Article

Applied Spectroscopy

Applied Spectroscopy
2021, Vol. 75(1) 34–45

DOI: 10.1177/0003702820955245
journals.sagepub.com/home/asp
SAGE

# ISREA: An Efficient Peak-Preserving Baseline Correction Algorithm for Raman Spectra

**Yunnan Xu[1], Pang Du[1], Ryan Senger[2], John Robertson[3], and James L. Pirkle[4]**

## Abstract

A critical step in Raman spectroscopy is baseline correction. This procedure eliminates the background signals generated by residual Rayleigh scattering or fluorescence. Baseline correction procedures relying on asymmetric loss functions have been employed recently. They operate with a reduced penalty on positive spectral deviations that essentially push down the baseline estimates from invading Raman peak areas. However, their coupling with polynomial fitting may not be suitable over the whole spectral domain and can yield inconsistent baselines. Their requirement of the specification of a threshold and the non-convexity of the corresponding objective function further complicates the computation. Learning from their pros and cons, we have developed a novel baseline correction procedure called the iterative smoothing-splines with root error adjustment (ISREA) that has three distinct advantages. First, ISREA uses smoothing splines to estimate the baseline that are more flexible than polynomials and capable of capturing complicated trends over the whole spectral domain. Second, ISREA mimics the asymmetric square root loss and removes the need of a threshold. Finally, ISREA avoids the direct optimization of a non-convex loss function by iteratively updating prediction errors and refitting baselines. Through our extensive numerical experiments on a wide variety of spectra including simulated spectra, mineral spectra, and dialysate spectra, we show that ISREA is simple, fast, and can yield consistent and accurate baselines that preserve all the meaningful Raman peaks.

## Keywords

Iterative smoothing-splines with root error adjustment, ISREA, Raman spectroscopy, baseline correction

Date received: 24 February 2020; accepted: 12 August 2020

## Introduction

Raman spectroscopy is an established tool used for both qualitative and quantitative analyses of molecular composition of macro- and nanomaterials and biological systems.[1] More recently, advances in instrumentation have improved sensitivity greatly, reduced interference from fluorescence and environmental sources, and have led to spectrometers that are relatively inexpensive and have a small footprint.[2,3] It has also found applications in a variety of medical studies, such as disease diagnosis and monitoring efficacy and progress of therapy.[4–8]

In Raman spectrum generation, a background signal generated by fluorescence or Rayleigh scattering can heavily interfere with accurate analysis of the underlying Raman spectrum. This background signal, commonly known as the baseline, often appears as a smooth curve in the raw spectrum. Therefore, one critical step in Raman spectroscopy is to perform a baseline correction that involves estimating the baseline by a smooth function and then removing it from the raw spectrum by subtraction.[9]

Numerous baseline correction methods have been proposed over the years. Almost all baseline correction methods implement an algorithm that employs a smoother to capture the smooth trend and a loss function to adjust the fitting. Commonly used smoothers include first- and second-order differentiation, Fourier transformation,[10,11]

[1]Department of Statistics, Virginia Tech, Blacksburg, VA, USA
[2]Department of Biological Systems Engineering, Virginia Tech, Blacksburg, VA, USA
[3]School of Biomedical Engineering and Sciences, Virginia Tech, Blacksburg, VA, USA
[4]School of Medicine, Wake Forest University, Winston-Salem, NC, USA

**Corresponding Author:**
Pang Du, Virginia Tech, 250 Drillfield Dr, Blacksburg, VA 24061-0131, USA.
Email: pangdu@vt.edu

polynomial fitting,[12,13] splines,[14–16] and wavelets.[17–19] For example, Zhang and Ben-Amotz used the Savitzky–Golay second-derivative method on spectral data.[20] Although differentiation is unbiased and efficient for fluorescence subtraction, it can severely distort shapes of Raman spectra and relies on complex fitting algorithms to reproduce conventional spectra according to Mosier-Boss et al.,[10] who applied the fast Fourier transform filtering technique to Raman spectra to eliminate interference due to fluorescence. Fourier transform filtering depends on direct human intervention to choose its upper and lower limits in the frequency domain each time. This process is tedious and does not lend well to automated analyses. In contrast, polynomials are simple and convenient and thus popular in the biomedical field. To preserve peak intensities of Raman spectra, they often rely on manual recognition of background points. Otherwise the fitted baseline would include both fluorescence background and Raman peaks. Furthermore, the manual selection of background points can be time consuming, which, again, does not lend well to automated spectra processing. To avoid the human-intervention input to the selection process, iterative procedures were considered.[11,13] Splines were used in baseline correction as early as the 1990s.[14] More recently, Cai et al. presented a method that combines penalized B-splines with vector transformation.[21] Wavelet transformation has been another popular tool for baseline correction. For example, Cai et al. applied the multi-resolution wavelet transformation and suppressed the empirical wavelet coefficients in groups with a blockwise threshold.[17] However, the major drawbacks of wavelets are their assumption of a well separated background from the rest of the signal[22] and the need of selecting the wavelet type and the wavelet coefficient threshold. Furthermore, wavelets may sometimes lead to a sub-optimal filter for experimental signals.[18] A second-generation adaptive wavelet transform, which makes use of a spatial domain to generate new wavelet filters, was also developed.[18] Yet, it is still complex and computationally expensive to implement, limiting its practical application. Also commonly seen are model-based approaches where baseline correction involves specifying models for additive and multiplicative forms of background noises. One example is the singular value decomposition-based method where multivariate loadings were used for background correction.[23] Another example is the extended multiplicative signal correction (EMSC), which was first proposed for the near infrared spectroscopy[24] and later extended to Raman spectra.[25–27] An EMSC model decomposes a raw spectrum into three components: a polynomial baseline function, a multiple of a reference spectrum, and the residual that actually contains the spectrum of interest for the scanned sample. Through an ordinary or weighted least squares estimation, it produces spectra similar to the reference spectrum. The choice of the reference spectrum thus depends on the ensuing analysis. When peak heights or related information are not required, a commonly used reference spectrum is the mean spectrum. Otherwise, a baseline-corrected reference spectrum should be used to achieve baseline correction for all spectra.[28]

In addition to the smoother, the choice of the loss function is also critical for baseline correction. Most existing methods rely on a symmetric loss function, generally the least square loss. This is now recognized as inappropriate for the baseline correction purpose[29] as it tends to produce a baseline that invades into Raman peaks. More specifically, a high peak in a raw spectrum is often expected to be made up of a Raman scattering signal represented by high peaks and a smooth baseline signal that forms the bottom of the peaks. A fitted function based on the least squares loss, however, often cuts into the peak areas instead of properly estimating the bottom of the peaks. Therefore, the heights of the peaks in the baseline-corrected spectrum would be smaller than their true values. This can create problems for ensuing quantitative analysis of Raman spectra since, according to Beer's Law, there is a proportional relationship between the height of a peak and the concentration of the molecule(s) creating it. The deficiency of the least squares loss lies in that it often pulls up the fitted curve to match up with a peak in order to minimize their squared difference.

