Analysis of Multicomponent Ionic Mixtures using Blind Source Separation - a Processing Case Study

Giovanni Maria Maggioni, Stefani Kocevska, Martha A. Grover,* and Ronald W. Rousseau*

Georgia Institute of Technology

E-mail: martha.grover@chbe.gatech.edu; rwr@chbe.gatech.edu

August 13, 2019

Abstract

The Python code and the Raw Raman and IR spectra can be found at: `https://github.com/john88gm/BSS_Analysis-Spectroscopy`.

1 Supplementary Information

1.1 Spectroscopy

The electromagnetic radiation travelling through a liquid solution and/or impinging on a solid can interact with the units (atoms, molecules, ions) constituting the material.[1–3] In this contribution, we consider only two spectroscopic techniques among all those possible: ATR-FTIR and Raman. They are based on the interaction of light with a chemical bond: the former through the absorption of infra-red radiation because of a dipole moment; the latter

through scattering of light related to molecular polarizability. Symmetric chemical bonds and vibration modes have no dipole moment (they are thus IR inactive), while strongly polar molecules are usually weakly (if at all) polarizable (and are thus Raman inactive or only weakly active): ATR-FTIR and Raman spectroscopy are therefore often complementary tools. Additionally, Raman scattering is active with both liquid and solid species, while ATR-FTIR is used mainly to monitor liquid solutions, since the signal of solids is rather weak. The mathematical treatment developed in this work applies equally to Raman and ATR-FTIR spectroscopy.

The intensity of the radiation (the absorbed one in ATR-FTIR and the scattered one in Raman) is related to the concentration of the active species. When the concentration of the species of interest in a medium is low, the intensity, X, follows a linear dependence on the concentration, C, known as the *Beer-Lambert law*:

$$X(\lambda, \mathbf{q}) = Cl(\lambda, \mathbf{q}) \tag{1}$$

where l is the absorption coefficient, which is dependent on the wavenumber, λ (with units of cm^{-1}), and possibly on other intensive properties, such as temperature, T, and pH, i.e. $\mathbf{q} = (T, \mathrm{pH})$. When there are n_K spectroscopically active species, the total intensity (absorbance/scattering) is usually obtained according to a linear superposition principle:

$$X(\lambda, \mathbf{q}) = \sum_{k=1}^{n_K} C_k L_k(\lambda, \mathbf{q}) \tag{2}$$

where L_k represents the intensity that the species k would have if it were the only one present in the medium. Note that Eqs. (1) and (2), resting on the assumption of linearity, are also valid in several systems at moderate, rather than low, concentrations. In actual experiments, one samples only a finite set of values of the wavenumbers, hence the spectrum is discretized into n_L values. When n_N different spectra are collected, Eq. (2) is written in matrix form

as a linear system:

$$\mathbf{X} = \mathbf{CL} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{n_N \times n_L}$, $\mathbf{C} \in \mathbb{R}^{n_N \times n_K}$, and $\mathbf{L} \in \mathbb{R}^{n_K \times n_L}$; the dependence on $\mathbf{q}$ is dropped to ease the notation. Since Eq. (2) is a linear problem, given the data $\mathbf{X}$ and if $\mathbf{L}$ (or $\mathbf{C}$) were known, one could compute $\mathbf{C}$ (or $\mathbf{L}$) by (pseudo-)inversion:

$$\mathbf{C} = \mathbf{X}\mathbf{L}^{-1} \tag{4}$$

$$\mathbf{L} = \mathbf{C}^{-1}\mathbf{X} \tag{5}$$

1.2 Standard Calibration

In a standard calibration approach, one uses various forms of supervised learning such as Partial Least Squares (PLS) or Principal Component Analysis (PCA).[3–5] Under the assumption that $n_N \gg n_K$, the sought approximate solution is formally written as:

$$\mathbf{C} = \mathcal{F}(\mathbf{X}) \approx \mathbf{X}\hat{\mathbf{L}}^{-1} \tag{6}$$

where $\mathcal{F}$ is the (non-linear) model used to correlate the input measurements $\mathbf{X}$ with concentration by means of calibration experiments in which the number of species, their identity, and their concentrations are known (i.e. one solves first for $\hat{\mathbf{L}} = \mathbf{C}_c^{-1}\mathbf{X}_c$, where the subscript c indicates the matrix of calibration concentration, then for any subsequent measurement the concentration is estimated as per Eq. (6)).

1.3 Data Preprocessing

Measured spectra are corrupted by several unwanted phenomena, such as baseline drift, spikes, and — more generally — noise that prevent a direct application of Eqs. (1) to (5) of the manuscript to the raw data. Preprocessing aims at removing these unwanted phenomena

from the spectra, before feeding them to the algorithms that calibrate and estimate the concentrations and/or the spectra of pure species.

Baseline correction is commonly performed on raw data of most spectroscopic techniques and several algorithms have been developed for that purpose. In this work, we have chosen two procedures, developed by Mazet *et al.*[6] and Zhang *et al.*[7] and based on a polynomial fit of the baseline: the former, though, uses nonquadratic cost functions adapted to the processed spectrum, while the latter uses reweighted penalized least squares (PLS). However, in this study we have found that both algorithms yield almost identical spectra after baseline correction, with no apparent differences in terms of performance and robustness. We have then adopted the PLS procedure.

Smoothing is also usually performed on raw data. Here, we have used the Savitzky-Golay filter,[8] because of its robustness and efficiency and since it can also perform numerical differentiation of the spectra with respect to the wavenumber (see Section 2 in the manuscript). The Savitzky-Golay filter requires two input parameters, the window size, S_w (i.e. the number of points to be used for estimating the smoothed value), and the degree of the locally interpolating polynomial, S_p.

Despiking is standard in pre-processing procedure particularly for Raman spectra and consists of the removal of spikes, i.e. sudden surges in the spectral intensity due to random cosmic rays. Several despiking algorithms have been proposed: in this work, we have adopted the simple, but rather effective approach recently proposed by Whitaker and Hayes[9] to identify a spike. Note that despiking may result in only a reduction, rather than a total removal, of spikes, particularly if two consecutive spikes are affecting the measurements.

