An Automated Peak Group Analysis for Vibrational Spectra Analysis

Mathias Sawall[a], Christoph Kubis[b], Benedict Leidecker[b], Lukas Prestin[a], Tomass Andersons[a], Martina Beese[a,b], Jan Hellwig[a,b], Robert Franke[c,d], Armin Börner[b], Klaus Neymeyr[a,b]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstraße 69, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse, Albert-Einstein-Straße 29a, 18059 Rostock*
[c] *Evonik Oxeno GmbH & Co. KG, Paul-Baumann Straße 1, 45772 Marl, Germany*
[d] *Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany*

Abstract

Peak Group Analysis (PGA) is a multivariate curve resolution technique that attempts to extract pure component spectra from time series of spectral mixture data. PGA requires that the mixture spectra consist of relatively sharp peaks, as is typical in IR and Raman spectroscopy. PGA aims to construct from individual peaks the associated pure component spectra in the form of nonnegative linear combinations of the right singular vectors of the spectral data matrix. In this paper, we propose an automated algorithm for peak detection, subsequent pure component spectra extraction and their correlation analysis. The algorithm is applied to FT-IR data sets on various rhodium carbonyl complexes and from an equilibrium of iridium complexes.

Key words: multivariate curve resolution, peak group analysis, peak detection

1. Introduction

Multivariate curve resolution (MCR) methods are used to extract pure component information from spectral mixture data sets, see for example [16, 2, 20, 17, 23]. Let the mixture spectra be stored row-wise in the matrix $D \in \mathbb{R}^{k\times n}$ with k being the number of spectra and n the number of frequency channels. Assuming that the spectral mixture data can be modeled by the bilinear Lambert-Beer law, it holds that

$$D = CS^T + E$$

with C and S being the matrices of the concentration profiles and the spectra of the pure components. The matrix E collects measurement errors, noise and deviations terms from the bilinear model. The matrix elements of E are assumed to be close to zero.

Depending on the spectroscopic techniques and on the chemical (reaction) system, a wide range of MCR methods is available for the data analysis. Some examples are:

- methods that attempt to analyze the overall data structure as the evolving factor analysis [9], the principal component analysis [3] or subspace angle analysis [21],
- techniques that impose the pure components to satisfy a particular hard model, such as a kinetic model [5, 17],
- soft-model-based pure component recovery techniques [13, 22],
- methods for the extraction of a subset of the pure component profiles [30, 25]
- and methods that make the full range of feasible solutions (C, S) available, namely techniques to compute factor ambiguity representations [10, 26].

1.1. Recovery of Pure Component Spectra

This work focuses on the extraction of pure component spectra from (time) series of mixture spectra where all spectra consist of a moderate number of partially overlapping relatively sharp peaks. A typical field of application is FT-IR spectroscopy in organometallic catalysis. A prominent algorithm for extracting such individual pure component spectra is band target entropy minimization (BTEM) by Garland and his coworkers. Many publications demonstrate the success of BTEM in the analysis of FT-IR spectra in organometallic catalysis, see for example [31, 30, 28]. Another algorithm with a similar goal is the Peak Group Analysis (PGA) [25, 27]. PGA is a computationally fast algorithm that aims to construct the associated pure component spectra from individual peaks. A major advantage