Based on the limitations noted in these observations, several asymmetric loss functions have been proposed. These asymmetric functions share the common feature of a reduced loss for large positive deviations compared with the least squares loss. The asymmetric least squares (ALS) loss was first proposed, which is essentially a weighted least squares with second-order derivatives as the penalty term.[30] However, the ALS may produce artificial negative peaks on the corrected spectrum.[31] Peng et al. generalized this approach for multiple spectra baseline correction taking advantage of means of the similarity among the multiple spectra.[32] They assumed that baseline stays the same or changes little for spectra of samples collected continuously over time. He et al. proposed an improved ALS method, which adds the first-order derivative of residuals to the least squared loss to achieve smoothness and used second-order polynomials as the smoother.[33] Mazet et al. replaced the symmetric squared loss with asymmetric Huber loss and asymmetric truncated quadratic loss to suppress the effect of large positive residuals, i.e., peak areas, and estimated the baseline by a low-order polynomial.[22] But selections of two tuning parameters, the threshold of loss functions and the order of polynomial, can be tricky. The order of polynomial demands manual selection based on the smoothness of background. The authors suggested that splines may provide a better fitting than polynomials.

More recently, Liu et al. developed the Goldindec method using polynomials with an asymmetric Indec loss and implemented it through a half-quadratic algorithm.[29]

This method was chosen for baseline correction in the initial versions of the Raman chemometrics (Rametrix) LITE and PRO Toolboxes for Matlab.[5,7] Among asymmetric losses, the asymmetric Indec loss has been shown to have the best performance. However, one major drawback of the Goldindec algorithm is its use of polynomials to represent a baseline signal. Polynomials, especially low-degree polynomials, are often used to fit simple smooth functions. Due to its small number of tuning parameters (the coefficients in a polynomial), polynomials may not be sufficiently flexible to capture complicated smooth trends and can be easily distorted by a few influential points. Another drawback is that Goldindec requires the selection of a change point in the asymmetric Indec loss function. This change point is critical for the success of the algorithm since it specifies the threshold such that a positive deviation beyond it will be less penalized than in the least squares loss. However, there is no systematic way to choosing this change point although some empirical (i.e., subjective) choices are suggested and their performance may not be satisfactory in practice.

As described above, a simple polynomial or smoothing spline fitted to a raw Raman spectrum is likely to produce a baseline that moves up with peaks, while a true baseline should leave the peaks untouched and take out only the background signals. That is, after the subtraction of a well-estimated baseline, the remaining spectrum should have the peak intensities preserved in the peak areas and intensities close to zero in the non-peak areas. Therefore, our goal is to develop a baseline correction method whose baseline estimate stays close to the true baseline with all the interesting peaks well-preserved.

In this article, we propose a new computationally efficient algorithm, called the iterative smoothing-splines with root error adjustment (ISREA). In ISREA, the baseline is fitted by smoothing splines, given their better flexibility in capturing the overall shape of the spectrum. To correct the aforementioned peak-invading problem of smoothing spline baseline estimate, we notice that peak invasion mostly happens in the regions of "positive deviations" or "positive prediction errors", that is, the regions where the observed intensity deviates from the fitted baseline intensity by a positive amount. Therefore, we propose the following iterative fitting procedure to adjust for the peak invasion. In each iteration, the prediction errors are adjusted down through a root transformation and added back to the fitted baseline intensities to form a new set of intensities. Then smoothing splines are applied to this new set of intensities to obtain a new baseline estimate, based on which a new set of prediction errors are calculated. This adjustment procedure is repeated until the errors drop to a negligible level. The error transformation used in the adjustment is motivated from asymmetric loss functions where large positive deviations are penalized less than in a least square loss. So, the ISREA baseline estimate inherits the nice properties of those estimates obtained from asymmetric losses. Furthermore, the simple root error adjustment avoids the tricky optimization of a non-conventional objective function otherwise required in all the methods based on asymmetric losses.

Therefore, it is much easier to implement, much more computationally efficient, and lends well to automated analyses.

In our numerical experiments, we compared ISREA with the ALS[30] and Goldindec algorithm[29] on both simulated and real Raman spectral data. In simulations, we considered both spiky and non-spiky data. In real applications, we studied spectra of pure mineral data from public databases and spectra of waste dialysate samples collected from patients undergoing hemodialysis treatments in a clinic. The spectrum of a pure mineral generally contains a small number of Raman peaks. Additionally, there is an expert-corrected spectrum available so that a "true" baseline can be recovered as the reference. On the other hand, a waste dialysate sample is a complicated solution containing many molecules, so its spectrum is expected to contain many Raman peaks. Furthermore, there is no ground truth or expert correction available as the reference although some signature chemicals such as urea are always present in the sample. Our experiments show that ISREA is adaptive to spectra with sparse or dense Raman peaks and has better performance on both simulated and real data.

## Methods

### *Notation and Asymmetric Losses for Baseline Correction*

Informed consent for the collection of urine specimens from healthy volunteers was obtained. A raw Raman spectrum consists of a sequence of intensity measurements $y_i$ at Raman shifts or wavenumbers (in $cm^{-1}$), $i = 1, \ldots, n$. Let $\mathbf{y}_{n\times 1} = (y_1, \ldots, y_n)^T$ be the vector of observed intensities. The model for a raw Raman spectrum is

$$y = m + a + \varepsilon \tag{1}$$

where $\boldsymbol{m}_{n\times 1} = (m_1, \ldots, m_n)^T, \boldsymbol{a}_{n\times 1} = (a_1, \ldots, a_n)^T$, and $\varepsilon_{n\times 1} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, are respectively vectors of unknown true baseline intensities, peak intensities, and random noises. In particular, the baseline intensities $m_i$ are often assumed to come from an unknown smooth baseline function $f$, say, defined on the interval $[0, 1]$, such that $m_i = f(i/n)$. The goal of a baseline correction procedure is thus to recover this unknown baseline function $f$.

While smoothers, such as polynomials or splines, are often used to model the baseline function $f$, a proper model for peak intensities $a_i$ is extremely hard. Each peak in the spectrum represents a specific molecular structure present in the scanned sample. When all the chemical

compositions of the sample are known, the peaks can be modeled as properly sized spikes at expected wavenumbers on the spectrum. In practice, the composition of the sample (e.g., the waste dialysate sample in our hemodialysis experiment) is often unknown, making it hard or even impossible to model the peak intensities properly. Instead, most baseline correction methods simply build up a smoother through the minimization of a loss function without any explicit modeling of peak intensities $a_i$. That is, the baseline is estimated essentially with peak intensities $a_i$ absorbed into the random error part of model in Eq. 1. For example, when the least squares loss $L_{lse}(x) = x^2$ is used, a smoother is trained through the minimization of $\sum_{n=1}^{n} \delta_i^2$ where $\delta_i = y_i - \hat{m}_i$ is the deviation of the fitted baseline from the observed intensity, and $\hat{m}_i = \hat{f}(i/n)$ is the fitted baseline intensity derived from the smooth function estimate $\hat{f}$ of $f$.

As reviewed in the Introduction section, the baseline function estimate $\hat{f}$ based on the least squares loss often cuts into the peak areas. To address this issue, various asymmetric loss functions $L_s(x)$, where $s > 0$ is a pre-specified threshold, are introduced such that $L_s(x) = x^2$ when $x < s$ but $L_s(x) < x^2$ for $x \geq s$. The principle for such a setup is to reduce the punishment otherwise enforced by the least squares loss when the difference $\delta_i$ is a big positive number, a phenomenon often observed at the peak area. For example, the left panel of Fig. 1 shows several such asymmetric loss functions. From top to bottom, the curves are respectively the asymmetric Huber function with $L_s(x) = 2sx - x^2$ when $x \geq s$, the asymmetric truncated quadratic function with $L_s(x) = s^2$ when $x \geq s$, and the asymmetric Indec function with $L_s(x) = \frac{s^3}{2x} + \frac{s^2}{2}$ when $x \geq s$.