Despiking aims at automatically detecting and removing the spikes, which could be otherwise interpreted during the data analysis as spectroscopic features of some species, leading to outlier values and spurious behaviors. Supposing that the spectra are collected in a sequence, this approach identifies a spike by using modified Z-scores, defined for each measured

spectrum as:

$$Z_{il} = 0.6745\frac{\Delta X_{il} - \langle \Delta X \rangle}{\mathcal{V}} \qquad \forall l = 1, ..., n_L, \forall i = 2, ..., n_N \tag{7}$$

where $\Delta X_{il} = X_{il} - X_{(i-1)l}$ is the difference between two consecutive measurements, $\langle \Delta X \rangle$ is the median of all ΔX, and $\mathcal{V}$ is the median of $|\Delta X_{il} - \langle \Delta X \rangle|$. Whitaker and Hayes suggest to identify as spike values those X_{il} for which $Z_{il} > 6$. As for spike removal, we adopted the method proposed by Li and Dai,[10] i.e. to compute the values replacing those of the spike by linear extrapolation from the immediate predecessors, i.e. extrapolating the values from the spectrum $\mathbf{X}_{i-1}$ to the spectrum $\mathbf{X}_i$. Figure 1 compares the raw spectrum (dashed blue line) with its despiked and baselined counterpart (solid black line).

Finally, pre-processing often includes spectra centering and rescaling. The former shifts the data, so that each spectrum has zero mean; the latter is used to decrease the relative difference between the importance of samples at high and at low concentration, thus diminishing the formation and propagation of spurious effects and numerical instabilities. It is worth mentioning that the use of so-called "internal standards", i.e. dividing the whole spectrum by its value at a specific wavenumber, is also frequently applied during calibration in Raman spectroscopy.[11–14] The use of internal standards, which can be considered a particular type of scaling, was based in this work on the peak of water at 1640 cm^{-1} for Raman spectra.

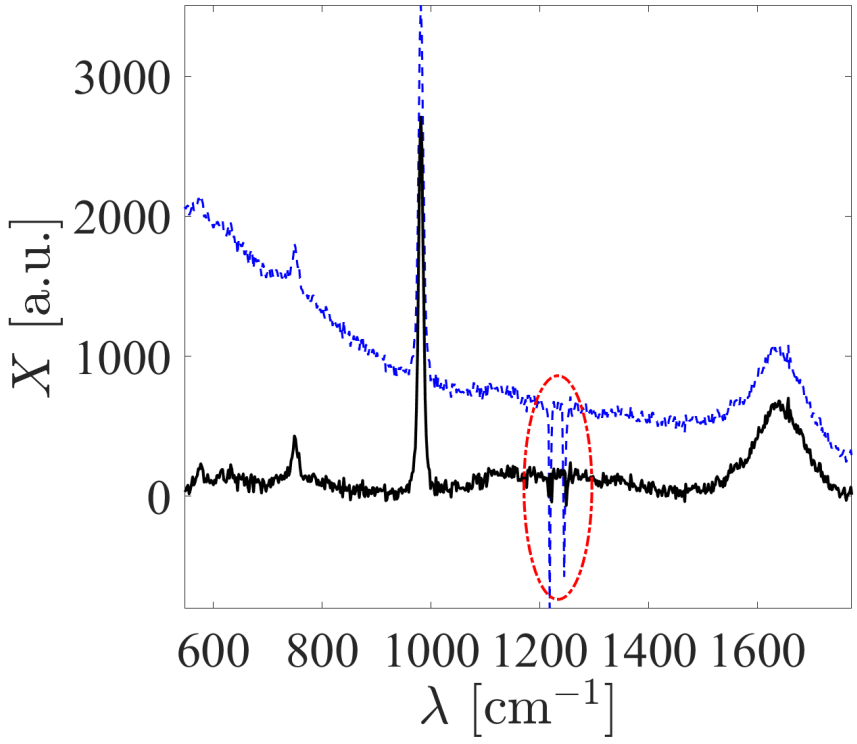

Figure 1: An example of a Raman spectrum with five different chemical species (water, sodium nitrite, sodium nitrate, sodium carbonate, and sodium aluminate) before (blue dashed line) and after despiking and baselining (black solid line). The attenuated, but not completely removed spike is highlighted by a red ellipse.