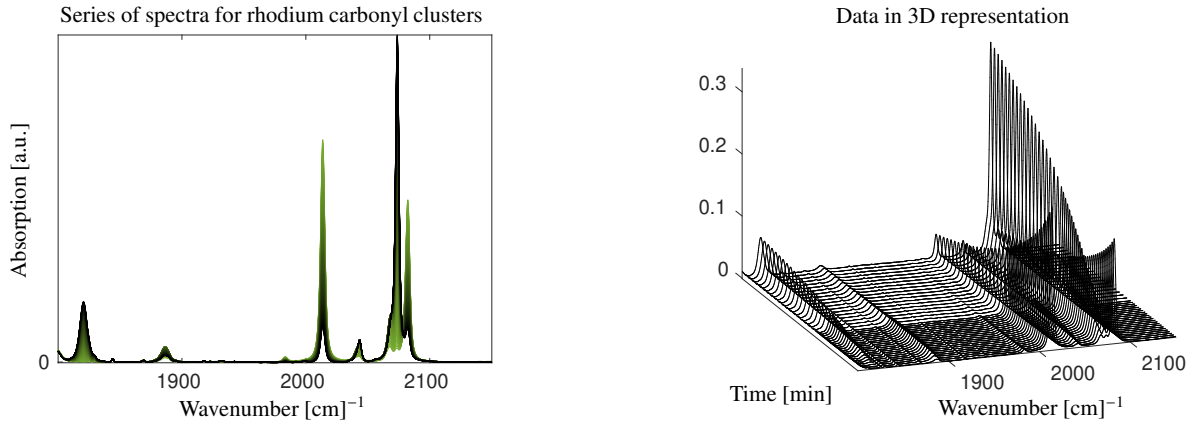

Figure 1: The FT-IR spectroscopic data set from a mixture of rhodium carbonyl complexes (Sec. 1.3). Left: Data in 2D with color grading from black for the first spectrum to green for the last spectrum. Right: The data set in 3D presentation.

of BTEM and PGA is that they are robust to data with a poor baseline or where no baseline correction has been performed beforehand. Typically, PGA allows the rapid detection of one or a few unknown chemical species.

PGA works with a singular value decomposition (SVD) of the spectral data matrix D, that is $D = U\Sigma V^T$. The columns of the orthogonal matrices U and V are the left and right singular vectors and the diagonal matrix Σ contains the singular values in decreasing order on its diagonal. It is a fundamental property of the SVD that the largest singular values and their associated singular vectors contain most of the structural information that is contained in the spectral data [16, 2, 20, 17]. PGA constructs the desired pure component spectra by linear combinations of the first few right singular vectors. In this construction process, the nonnegativity constraint is essential to form a component-wise nonnegative profile. Other constraints can be applied in order to enforce a certain degree of profile smoothness or sharp peaks, as can be expected in vibrational spectroscopy.

PGA in its original form [25] requires three manual inputs. First, the user must select a frequency window containing a peak. Second, the number of singular vectors for the profile expansion must be selected, and third, weighting factors for the various constraints are required. If classical PGA is used to compute all pure component spectra, then the algorithm must be manually applied to all major peaks.

1.2. Aim and Organization of the Paper

This work introduces an automated peak detection combined with an automated PGA, which includes a correlation analysis between the constructed spectra in order to find maximally uncorrelated profiles. These spectra may well be pure component profiles. We would also like to present the software implementation of PGA as a module included in the free software package FACPACK. The PGA module comes with a graphical user interface that requires a minimum of manual input and that suggests default parameter settings.

The paper is organized as follows. First, we present an experimental FT-IR data set of rhodium carbonyl complexes, which is used throughout the paper to demonstrate the capabilities of the PGA implementation. Sec. 2 briefly explains the PGA and its main algorithmic steps. Then, Sec. 3 introduces the automated PGA. Sections 4.1 and 4.2 present results of PGA analyses of FT-IR data taken from a mixture of rhodium carbonyl complexes and from an equilibrium of iridium complexes. Finally, Sec. 5 introduces the PGA graphical user interface that is a part of the FACPACK software.

1.3. Sample Data for Demonstrations

This series of FT-IR spectra was collected during the treatment of $Rh(acac)(CO)_2$ ([Rh] = 1 mM) with synthesis gas (20 bar CO/H_2) at 100°C. Under these conditions clustering occurs to form $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$. During the decomposition of $Rh(acac)(CO)_2$ acetylacetone is also formed, whose vibrational bands are non-absorbing in the spectral region examined. Fig. 1 shows the series of spectra in 2D and 3D presentations. The series contains $k = 208$ spectra each at $n = 1453$ wavenumbers in the spectral window of $[1800,\ 2150]\,\mathrm{cm}^{-1}$.

2. Peak Group Analysis

Next, we briefly review on the original form of PGA as presented in [25]. The first step is to find individual peaks in the spectral mixture data. For each such peak, PGA constructs a potential pure component spectrum that includes the previously selected peak. PGA requires that all peaks in the mixture spectra are relatively sharp with only

a moderate amount of mutual overlap. These conditions are typically met in FT-IR or Raman spectroscopy. The basis for the reconstruction of the pure component spectrum is the truncated SVD of D. The truncated SVD uses only a number z of dominant singular values of D and the associated left and right singular vectors. All other singular values and singular vectors are considered to have originated from noise and do not contain important chemical information. For a $k \times n$ matrix D, the truncated SVD reads