These asymmetric loss functions all share the same shape with the least squared loss for negative deviations while they have reduced loss for large positive deviations. This helps discourage invasions into the high peaks commonly seen in a baseline minimizing the least squares loss. However, these asymmetric loss functions have a couple of critical drawbacks: (1) the selection of threshold $s$ can be tricky; (2) the optimization of these loss functions is often time-consuming due to their nonstandard forms. To address the issue of threshold selection, we propose the following threshold-free asymmetric root error loss function

$$L(x) = \begin{cases} x^2, & \text{if } x \leq 0 \\ \sqrt{x}, & \text{if } x > 0 \end{cases} \tag{2}$$

As plotted in the right panel of Fig. 1, the asymmetric square root error function clearly preserves the discouragement of large positive deviations. And it has the advantage of not requiring any manual selection of a threshold. To further improve the computational efficiency, we shall avoid the direct optimization of the loss function in Eq. 2. Instead, we propose a computational procedure that iterates between smoothing and updating "observations" with positive deviations by the sums of current smooth baseline estimates and square-root-adjusted deviations.

## The ISREA Algorithm

We now introduce our algorithm called the ISREA. The basic idea is to yield a baseline function estimate that targets at minimizing a loss function similar to Eq. 2 without really resorting to its direct optimization which can be complicated and time consuming.

In the ISREA algorithm, we first obtain an initial baseline estimate by fitting smoothing splines to the raw spectrum. Recall that $\delta_i = y_i - \hat{m}_i$ is the deviation of the smooth baseline estimate $\hat{m}_i$ from the observed raw spectral intensity value $y_i$. Then, intensities are adjusted in a way such that in the areas with $\delta_i \leq 0$ the intensities remain the same, while

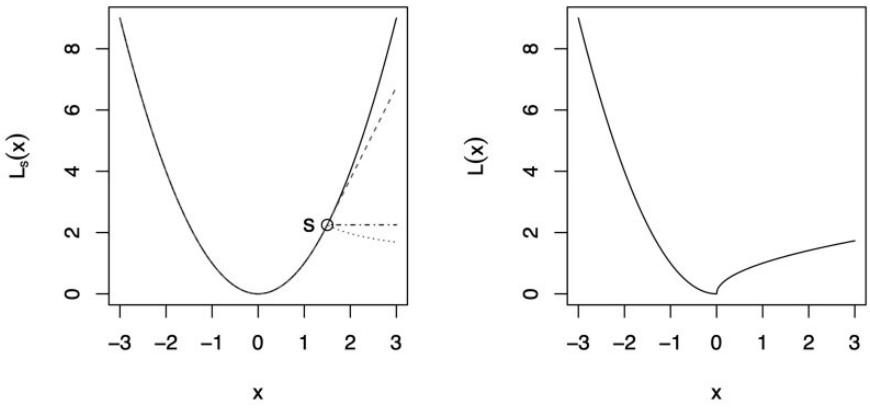


**Figure 1.** Loss functions in baseline correction methods. Left: Existing asymmetric loss functions (from top to bottom, the asymmetric Huber function as the dashed line, the asymmetric truncated quadratic function as the dash-dotted line, and the asymmetric Indec function as the dotted line) with the least squares loss function (solid line) imposed. Right: Proposed asymmetric root error loss function.

in the areas with $\delta_i > 0$ intensities are updated as $y_i^{(new)} = \hat{m}_i + \sqrt[4]{\delta_i}$. To see how this adjustment is related to the asymmetric root error loss function $L(x)$ in Eq. 2, we note that squared errors at areas with non-positive deviations remain $\delta_i^2$, while squared errors at areas with positive deviations become $\sqrt{\delta_i}$. We then feed the new intensities $y_i^{(new)}$ into smoothing spline estimation again to get an updated baseline function estimate $\hat{f}$ and thus updated baseline intensity estimates $\hat{m}_i = \hat{f}(i/n)$.

These steps are repeated until the difference between two consecutive fitted baselines is small. The complete algorithm of ISREA is shown in Algorithm 1 in Fig. 2.

In each iteration, intensities at peak areas are reduced and intensities at non-peak areas remain the same. So smoothing splines are actually fitted to a modified spectrum with reduced peak heights. Consequently, the fitted baseline can stay low at the bottom of peaks like the true baseline and provides a properly corrected spectrum with retained peak heights.

Function *Smooth.spline* in the R package stats is used to fit a cubic smoothing spline to the data. For the number of knots, we use five for general raw spectra and 15 for really spiky raw spectra. For the constant $\varepsilon$ in the convergence criterion, we have tested different choices of $\varepsilon$ in a range from 10 to 0.0001 on simulated spectra, dialysis spectra, and mineral spectra. The results, collected in the supplemental material, are quite similar and do not appear to be sensitive to its choice. More time is taken as $\varepsilon$ decreases since more iterations are needed.

## Numerical Studies

In this section, we compare the ISREA with two existing baseline correction methods, the ALS and Goldindec. For the ALS, we use the "als" method of the baseline function in the R package baseline with the options lambda = 7.5 and weight = 0.05, as found by optimizing baseline accuracy over a grid of plausible values.[34] For the Goldindec, we use degree three polynomials and a peak ratio of 0.5 as suggested in the paper by Liu et al.[29] We test the ISREA with different choices of the number of knots.

### Simulated Spectra

Simulated data were generated from the assumed Raman spectrum model Eq. 1. In the simulation, the true baseline intensity $m$ was set to be a polynomial function of degree five whose coefficients were generated from Normal distributions. The first five coefficients were generated from $N(0, 10^2)$ while last one was generated from $N(0, 0.01^2)$. Peaks were simulated from groups of Gaussian distributions, with the number of them randomly selected between 1 and 10. Their central locations were randomly set within the spectral range. Standard deviations of those peaks were independently selected in the range from 1 to 15 such that the width of those artificial peaks was similar to peaks on real Raman spectra. Noise signals were generated from $N(0, 1)$. One-thousand Raman spectra were simulated this way and tested with the ISREA and Goldindec.

We used a statistical measure, called AC_rate, to evaluate the performance of ISREA. It was proposed by Liu et al.[29] to assess the accuracy of the Goldindec method and defined as

$$AC\_rate = 1 - \frac{\_\boldsymbol{m} - \hat{\boldsymbol{m}}\_2}{\_\boldsymbol{m}\_2} \tag{3}$$

**Algorithm 1:** The ISREA algorithm

**input :** Raw spectrum $\mathbf{y} = (y_1, \ldots, y_n)^T$ with intensity $y_i$ at Raman shifts $i = 1, \ldots, n$; convergence constant $\epsilon$.
**output:** Fitted baselines $\hat{\mathbf{m}} = (\hat{m}_1, \ldots, \hat{m}_n)^T$.