1.4 Simulated Data

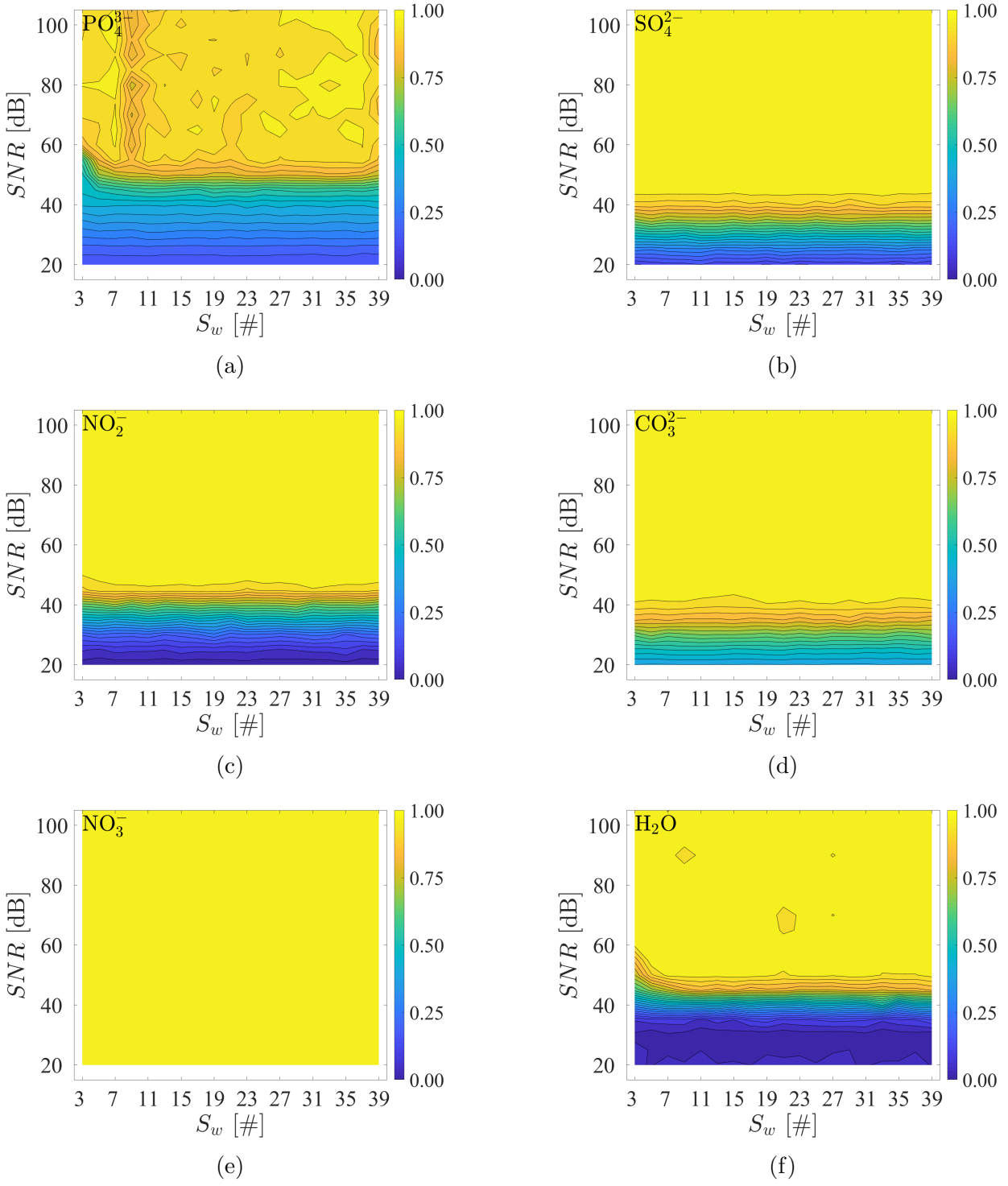

Figure 2: The value of the correlation coefficients between the spectra obtained from ICA/MCR-ALS and the reference spectra for the true components, for water and the five anions of interest, at different levels of noise and for different values of S_w for the Savitzky-Golay filter. The composition of the synthetic mixture follows the values reported in Table 1, with $\kappa = 0.10$. The bright yellow areas indicate the regions where the correlation is highest.

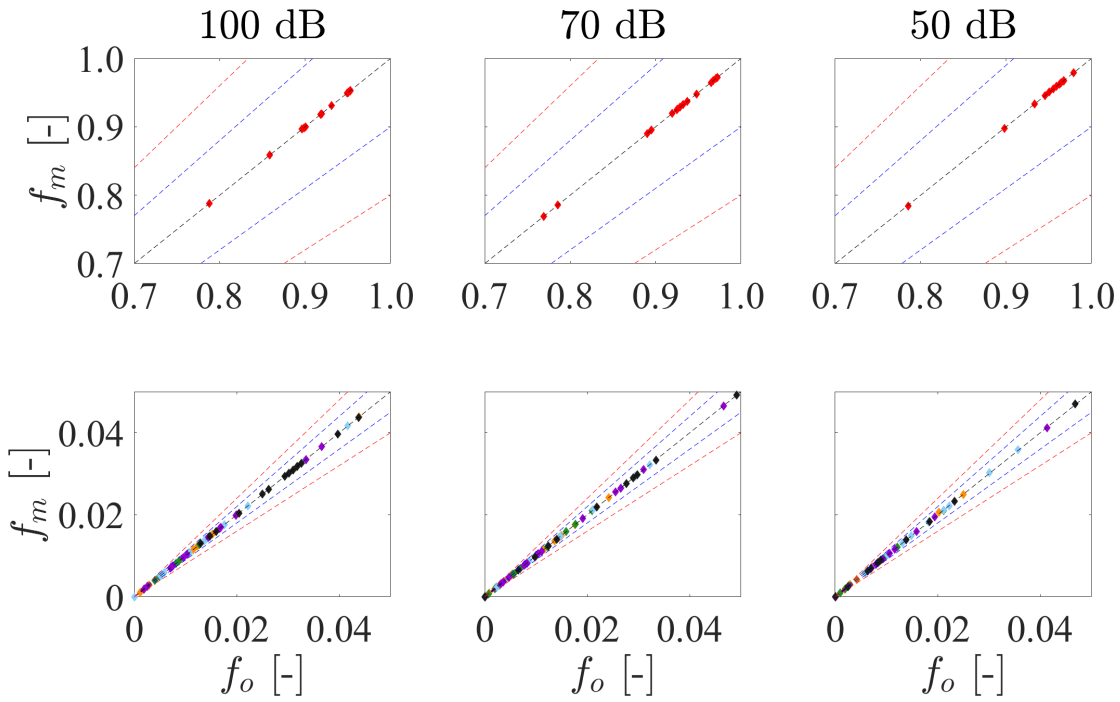

Figure 3: Estimation of the composition for a set of 15 mixtures, with $\kappa = 0.70$, with $SNR =$ 100, 70, 50 dB, from left to right; the values have been computed enforcing the spectra non-negativity and choosing a Savitsky-Golay window of 11 for all values of SNR. The black, blue, and red dashed lines in each plot indicate a perfect match, the $\pm 10\%$ boundaries, and the $\pm 20\%$ boundaries, respectively. H_2O, NO_3^-, NO_2^-, CO_3^{2-}, SO_4^{2-}, PO_4^{3-} are reported as red, black, light blue, violet, orange, and green symbols, respectively.