$$D \approx U\Sigma V^T = \underbrace{U\Sigma T^{-1}}_{C} \underbrace{TV^T}_{S^T} \tag{1}$$

with orthogonal $U \in \mathbb{R}^{k\times z}$, orthogonal $V \in \mathbb{R}^{n\times z}$ and a regular diagonal matrix $\Sigma \in \mathbb{R}^{z\times z}$ with the singular values on its diagonal. In addition, Eq. (1) shows how a regular $z \times z$ matrix T can be used to form potential pure component factors C and S with the bases of the dominant left and right singular vectors as contained in the columns of U and V. Each run of PGA can be interpreted as generating one column of T. Ideally, running PGA z times results in a complete matrix T, and thus the pure component factors C and S are determined. The number z is typically equal to the number of chemical species in the reaction system, but sometimes, depending on the level of noise, additional columns of V may contain relevant chemical information. A spectrum s is represented with respect to the basis of right singular values as

$$s = V\begin{pmatrix}1\\ t\end{pmatrix} \tag{2}$$

with $t \in \mathbb{R}^{z-1}$. (This representation of s gives the expansion coefficient of the first right singular vector the value 1, for which there is a well-known justification based on the Perron-Frobenius theory of nonnegative and irreducible matrices [26].) Further, let $[\nu_0, \nu_1]$ be a frequency interval around a given peak with the index interval $I = [i_0, i_1]$ of the frequency channels. The profile s is normalized to 1 over the interval I as

$$\tilde{s} = \frac{s}{\max\limits_{i\in I} |s_i|}. \tag{3}$$

The vector t of expansion coefficients is determined subject to some constraints as given in terms of penalty functions. The different penalty functions are combined in a weighted sum to form an objective function, which is then minimized numerically. The most important penalty function drives the numerical optimization towards a nonnegative spectral profile

$$g_1(\tilde{s}) = \sum_{i=1}^{n} \min\left(\frac{\tilde{s}_i}{\max \tilde{s}} + \varepsilon, 0\right)$$

with a small control parameter $\varepsilon \geq 0$. The function

$$g_2(\tilde{s}) = \sum_{j=1}^{k} \sum_{i\in I} (D_{ji} - c_j \tilde{s}_i)^2.$$

works with the local reconstruction of the concentration profile c according to

$$c = U\Sigma v^* \quad \text{with} \quad v^* = \frac{V(I,:)^T\, s(I)^T}{\|s(I)\|_2^2}$$

(see [25] for details) and measures the deviation of the selected peak from its local reconstruction. The penalty function g_2 is less important than the essential nonnegativity represented by g_1. The following two functions favor sharp peaks and smooth profiles as expected for pure component spectra of FT-IR and Raman data

$$f_1(\tilde{s}) = \sum_{i=1}^{n} \tilde{s}_i^2,$$

$$f_2(\tilde{s}) = \sum_{i=2}^{n-1} \frac{\tilde{s}_{i-1} - 2\tilde{s}_i + \tilde{s}_{i+1}}{(\Delta\nu)^2}.$$

Therein $\Delta\nu = \nu_{i+1} - \nu_i$ is the step-size in frequency direction. The objective function for the constrained optimization reads

$$f(t) = \gamma_1 g_1(\tilde{s}(t)) + \gamma_2 g_2(\tilde{s}(t)) + \gamma_3 f_1(\tilde{s}(t)) + \gamma_4 f_2(\tilde{s}(t))$$

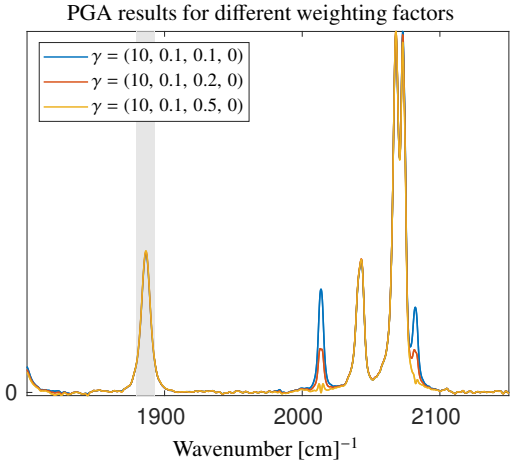

Figure 2: PGA is applied to the peak in the frequency window $I = [1879.6, 1892.6]\,\mathrm{cm}^{-1}$ with fixed weighting factors $\gamma_1, \gamma_2, \gamma_4$, but different values of γ_3. The solution with $\gamma_3 = 0.1$ is not a sufficiently accurate approximation of the spectrum of $Rh_4(CO)_{12}$. The solutions with $\gamma_3 \geq 0.2$ are better.