**begin**
  compute an initial estimate $\hat{\boldsymbol{m}}^{(0)}$ from the raw spectrum by smoothing splines.
  Set $D = 2\epsilon$ (or any number $> \epsilon$)
  **while** $D > \epsilon$ **do**
    $\boldsymbol{\delta}^{(t)} = \boldsymbol{y} - \hat{\mathbf{m}}^{(t)}$
    **if** $\delta_i^{(t)} \leq 0$ **then**
      $y_i^{(t)} = y_i$
    **else**
      $y_i^{(t)} = \hat{\boldsymbol{m}}^{(t)} + \sqrt[4]{\delta_i^{(t)}}$
    Replace $\boldsymbol{y}$ with $\boldsymbol{y}^{(t)}$ and fit a new baseline $\hat{\boldsymbol{m}}^{(t+1)}$ by smoothing spline
    $D = \|\hat{\boldsymbol{m}}^{(t+1)} - \hat{\boldsymbol{m}}^{(t)}\|_2^2$

**Figure 2.** The ISREA algorithm procedure.

where $\boldsymbol{m}$ represents the true/expert-fitted baseline intensity, $\hat{\boldsymbol{m}}$ represents the fitted baseline intensity, and $||\cdot||_2$ is the $L^2$ norm of a function on the spectral range. The AC_rate compares the true/expert-corrected baseline with the fitted baseline. Generally, a bigger AC rate that is close to one is desired. Besides the AC_rate, we also considered the root mean square error (RMSE) between the true and estimated baselines.[33] For simulated data and minerals data, we used AC_rate and RMSE to compare the ALS, ISREA, and Goldindec baseline estimates against the true baselines or expert-corrected baselines. For dialysis data, since there are no true/expert-correction baselines, we plotted baseline-corrected spectra of dialysis samples to compare the ALS, ISREA, and Goldindec methods.

Tables I and II present the AC_rate and RMSE percentiles for the ALS, Goldindec, and ISREA with different numbers of knots (NK). While the Goldindec and ALS methods yielded a satisfactory AC_rate for 50% of the simulated spectra, we also note the much-deteriorated AC_rate for the lower 25% of the spectra. In particular, the smallest AC_rates were even negative for both methods, indicating a complete miss of the true baseline. On the other hand, the AC_rates of the ISREA with different numbers of knots are consistently better than those of the Goldindec and ALS. In terms of the RMSEs in Table II, the ISREA completely dominated the other two methods with much smaller RMSEs. In general, although $NK = 5$ generally gave the best performance for ISREA, the AC_rates and RMSEs for $NK = 10$ and $NK = 15$ were remarkably close. This indicates that the ISREA method is not sensitive to different choices of the number of knots.

## *Mineral Spectra*

In this section, we compare the ALS, Goldindec, and ISREA methods on Raman spectra of minerals obtained from the RRUFF database.[35] The database provides both the raw spectra and expert-corrected spectra for many minerals, the differences of which can be treated as the "true" baselines.

The ALS, ISREA, and Goldindec methods were applied to spectra of six minerals, namely, andersonite, eastonite, marialite, parascholzite, sugilite, and wadeite. These six minerals, two of which also appeared in the Goldindec paper,[29] were selected to represent some typical variations of Raman peak locations (in the middle versus at the ends of the spectral domain), sharpness (sharp peaks versus low peaks), and spread (clustered peaks versus spread-out peaks). Compositions of these minerals are simple and pure. Therefore, their Raman spectra are also simple with clear Raman peaks. Fig. 3 compares the three versions of baseline-corrected spectra against the expert-corrected spectra. Clearly, both the ALS and the ISREA-corrected spectra almost overlap the expert-corrected spectra for all the six minerals, whereas Goldindec sometimes generated spectra that deviated from the expert-corrected spectra by large margins at parts of the spectral domain. Further examination of the AC_rates in Table III and RMSEs in Table IV confirmed that Goldindec did not do so well in baseline correction for three of the six minerals, namely, andersonite, eastonite, and wadeite. A close inspection of the raw spectra for these minerals revealed that their underline baselines have complicated shapes which are generally hard to be captured by polynomials. Next, we compare the ALS with the ISREA. In terms of the AC_rates in

**Table II.** Percentiles of the RMSE for ALS, Goldindec, and ISREA on 1000 simulated spectra.

| Percentile | 0% | 5% | 25% | 50% | 75% | 95% | 100% |
|---|---|---|---|---|---|---|---|
| ALS | 0.85 | 0.95 | 1.10 | 1.31 | 1.80 | 2.75 | 4.80 |
| Goldindec | 0.90 | 1.13 | 1.27 | 1.55 | 2.36 | 4.10 | 15.32 |
| ISREA | | | | | | | |
| $NK = 5$ | 0.05 | 0.10 | 0.18 | 0.27 | 0.35 | 0.45 | 0.68 |
| $NK = 10$ | 0.08 | 0.13 | 0.22 | 0.33 | 0.42 | 0.54 | 0.76 |
| $NK = 15$ | 0.08 | 0.15 | 0.26 | 0.38 | 0.48 | 0.61 | 0.91 |

Note: *NK* is the number of knots used in ISREA.
ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

**Table I.** Percentiles of the AC_rates for ALS, Goldindec, and ISREA on 1000 simulated spectra.

| Percentile | 0% | 5% | 25% | 50% | 75% | 95% | 100% |
|---|---|---|---|---|---|---|---|
| ALS | –0.6198 | 0.7748 | 0.9322 | 0.9555 | 0.9789 | 0.9853 | 0.9896 |
| Goldindec | –0.8795 | 0.6959 | 0.9107 | 0.9515 | 0.9700 | 0.9833 | 0.9895 |
| ISREA | | | | | | | |
| $NK = 5$ | 0.6909 | 0.9355 | 0.9843 | 0.9935 | 0.9965 | 0.9985 | 0.9995 |
| $NK = 10$ | 0.6178 | 0.9214 | 0.9812 | 0.9921 | 0.9957 | 0.9981 | 0.9994 |
| $NK = 15$ | 0.5832 | 0.9123 | 0.9785 | 0.9906 | 0.9950 | 0.9979 | 0.9992 |

Note: *NK* is the number of knots used in ISREA.
ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