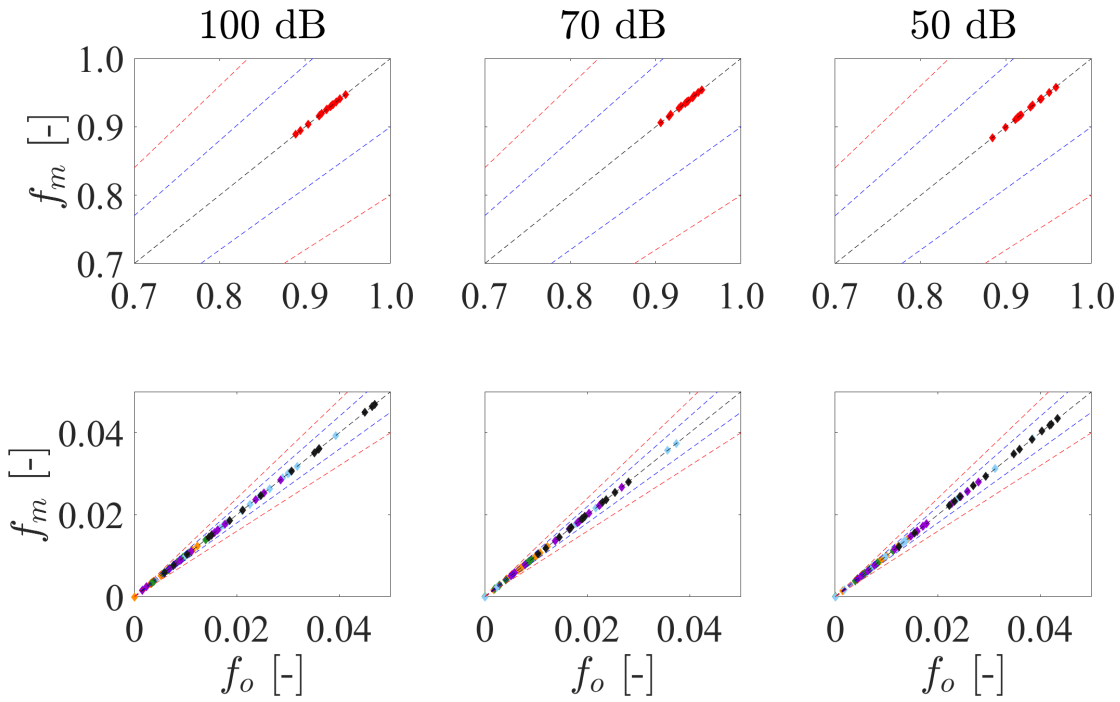

Figure 4: Estimation of the composition for a set of 15 mixtures, with $\kappa = 0.50$, with $SNR =$ 100, 70, 50 dB, from left to right; the values have been computed enforcing the spectra non-negativity and choosing a Savitsky-Golay window of 11 for all values of SNR. The black, blue, and red dashed lines in each plot indicate a perfect match, the $\pm 10\%$ boundaries, and the $\pm 20\%$ boundaries, respectively. H_2O, NO_3^-, NO_2^-, CO_3^{2-}, SO_4^{2-}, PO_4^{3-} are reported as red, black, light blue, violet, orange, and green symbols, respectively.

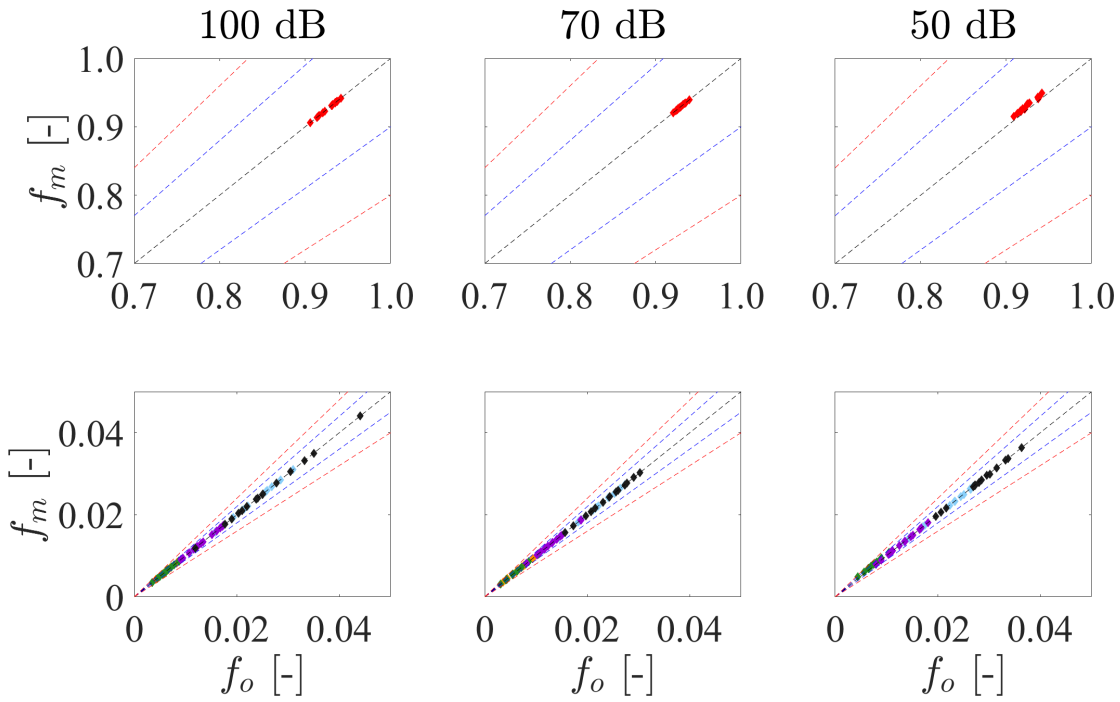

Figure 5: Estimation of the composition for a set of 15 mixtures, with $\kappa = 0.25$, with $SNR =$ 100, 70, 50 dB, from left to right; the values have been computed enforcing the spectra non-negativity and choosing a Savitsky-Golay window of 11 for all values of SNR. The black, blue, and red dashed lines in each plot indicate a perfect match, the $\pm 10\%$ boundaries, and the $\pm 20\%$ boundaries, respectively. H_2O, NO_3^-, NO_2^-, CO_3^{2-}, SO_4^{2-}, PO_4^{3-} are reported as red, black, light blue, violet, orange, and green symbols, respectively.