with weighting factors $\gamma_1 > 0$ and $\gamma_2, \gamma_3, \gamma_4 \geq 0$. The vector $t \in \mathbb{R}^{z-1}$ defines the profile $\tilde{s}$ as in (2). The goal is to minimize the objective function $f(t)$ to obtain a potential pure component spectrum. The objective function is nonlinear and typically has many local minima. Therefore, the numerical minimization can be difficult. A genetic algorithm is used to start the minimization of $f(t)$, and then the nonlinear least squares solver NL2SOL [6] refines the results. For this soft modeling approach, the final profile depends strongly on the selected weighting factors. Fig. 2 shows the results for different settings of the parameter vector $\gamma = (\gamma_1, \ldots, \gamma_4)$ for the data set on rhodium carbonyl complexes. The selected peak range is $[1879.6,\ 1892.9]\,\mathrm{cm}^{-1}$. We look for sharp peaks and select only f_1 as the penalty function in addition to g_1 and g_2. To demonstrate the effect of different weighting factors, we vary $\gamma_3 \in \{0.1,\ 0.2,\ 0.5\}$ in a first calculation. The other weighting factors are fixed at $\gamma_1 = 10$, $\gamma_2 = 0.1$ and $\gamma_4 = 0$. The results vary, and in particular only $\gamma_3 \geq 0.2$ leads to the pure component spectrum of $Rh_4(CO)_{12}$.

See [25] for more details on the PGA as well as [27] for its application to rhodium-catalyzed hydroformylation, including a comparison with results from other MCR methods.

3. Automatic Peak Detection

Peak detection or peak finding algorithms are well known in the chemometric literature. Some of these algorithms combine peak detection with peak modeling, i.e., fitting the peaks to a particular peak model, while others focus on finding the peak positions in high-dimensional data, including multiple characteristic peaks. Finding peaks in noisy data is also a challenge. Some references with different application areas are [32, 8, 7, 29, 28]. GITHUB provides program codes for peak detection as the program codes _peak_finding.py or peakdet.m. Not surprisingly, all these peak finding strategies share common ideas.

In our first work on PGA, see [25], and in the associated original FACPACK implementation, PGA required several manual inputs. The initial peaks had to be selected by marking a frequency window, then the number z of right singular vectors had to be fixed, and finally the weighting factors for the constrained optimization had to be set. The first step took the most time and had to be done for each peak/species. For the second step, a subsystem, subspace or evolving factor analysis [19, 18] provided indicators to determine the correct values for z. For the third step, the program provided default settings that could be modified by the user. The definition of a proper frequency window is important for PGA, as otherwise the normalization in (3) could result in large entries outside the window, which can make it difficult for the constrained minimization to determine a global minimum. Next, we propose a strategy to automate these steps. We modify the algorithm so that we no longer consider frequency windows, but instead use only isolated frequency channels to select peak positions. The exact selection of the channels is not important as long as the channel indices fall within the frequency ranges where the peaks are clearly visible. The peak detection algorithm works on the complete (time) series of spectra. Peak selection techniques are developed in frequency and time directions.

3.1. Techniques for the Peak Detection

In the next subsections, we propose four peak detection techniques, all of which aim to decide whether a peak is worthy of being analyzed by PGA. The first technique locates minima of the (discrete) second derivative of all