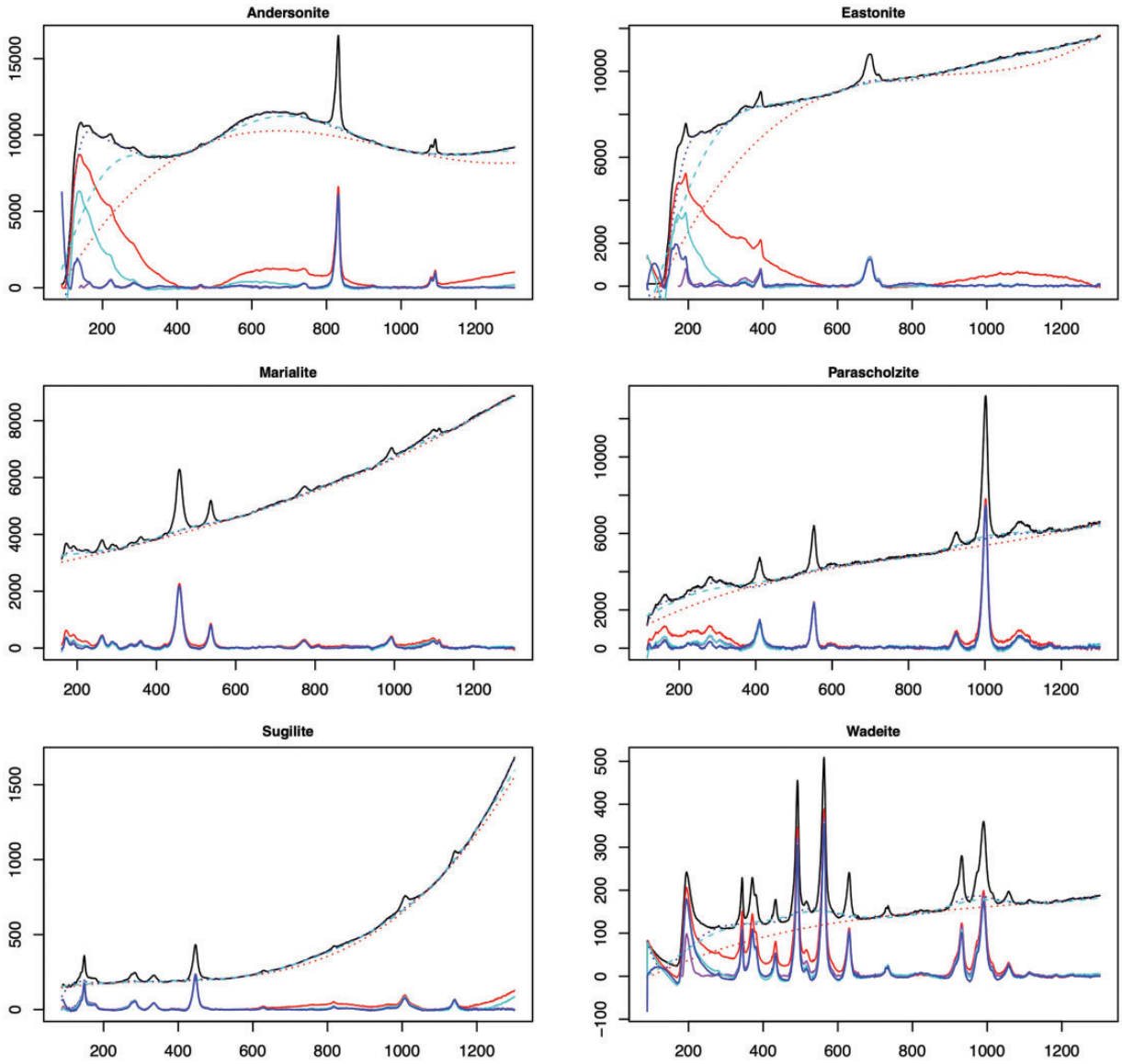


**Figure 3.** Comparison of ALS, Goldindec, and ISREA (with $NK = 15$) baseline correction on Raman spectra of six minerals against the expert-corrected spectra (purple solid). Horizontal axis: Raman shift (in $cm^{-1}$); vertical axis: intensity. The curves are respectively: raw spectra (black solid), ALS baseline estimates (cyan dashed), ALS-corrected spectra (cyan solid), Goldindec baseline estimates (red dashed), Goldindec-corrected spectra (red solid), ISREA baseline estimates (blue dashed), and ISREA-corrected spectra (blue solid).

Table III, the ALS and ISREA had similar AC_rates although the ISREA with different choices of *NK* had slightly higher AC_rates for five of the six minerals, except for paraschol-zite where both methods essentially had the same AC_rate. In terms of the RMSEs in Table IV, the ISREA with $NK = 15$ had smaller RMSEs than the ALS for five of the six minerals, except for parascholzite where the ALS had a slightly lower RMSE.

In our numerical experiments, we also considered different choices of NK for the ISREA. As Tables III and IV show, in general ISREA is not sensitive to the choice of NK with different NKs yielding similar AC_rates and RMSEs. This confirmed our finding in the previous simulations. One noticeable exception is the $NK = 5$ case for andersonite, where the sharp peak at the left end of the raw spectrum might be hard to capture when the number of knots is too small. As demonstrated by our results in Tables III and IV, this can be easily handled by using a slightly larger number of knots.

## *Dialysate Spectra*

The last set of spectra was collected in our hemodialysis experiment. Hemodialysis is one of the most common treatments for patients with end-stage kidney diseases. In a hemodialysis treatment session, typically about four hours

**Table III.** AC_rate comparisons of ALS, Goldindec, and ISREA for six minerals.

| | Andersonite | Eastonite | Marialite | Parascholzite | Sugilite | Wadeite |
|---|---|---|---|---|---|---|
| ALS | 0.8891 | 0.9412 | 0.9935 | 0.9825 | 0.9815 | 0.8916 |
| Goldindec | 0.7822 | 0.8546 | 0.9878 | 0.9413 | 0.9598 | 0.8044 |
| ISREA | | | | | | |
| $NK = 5$ | 0.8789 | 0.9479 | 0.9936 | 0.9740 | 0.9923 | 0.8910 |
| $NK = 10$ | 0.9653 | 0.9484 | 0.9966 | 0.9806 | 0.9913 | 0.9055 |
| $NK = 15$ | 0.9872 | 0.9818 | 0.9951 | 0.9824 | 0.9875 | 0.8938 |
| $NK = 20$ | 0.9800 | 0.9884 | 0.9924 | 0.9814 | 0.9823 | 0.9152 |
| $NK = 25$ | 0.9861 | 0.9805 | 0.9888 | 0.9717 | 0.9781 | 0.9140 |

ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

**Table IV.** RMSE comparisons of ALS, Goldindec, and ISREA for six minerals.

| | Andersonite | Eastonite | Marialite | Parascholzite | Sugilite | Wadeite |
|---|---|---|---|---|---|---|
| ALS | 1075.25 | 559.36 | 36.96 | 81.85 | 11.41 | 16.02 |
| Goldindec | 2110.84 | 1382.52 | 69.68 | 274.02 | 24.74 | 28.90 |
| ISREA | | | | | | |
| $NK = 5$ | 1174.22 | 494.93 | 36.68 | 121.57 | 4.77 | 16.11 |
| $NK = 10$ | 336.14 | 491.04 | 19.35 | 90.67 | 5.37 | 13.96 |
| $NK = 15$ | 124.54 | 173.34 | 27.91 | 82.24 | 7.72 | 15.69 |
| $NK = 20$ | 194.33 | 110.57 | 43.78 | 86.93 | 10.92 | 12.53 |
| $NK = 25$ | 135.10 | 185.63 | 63.91 | 131.90 | 13.47 | 12.71 |

ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

long, a machine called a dialyzer is connected to the patient to help partially purify blood by removing metabolic waste products and rebalancing water and electrolytes. The dialyzer pumps the blood to a filter chamber, composed of a semipermeable polysulfone membrane that separates the patient's blood from a fluid called dialysate. Wastes in the blood are released to the dialysate in the chamber, as fresh blood flows out of the chamber. Our hemodialysis study consisted of a total of 30 treatment sessions for 10 chronic kidney disease patients. The detailed description of the cohort is in the supplemental material. We now describe the analysis of the Raman spectral data from one treatment session since our analysis of other treatment sessions yield similar results. We collected used dialysate samples (containing metabolic wastes) at 10, 60, 120, 180, and 240 min (the end) of the session. Each sample was divided into 10 portions and each portion was analyzed by a Raman spectrometer (Peakseeker Pro 785, Agiltron Inc.) to produce a raw spectrum. Therefore, we generated a total of 50 raw Raman spectra with 10 spectra associated with each time point. As discussed earlier, there were no expert-corrected spectra of waste dialysate for references.