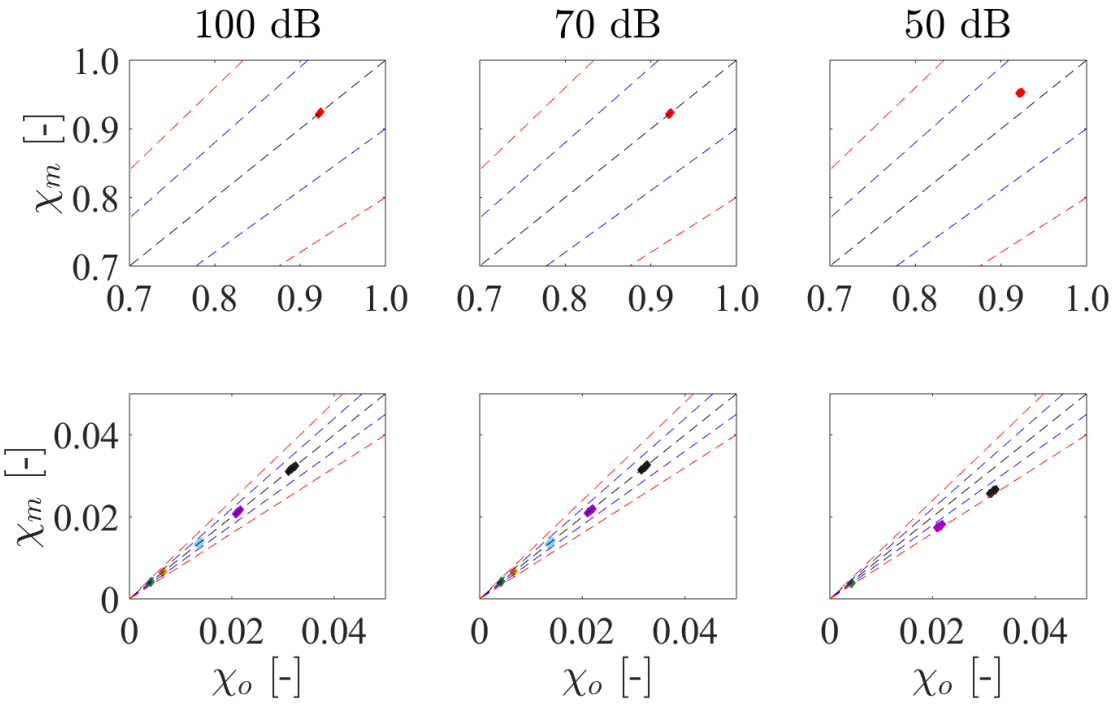

Figure 6: Estimation of the composition for a set of 15 mixtures, with $\kappa = 0.01$, with $SNR =$ 100, 70, 50 dB, from left to right; the values have been computed enforcing the spectra non-negativity and choosing a Savitsky-Golay window of 11 for all values of SNR. The black, blue, and red dashed lines in each plot indicate a perfect match, the $\pm 10\%$ boundaries, and the $\pm 20\%$ boundaries, respectively. H_2O, NO_3^-, NO_2^-, CO_3^{2-}, SO_4^{2-}, PO_4^{3-} are reported as red, black, light blue, violet, orange, and green symbols, respectively.

1.5 Experimental Values

Table 1: Compositions used for the experimental measurements with Raman and IR.

species/ sample	Na_3PO_4	Na_2SO_4	$NaNO_2$	Na_2CO_3	Na_NO_3	H_2O
1	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
2	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
3	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
4	0.00%	0.16%	0.00%	0.00%	0.00%	99.84%
5	0.24%	0.15%	0.00%	0.00%	0.00%	99.60%
6	0.24%	0.15%	0.00%	0.64%	0.00%	98.97%
7	0.24%	0.15%	0.00%	0.63%	1.20%	97.78%
8	0.23%	0.15%	1.00%	0.63%	1.19%	96.81%
9	0.23%	0.29%	0.97%	0.61%	1.15%	96.66%
10	0.46%	0.30%	0.99%	0.62%	1.18%	96.43%
11	0.45%	0.29%	0.97%	1.21%	1.15%	93.81%
12	0.46%	0.30%	0.98%	1.22%	2.32%	94.73%
13	0.45%	0.29%	1.94%	1.21%	2.30%	93.81%
14	0.45%	0.44%	1.94%	1.21%	2.29%	93.67%
15	0.68%	0.44%	1.93%	1.20%	2.29%	93.45%
16	0.68%	0.44%	1.92%	1.79%	2.27%	92.90%
17	0.67%	0.44%	1.90%	1.77%	3.38%	91.85%
18	0.67%	0.43%	2.83%	1.75%	3.35%	90.97%

1.6 Error Analysis

Table 2: Compositions estimated from the one-point calibration from Raman spectra.

species/ samples						
1	0.00%	0.00%	0.00%	0.00%	0.00%	100.01%
2	0.00%	0.00%	0.00%	0.00%	0.00%	100.05%
3	0.00%	0.00%	0.00%	0.00%	0.00%	100.03%
4	0.00%	0.17%	0.00%	0.00%	0.00%	99.88%
5	0.12%	0.18%	0.00%	0.00%	0.00%	99.72%
6	0.10%	0.17%	0.00%	0.65%	0.00%	99.10%
7	0.07%	0.14%	0.00%	0.56%	1.30%	97.93%
8	0.04%	0.16%	0.88%	0.67%	1.47%	96.77%
9	0.02%	0.33%	0.83%	0.69%	1.42%	96.70%
10	0.25%	0.35%	0.89%	1.43%	1.50%	95.59%
11	0.26%	0.36%	0.90%	1.46%	1.54%	95.48%
12	0.24%	0.33%	0.87%	1.55%	2.85%	94.14%
13	0.22%	0.32%	1.77%	1.55%	2.72%	93.42%
14	0.20%	0.52%	1.81%	1.67%	2.83%	92.97%
15	0.48%	0.51%	1.78%	1.65%	2.78%	92.80%
16	0.47%	0.53%	1.88%	2.46%	2.90%	91.76%
17	0.40%	0.48%	1.76%	2.60%	3.92%	90.83%
18	0.42%	0.48%	2.70%	2.73%	3.89%	89.78%

