%		1052	66	54	28
99%		873	64	52	25
99.9%		716	62	51	17
99.99%		578	60	47	13
Regular	20%	1580	34	31	18
	50%		11	12	7
	80%		5	6	4
	90%		3	5	4

Significant reductions in runtime are linked to the size of the *q(%)* parameter. This is unsurprising since the *q(%)* parameter determines the number of total iterations as well as the size of each predictor subset at each pass of the routine. Notably, the *q(%)* parameter also influences the number of potentially redundant passes of the routine. A smaller *q(%)* requires fewer passes than indicated by Equation 6.3. This indicates that smaller *q(%)* are more susceptible to redundancy, requiring the more frequent execution of the auxiliary elimination function (see **Methods**, **Step 4: Backward Elimination**).

Other considerations

Because the $PICP_{1-\alpha}$ statistic assumes *t*-distributed residuals, normality for each PLS model's residuals was checked using one-sample Kolmogorov-Smirnoff tests (α = 0.05), normal q-q plots, and histograms of the model residuals.

References

[1] A. T. Weakley, P. R. Griffiths, and D. E. Aston, "Automatic Baseline Subtraction of Vibrational Spectra Using Minima Identification and Discrimination via Adaptive, Least-Squares Thresholding," vol. 66, pp. 519-529, 2012.

[2] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, pp. 433-459, 2010.

[3] T. Næs, Isaksson, T., Fearn, T. Davies, T., An User-friendly Guide to Multivariate Calibration and Classification. Chichester, West Sussex: Nir Publications, 2002.

[4] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," Technometrics, vol. 11, pp. 137-148, 1969.

[5] H. van der Voet, "Pseudo-degrees of freedom for complex predictive models: the example of partial least squares," Journal of Chemometrics, vol. 13, pp. 195-208, 1999.

[6] N. Krämer and M. Sugiyama, "The Degrees of Freedom of Partial Least Squares Regression," Journal of the American Statistical Association, vol. 106, pp. 697-705, 2011.