Visual comparisons of ALS, Goldindec, and ISREA are displayed in Fig. 4. Clearly, both $NK = 5$ and $NK = 10$ for ISREA produced consistent baseline estimates and yielded similar baseline-corrected spectra for the 50 raw spectra. ALS also generated consistent baseline estimates, with slightly lower peaks than those from ISREA. On the other hand, a good portion of the Goldindec baselines completely missed the trend at the right end of the spectral domain. Since peaks in this area can be associated with important molecules in used dialysate, this can cause serious trouble for ensuing quantitative analysis of the compositions of used dialysate samples.

To study this further, we also numerically examined how much ALS, ISREA, and Goldindec respectively preserved the similarities between spectra. As shown in Fig. 4, since all the raw spectra were from used dialysate samples collected in the same treatment session, they are all displayed similar trends. Such similarity is expected to be preserved even after the baseline correction. To represent this similarity, we paired the first raw spectrum with each of the other raw spectra, resulting in 49 pairs of spectra. For each pair, we calculated the correlation between the intensities of the two spectra. Then these 49 correlations represented the similarity between the 50 raw spectra. Similarly, we obtained 49 correlations each for the ALS baseline-corrected spectra, the Goldindec baseline-corrected spectra,

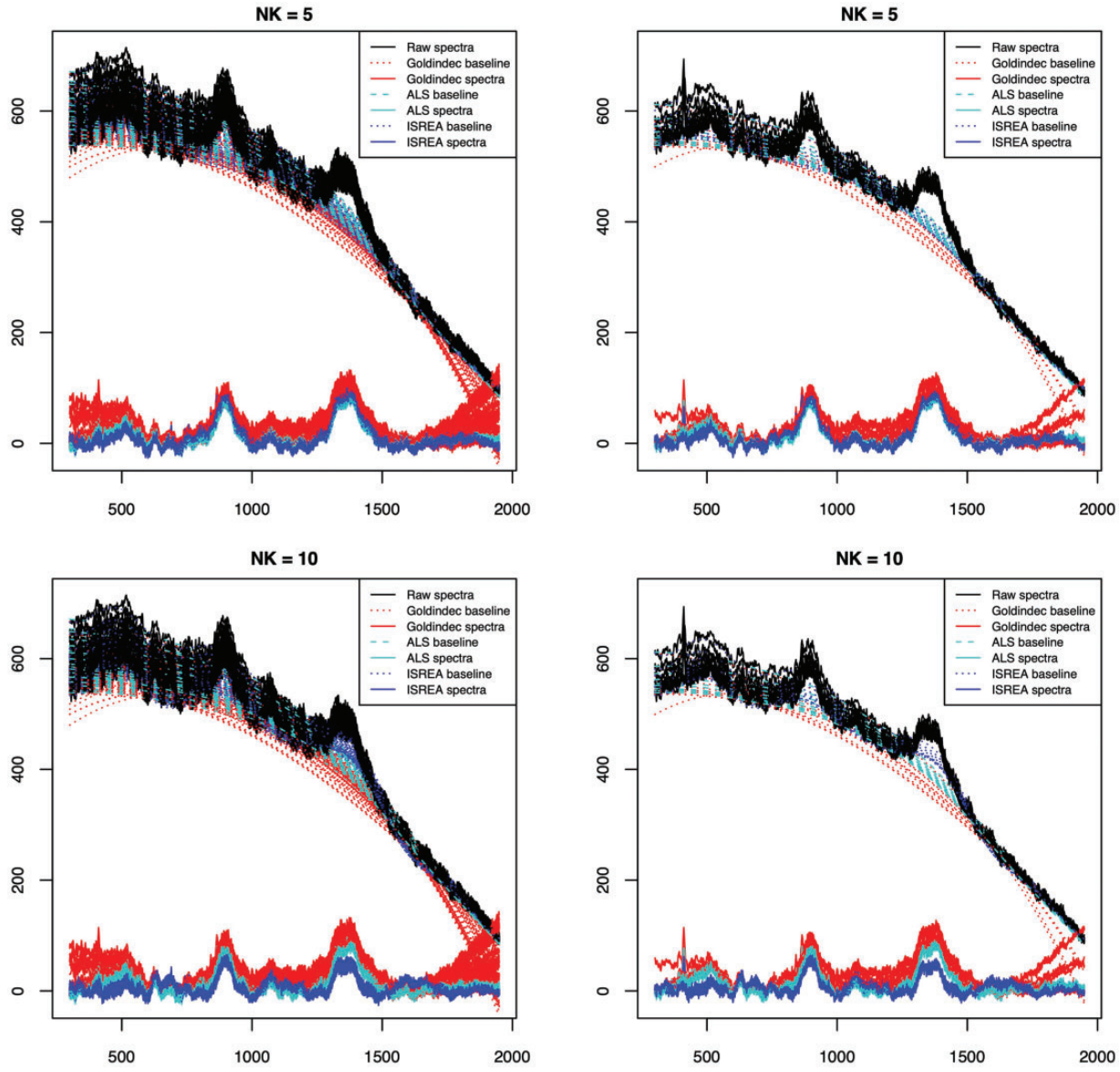


**Figure 4.** Comparison with ALS, Goldindec, and ISREA (with $NK = 5$ and $NK = 10$) on dialysate spectra at 10 min of the session. *NK* refers to number of knots selected. Right: 10 dialysate spectra. Left: All 50 dialysate spectra. Horizontal axis: Raman shift (in $cm^{-1}$); vertical axis: intensity. The curves are respectively: raw spectra (black solid), ALS baseline estimates (cyan dashed), ALS-corrected spectra (cyan solid), Goldindec baseline estimates (red dashed), Goldindec-corrected spectra (red solid), ISREA baseline estimates (blue dashed), and ISREA-corrected spectra (blue solid).

the ISREA baseline-corrected spectra with $NK = 5$, and the ISREA baseline-corrected spectra with $NK = 10$. We then calculated the similarity changes by subtracting each of these three sets of correlations for baseline-corrected spectra from the correlations for raw spectra. Hence, these correlation differences represented the similarity changes after each baseline correction procedure. The means and standard deviations of these correlation differences are summarized in Table V. It shows that Goldindec baseline correction dramatically decreased the similarities between raw spectra, whereas ALS and the ISREA baseline

**Table V.** Similarity change comparisons of ALS, Goldindec, and ISREA on dialysate spectra.

| | Mean | Standard deviation |
|---|---|---|
| ALS | 0.0507 | 0.0082 |
| Goldindec | 0.2559 | 0.2282 |
| ISREA | | |
| $NK = 5$ | 0.0446 | 0.0072 |
| $NK = 10$ | 0.1238 | 0.0206 |

ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

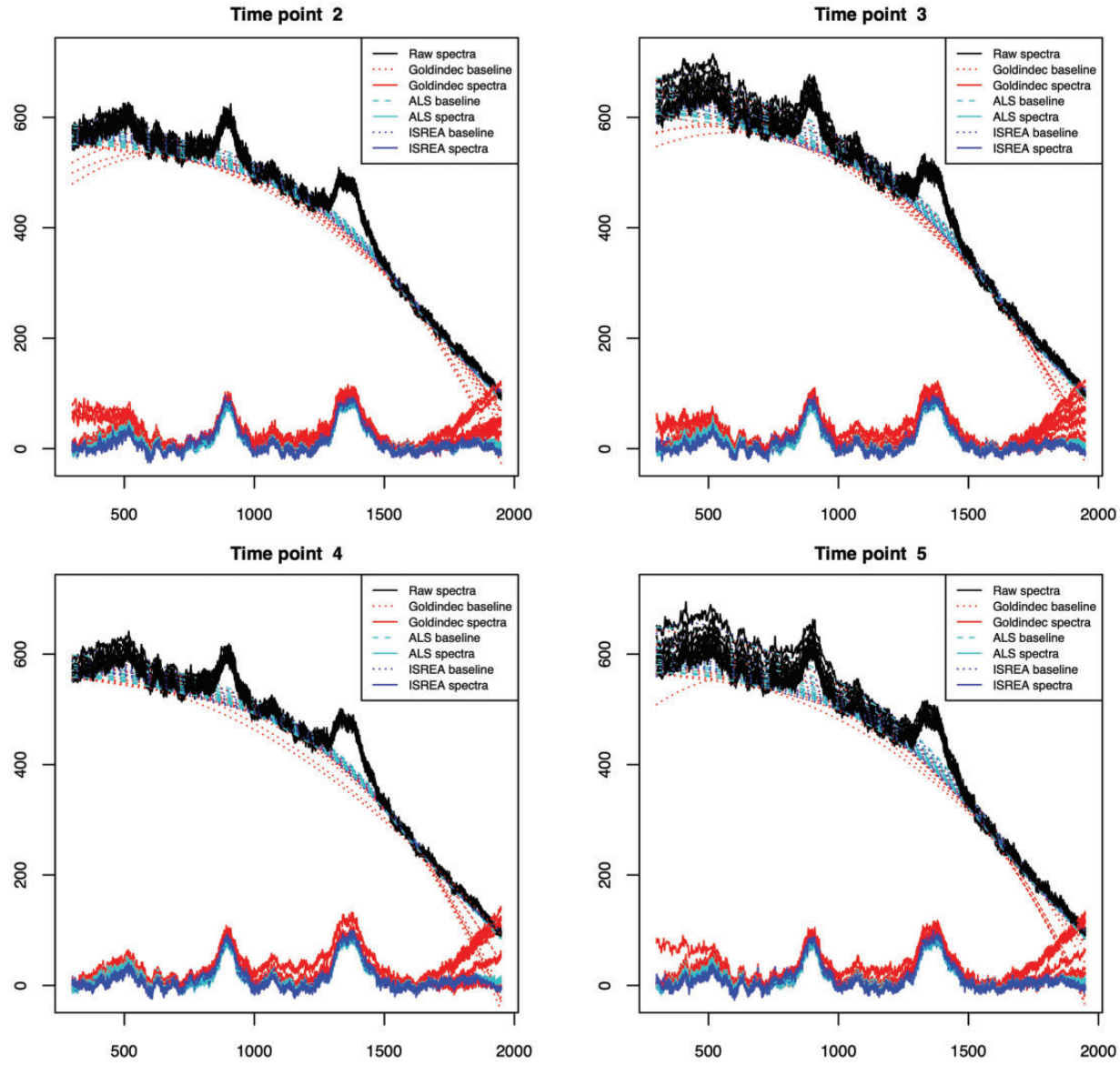


**Figure 5.** Comparison with ALS, Goldindec, and ISREA (with $NK = 5$) on dialysate spectra at different time points in one treatment session. *NK* refers to number of knots selected. Time points 2, 3, 4, 5 refer to 60, 120, 180, 240 min of the session. Horizontal axis: Raman shift (in $cm^{-1}$); vertical axis: intensity. The curves are respectively: raw spectra (black solid), ALS baseline estimates (cyan dashed), ALS-corrected spectra (cyan solid), Goldindec baseline estimates (red dashed), Goldindec-corrected spectra (red solid), ISREA baseline estimates (blue dashed), and ISREA-corrected spectra

correction procedure with $NK = 5$ or $NK = 10$ preserved the similarities between raw spectra much better. The mean similarity difference was the smallest for the ISREA with $NK = 5$ while the mean similarity difference for the ALS was only slightly larger.

To illustrate the trend over the treatment session, we also provide the raw and baseline-corrected spectra in Fig. 5 for dialysate samples collected at 60, 120, 180 and 240 minutes of the session. A more formal study of the trend would require new statistical analysis tools that are currently under development.

**Table VI.** Computational time (s) of ALS, Goldindec, and ISREA when processing batches of simulated spectra and dialysate spectra.

| | 1000 simulated spectra | 50 dialysis spectra |
|---|---|---|
| ALS | 3.16 | 0.16 |
| Goldindec | 885.36 | 38.54 |
| ISREA | 24.13 | 8.76 |

ALS: asymmetric least squares; ISREA: iterative smoothing-splines with root error adjustment.

### *Computational Cost Comparison*

Another advantage of ISREA is its relatively low computational cost, making it ideal for efficient batch processing of a large number of raw Raman spectra. Such computational efficiency in batch processing is critical for applications like real-time monitoring by Raman spectroscopy. In Table VI, we present the total computational times that ALS, Goldindec, and ISREA respectively take for 1000 simulated spectra and 50 spectra of dialysate samples. As the table shows, both ALS and ISREA are clearly efficient in computation and ideal for batch processing.

## Conclusion

Motivated from the success of asymmetric loss functions, ISREA is a simple and fast baseline correction procedure that mimics the threshold-free asymmetric square-root loss and well preserves all the meaningful Raman peaks. Its computational efficiency is a natural by-product deriving from the elimination of the threshold selection and non-convex optimization required in existing asymmetric loss methods. Our numerical experiments have demonstrated that ISREA has excellent and consistent performance on a wide variety of spectra, including mineral spectra that consist of sharp or low but sparse peaks and spectra of complicated compounds like used dialysate that contain many meaningful peaks over the whole spectral domain. Although we have implemented ISREA in R, it can be easily translated to other common languages like Python, Matlab, and others to facilitate automation and processing of large Raman spectral datasets.

Our comparisons of ISREA with the ALS and Goldindec procedures indicate that both ISREA and ALS can perform better baseline correction than the Goldindec. Against each other, ISREA and ALS each have slight advantages from different aspects.

As described in the Introduction, EMSC is another commonly used baseline correction procedure which produces spectra similar to the chosen reference spectrum. In our hemodialysis experiment, the baseline-corrected spectra generated at different time points will be used in group comparisons to determine whether there are any significant changes in the chemical compositions of waste dialysate samples collected at these time points. The comparison analysis, to be reported in a later manuscript, would require accurate representations of the Raman peaks of constituent chemicals of waste dialysate. The EMSC method with the mean spectrum as the reference spectrum at each time point would not yield spectra with such kind of representations. A reference spectrum that is already baseline corrected would be necessary. This would require a combination of the EMSC with another baseline correction method such as ALS or ISREA. The performance of such a combination certainly merits further research but is beyond the scope of this paper.

### Acknowledgments

The authors are grateful to the Associate Editor and two reviewers for their insightful comments that have significantly improved the paper.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The corresponding author's research has been supported by Virginia Tech College of Science Dean's Discovery Fund and U.S. National Science Foundation (DMS-1620945 and DMS-1916174).