Table 3: Compositions estimated from the one-point calibration from IR spectra.

species/ samples						
1	0.02%	0.00%	0.00%	0.07%	0.00%	99.91%
2	0.02%	0.00%	0.00%	0.07%	0.00%	99.91%
3	0.02%	0.00%	0.00%	0.07%	0.00%	99.91%
4	0.03%	0.20%	0.00%	0.06%	0.00%	99.71%
5	0.31%	0.18%	0.00%	0.06%	0.00%	99.45%
6	0.26%	0.15%	0.00%	0.83%	0.00%	98.76%
7	0.22%	0.12%	0.00%	0.69%	1.23%	97.73%
8	0.23%	0.14%	0.94%	0.75%	1.13%	96.81%
9	0.25%	0.33%	0.93%	0.79%	1.13%	96.57%
10	0.54%	0.30%	0.85%	0.78%	1.12%	96.43%
11	0.49%	0.27%	0.82%	1.31%	1.15%	95.96%
12	0.50%	0.28%	0.89%	1.19%	2.32%	94.81%
13	0.49%	0.29%	1.98%	1.26%	2.15%	93.83%
14	0.53%	0.47%	1.95%	1.31%	2.13%	93.60%
15	0.78%	0.43%	1.87%	1.31%	2.12%	93.49%
16	0.77%	0.42%	1.88%	1.72%	2.21%	93.01%
17	0.77%	0.42%	1.95%	1.62%	3.25%	91.99%
18	0.76%	0.42%	2.98%	1.71%	3.03%	91.10%

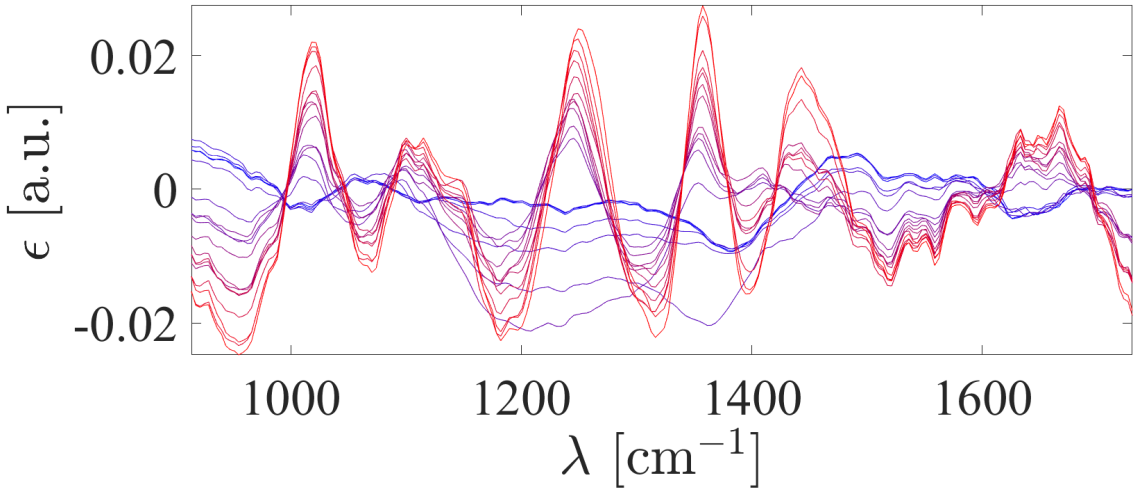

Figure 7: The error between the measured and the reconstructed IR spectra, computed as an element-wise difference.

1.7 Incomplete Library

The BSS part of the algorithm does not depend on the library, hence the spectra reconstructed by the BSS procedure are not affected by any missing component. Figure 8 shows the case discussed in the main text, where the reference spectrum of carbonate is missing from the library. In this figure, the left column reports the spectra reconstructed using BSS (top), CLS (center), and LASSO (bottom); the right column shows the residual errors calculated as discussed in Section 3.3.2 for each method.

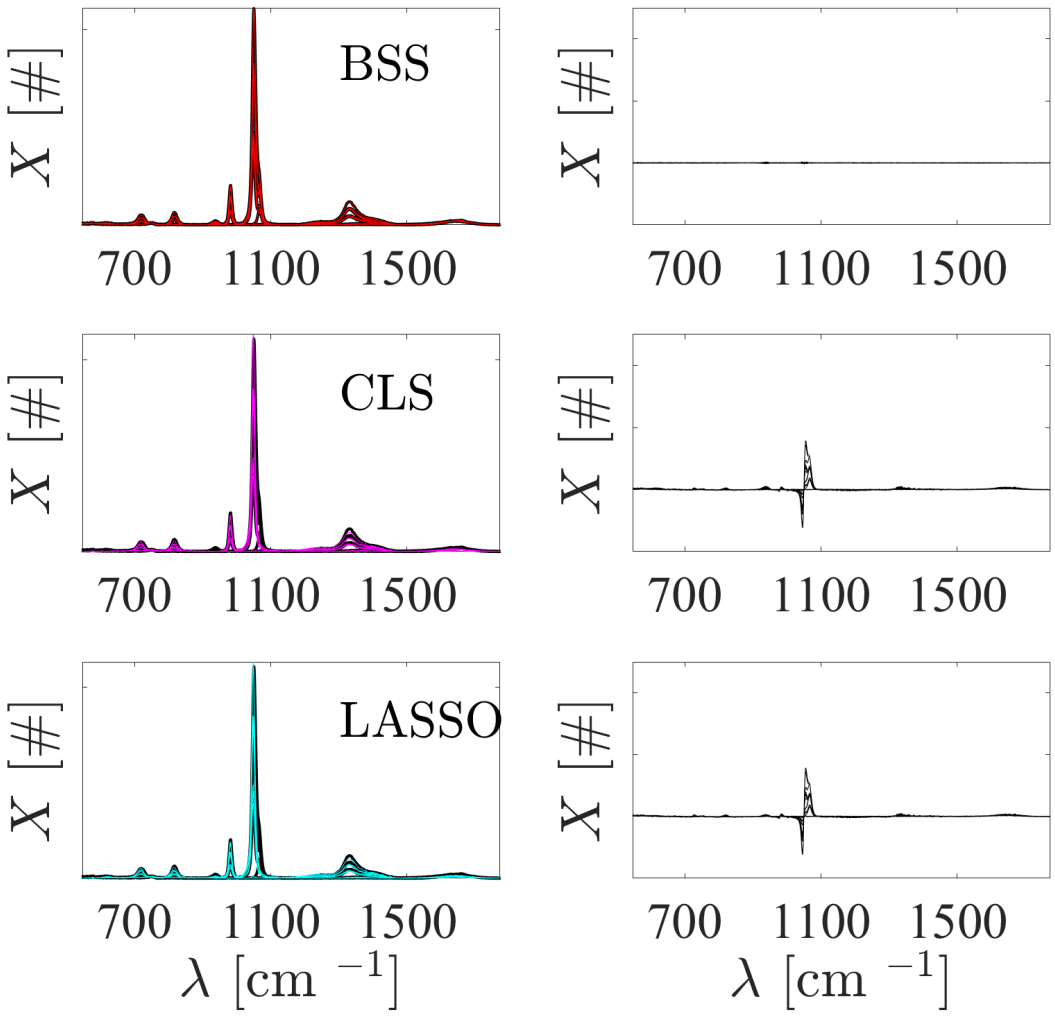

Figure 8: On the left column, we have reported the measured Raman spectra for the data set (black) and the reconstructed spectra using BSS, CLA, LASSO, in red, magenta, and light blue (from top to bottom). On the right column, the corresponding element-wise error: the BSS residual is basically background noise, whereas CLS and LASSO capture neither the peak drift of nitrate nor the peak of carbonate.

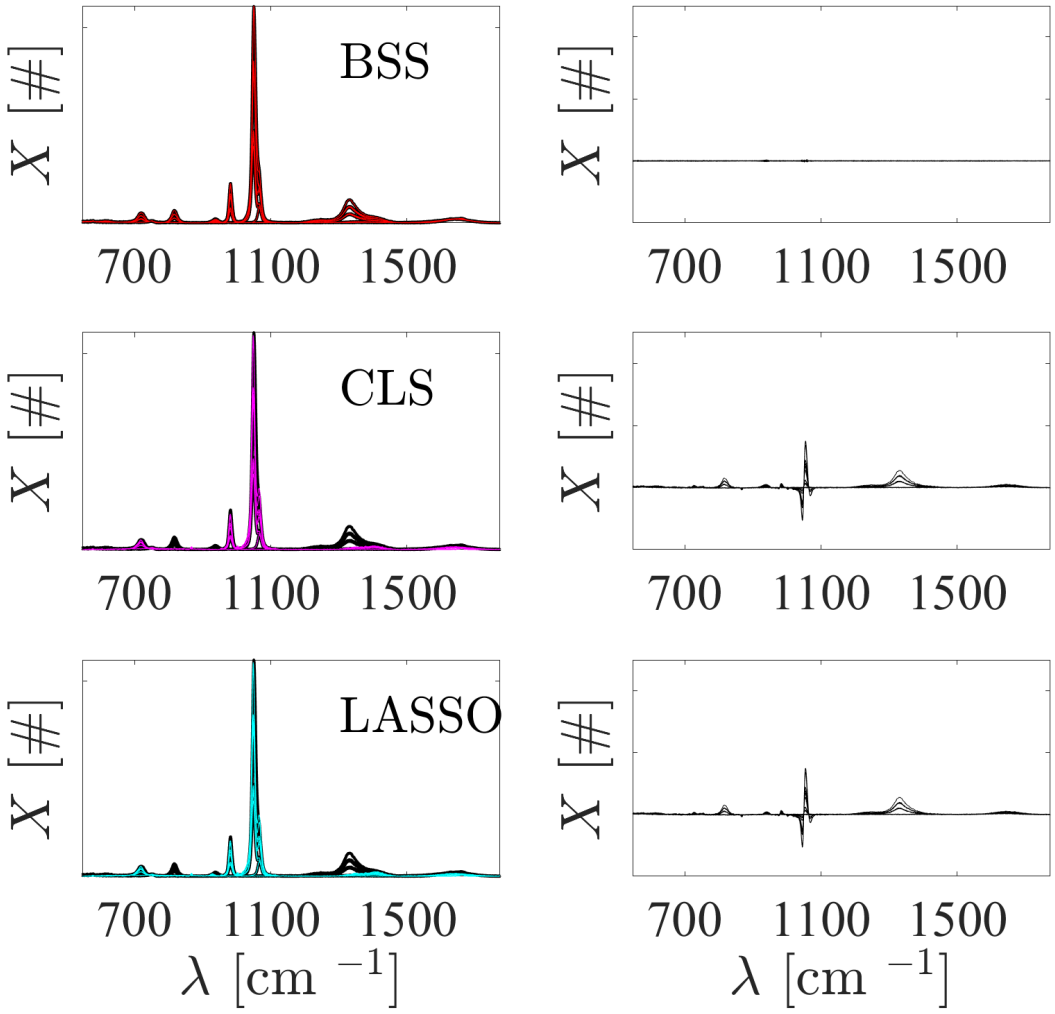

Figure 9: Library without nitrate. On the left column, we have reported the measured Raman spectra for the data set (black) and the reconstructed spectra using BSS, CLA, LASSO, in red, magenta, and light blue (from top to bottom). On the right column, the corresponding element-wise error.