### Supplemental material

All supplemental material mentioned in the text, consisting of cohort description for the hemodialysis study and numerical experiments on the sensitivity of $\varepsilon$, is available in the online version of the journal.

### ORCID iD

Pang Du https://orcid.org/0000-0003-1365-4831

### References

1. A. Kudelski. "Analytical Applications of Raman Spectroscopy". Talanta. 2008. 76(1): 1–8.
2. R.S. Das, Y.K. Agrawal. "Raman Spectroscopy: Recent Advancements, Techniques, and Applications". Vib. Spectrosc. 2011. 57(2): 163–176.
3. N. Wang, H. Cao, L. Wang, et al. "Recent Advances in Spontaneous Raman Spectroscopic Imaging: Instrumentation and Applications". Curr. Med. Chem. 2020. 26: 1.
4. J. Depciuch, E. Kaznowska, I. Zawlik, et al. "Application of Raman Spectroscopy and Infrared Spectroscopy in the Identification of Breast Cancer". Appl. Spectrosc. 2016. 70(2): 251–263.
5. A.K. Fisher, W.F. Carswell, A.I.M. Athamneh, et al. "The Rametrix LITE Toolbox V1.0 for MATLAB". J. Raman Spectrosc. 2018. 49(5): 885–896.
6. R.S. Senger, V. Kavuru, M. Sullivan, et al. "Spectral Characteristics of Urine Specimens from Healthy Human Volunteers Analyzed Using Raman Chemometric Urinalysis (Rametrix)". PLoS One. 2019. 14(9): e0222115.
7. R.S. Senger, J.L. Robertson. "The Rametrix PRO Toolbox V1.0 for Matlab". PeerJ. 2020. 8: e8179.
8. R.S. Senger, M. Sullivan, A. Gouldin, et al. "Spectral Characteristics of Urine from Patients with End-Stage Kidney Disease Analyzed Using Raman Chemometric Urinalysis (Rametrix)". PLoS One. 2020. 15(1): E0227281.
9. R. Gautam, S. Vanga, F. Ariese, et al. "Review of Multidimensional Data Processing Approaches for Raman and Infrared Spectroscopy". EPJ Technol. Instrum. 2015. 2(1): 8.
10. 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.
11. J. Zhao, H. Lui, D.I. McLean, et al. "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy". Appl. Spectrosc. 2007. 61(11): 1225–1232.

12. A. Mahadevan-Jansen, R.R. Richards-Kortum. "Raman Spectroscopy for the Detection of Cancers and Precancers". J. Biomed. Opt. 1996. 1(1): 31–70.
13. C.A. Lieber, A. Mahadevan-Jansen. "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra". Appl. Spectrosc. 2003. 57(11): 1363–1367.
14. C.M. Stellman, K.S. Booksh, M.L. Myrick. "Multivariate Raman Imaging of Simulated and 'Real World' Glass-Reinforced Composites". Appl. Spectrosc. 1996. 50(5): 552–557.
15. V. Shusterman, S.I. Shah, A. Beigel, et al. "Enhancing the Precision of ECG Baseline Correction: Selective Filtering and Removal of Residual Error". Comput. Biomed. Res. 2000. 33(2): 144–160.
16. G. Schulze, A. Jirasek, M.M. Yu, et al. "Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation". Appl. Spectrosc. 2005. 59(5): 545–574.
17. T.T. Cai, D. Zhang, D. Ben-Amotz. "Enhanced Chemical Classification of Raman Images Using Multiresolution Wavelet Transformation". Appl. Spectrosc. 2001. 55(9): 1124–1130.
18. D. Chen, Z. Chen, E. Grant. "Adaptive Wavelet Transform Suppresses Background and Noise for Quantitative Analysis by Raman Spectrometry". Anal. Bioanal. Chem. 2011. 400(2): 625–634.
19. J. Li, B. Yu, H. Fischer. "Wavelet Transform Based on the Optimal Wavelet Pairs for Tunable Diode Laser Absorption Spectroscopy Signal Processing". Appl. Spectrosc. 2015. 69(4): 496–506.
20. D. Zhang, D. Ben-Amotz. "Enhanced Chemical Classification of Raman Images in the Presence of Strong Fluorescence Interference". Appl. Spectrosc. 2000. 54(9): 1379–1383.
21. Y. Cai, C. Yang, D. Xu, et al. "Baseline Correction for Raman Spectra Using Penalized Spline Smoothing Based on Vector Transformation". Anal. Methods. 2018. 10(28): 3525–3533.
22. V. Mazet, C. Carteret, D. Brie, et al. "Background Removal from Spectra by Designing and Minimising a Non-Quadratic Cost Function". Chemom. Intell. Lab. Syst. 2005. 76(2): 121–133.
23. J.R. Beattie, J.J. McGarvey. "Estimation of Signal Backgrounds on Multivariate Loadings Improves Model Generation in Face of Complex Variation in Backgrounds and Constituents". J. Raman Spectrosc. 2013. 44(2): 329–338.
24. H. Martens, E. Stark. "Extended Multiplicative Signal Correction and Spectral Interference Subtraction: New Preprocessing Methods for Near Infrared Spectroscopy". J. Pharm. Biomed. Anal. 1991. 9(8): 625–635.
25. K. Liland, A. Kohler, N. Afseth. "Model-Based Pre-Processing in Raman Spectroscopy of Biological Samples". J. Raman Spectrosc. 2016. 47(6): 643–650.
26. M. Scholtes-Timmerman, H. Willemse-Erix, T.B. Schut, et al. "A Novel Approach to Correct Variations in Raman Spectra Due to Photo-Bleachable Cellular Components". Analyst. 2009. 134(2): 387–393.
27. P. Candeloro, E. Grande, R. Raimondo, et al. "Raman Database of Amino Acids Solutions: A Critical Study of Extended Multiplicative Signal Correction". Analyst. 2013. 138(24): 7331–7340.
28. N.K. Afseth, A. Kohler. "Extended Multiplicative Signal Correction in Vibrational Spectroscopy, A Tutorial". Chemom. Intell. Lab. Sys. 2012. 117: 92–99.
29. J. Liu, J. Sun, X. Huang, et al. "Goldindec: A Novel Algorithm for Raman Spectrum Baseline Correction". Appl. Spectrosc. 2015. 69(7): 834–842.
30. P.H.C. Eilers. "A Perfect Smoother". Anal. Chem. 2003. 75(14): 3631–3636.
31. K. Liland, T. Almøy, B. Mevik. "Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra". Appl. Spectrosc. 2010. 64(9): 1007–1016.
32. J. Peng, S. Peng, A. Jiang, et al. "Asymmetric Least Squares for Multiple Spectra Baseline Correction". Anal. Chim. Acta. 2010. 683(1): 63–68.
33. S. He, W. Zhang, L. Liu, et al. "Baseline Correction for Raman Spectra Using an Improved Asymmetric Least Squares Method". Anal. Methods. 2014. 6(12): 4402–4407.
34. H.F.M. Boelens, P.H.C. Eilers. "Baseline Correction with Asymmetric Least Squares Smoothing". Leiden University Medical Centre Report. 2005. 1(1): 5.
35. B. Lafuente, R.T. Downs, H. Yang, et al. "The Power of Databases: The RRUFF Project". In: T. Armbruster, R.M. Danisi, editors. Highlights in Mineralogical Crystallography. Berlin: W. De Gruyter, 2015. Chap. 1, 1–30.