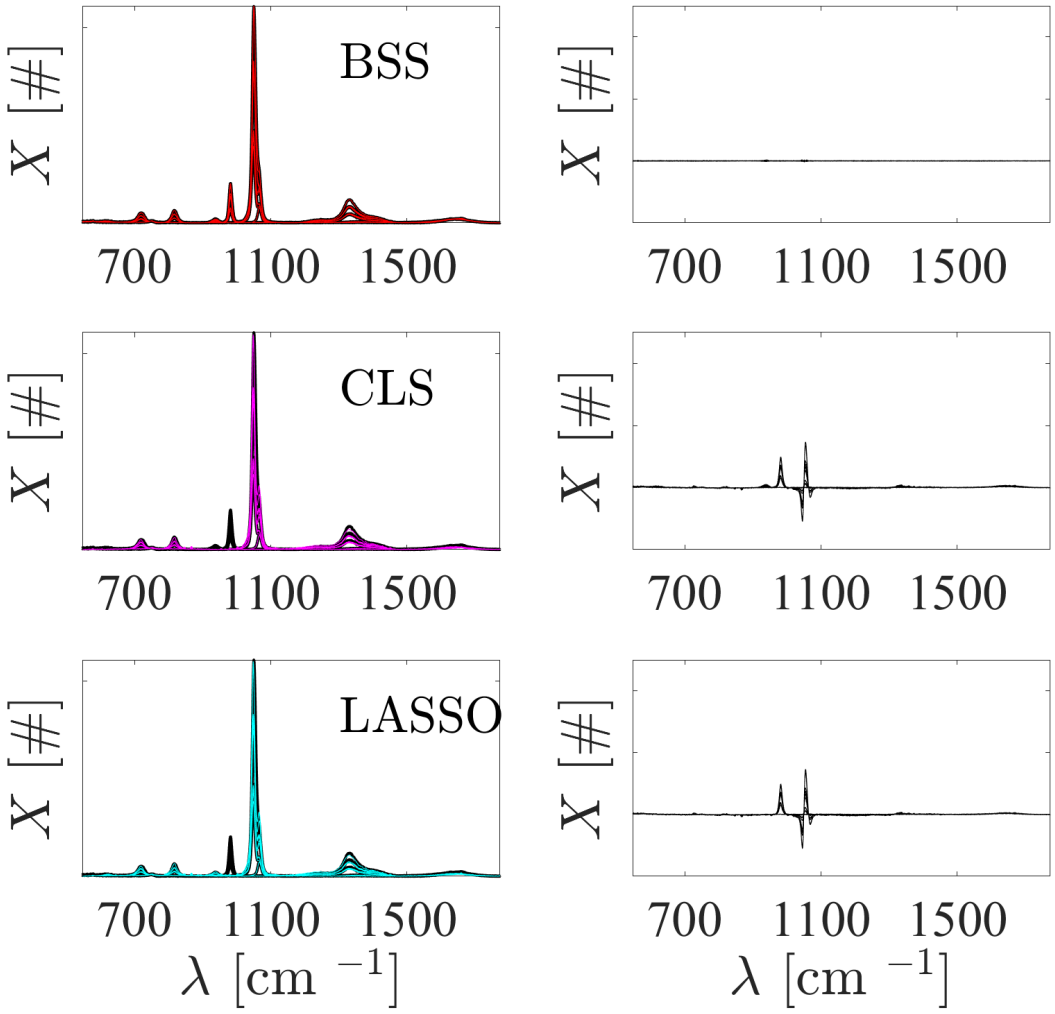

Figure 10: Library without sulfate. On the left column, we have reported the measured Raman spectra for the data set (black) and the reconstructed spectra using BSS, CLA, LASSO, in red, magenta, and light blue (from top to bottom). On the right column, the corresponding element-wise error.

References

(1) Adar, F.; Geiger, R.; Noonan, J. Raman spectroscopy for process/quality control. *Appl. Spectrosc. Rev.* **1997**, *32*, 45–101.

(2) Dunuwila, D. D.; Berglund, K. A. ATR FTIR spectroscopy for in situ measurement of supersaturation. *J. Cryst. Growth* **1997**, *179*, 185–193.

(3) Cornel, J.; Lindenberg, C.; Mazzotti, M. Quantitative application of in situ ATR-FTIR and Raman spectroscopy in crystallization processes. *Ind. Eng. Chem. Res.* **2008**, *47*, 4870–4882.

(4) Siesler, H. W.; Ozaki, Y.; Kawata, S. *Wiley –VCH*; 2002.

(5) Griffin, D. J.; Grover, M. A.; Kawajiri, Y.; Rousseau, R. W. Robust multicomponent IR-to-concentration model regression. *Chem. Eng. Sci.* **2014**, *116*, 77–90.

(6) Mazet, V.; Carteret, C.; Brie, D.; Idier, J.; Humbert, B. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 121–133.

(7) Zhang, Z.-M. M.; Chen, S.; Liang, Y.-Z. Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **2010**, *135*, 1138–1146.

(8) Savitzky, A.; Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.

(9) Whitaker, D. A.; Hayes, K. A simple algorithm for despiking Raman spectra. *Chemom. Intell. Lab. Syst.* **2018**, *179*, 82–84.

(10) Li, S.; Dai, L. An improved algorithm to remove cosmic spikes in Raman spectra for online monitoring. *Appl. Spectrosc.* **2011**, *65*, 1300–1306.

(11) Mosier-Boss, P. A.; Lieberman, S. H. Detection of nitrate and sulfate anions by normal Raman spectroscopy and SERS of cationic-coated, silver substrates. *Appl. Spectrosc.* **2000**, *54*, 1126–1135.

(12) Ianoul, A.; Coleman, T.; Asher, S. A. UV resonance Raman spectroscopic detection of nitrate and nitrite in wastewater treatment processes. *Anal. Chem.* **2002**, *74*, 1458–1461.

(13) Pelletier, M. J. Quantitative analysis using Raman spectrometry. *Appl. Spectrosc.* **2003**, *57*, 20A–42A.

(14) Sun, Q.; Qin, C. Raman OH stretching band of water as an internal standard to determine carbonate concentrations. *Chem. Geol.* **2011**, *283*, 274–278.