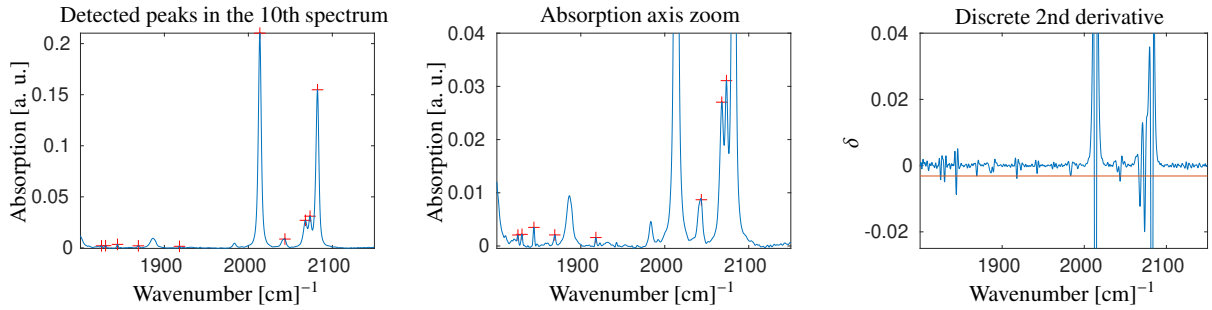

Figure 3: Peak detection by the second derivative along the wavenumber axis. Left: The 10th spectrum is shown after Savitzky-Golay smoothing. The detected peak centers are marked with red plus symbols. Center: Zoom in along the axis of absorption. Right: The discrete second derivative (blue) and the threshold level $\epsilon_1 = -3.1 \cdot 10^{-3}$ (red).

measured spectra in the frequency direction. Another criterion does the same for the right singular vectors. Two other techniques evaluate the first discrete derivatives of the spectra series in the time direction to detect characteristic changes. All of these techniques produce indicator values for the presence of a peak.

Typically, adjacent frequency channels have similar indicator values. Therefore, indicator value thresholds are defined and the algorithm returns all frequency channels whose indicator values exceed the respective threshold. In addition, adjacent channel indices clustered around a peak are combined into a single index, possibly at best at the peak center. Typically, noise makes this approach difficult. To reduce the effects of noise, we first apply Savitzky-Golay filtering [24] to smooth the data. Finally, PGA is applied to each of the selected peaks. Section 3.7 describes the details of the implementation in FACPACK.

3.2. Strategy 1: Second Derivative in Frequency Direction

The second order derivative of spectral line shapes such as Gauss, Lorentz or Voigt profiles are well known to have a minimum at their peak center. If an experimentally obtained peak is close to these line shapes and if it is a sufficiently smooth function, then its peak center can be found by locating a minimum of its second derivative. In the frequency-discrete representation of the mixture spectra, the second derivative is approximated by the second-order finite difference

$$\delta_j = \frac{D_{i,j-1} - 2D_{i,j} + D_{i,j+1}}{(\Delta\nu)^2}, \quad j = 2, \ldots, n-1.$$

Therein $\Delta\nu = \nu_{i+1} - \nu_i$ is the step size in the frequency direction, or the wavenumbers axis for FT-IR spectra. A peak center is located if $\delta_j < \epsilon$ for an appropriate threshold value $\epsilon < 0$. Sec. 3.7 explains how to determine a proper ϵ. Fig. 3 illustrates the application of the discrete second derivative to the spectrum no. 10 of the experimental data set given in Sec. 1.3.

To prevent noise from destabilizing the computation of the second derivative, we have applied to each spectrum a Savitzky-Golay smoothing filter based on polynomials of order 2 and least-squares fitting to 9 neighboring absorbance values. If the jth channel of the spectrum $D(i,:)$ is considered, then a 2nd order polynomial is fitted to $D_{i,j-4}, D_{i,j-3}, \ldots, D_{i,j+4}$. The second derivative of this polynomial, evaluated at ν_j, is taken as the value of δ_j that is used for the computation.

Peak center detection by computing the second derivative of the smoothed spectrum is applied to each spectrum and results in a preliminary indicator value for each frequency channel. The second step is to compute the minimum of the discrete second derivatives at each frequency channel along the series of spectra. Clusters of adjacent minima are processed in a third step. A user-definable sensitivity level is employed to extract characteristic and separate minima of the second derivatives representing the detected peaks. Fig. 4 illustrates these steps for the given data set.

3.3. Strategy 2: Extrema of the Leading Singular Vectors

The right singular vectors of spectroscopic data matrices from vibrational spectroscopy reproduce in some sense the peak pattern of the mixture spectra. However, the peaks in the singular vectors are oriented in the positive and the negative directions. Since PGA constructs the spectral profiles with respect to the basis of the right singular vectors, it is natural to use them for an automated peak detection. This peak detection approach works with series of spectra for which multiple singular vectors are available. The peak detection is similar to the second-order derivative criterion described earlier. However, the entries of the singular vectors can also be negative. Thus, we search not only for minima, but also for maxima of the second derivative.

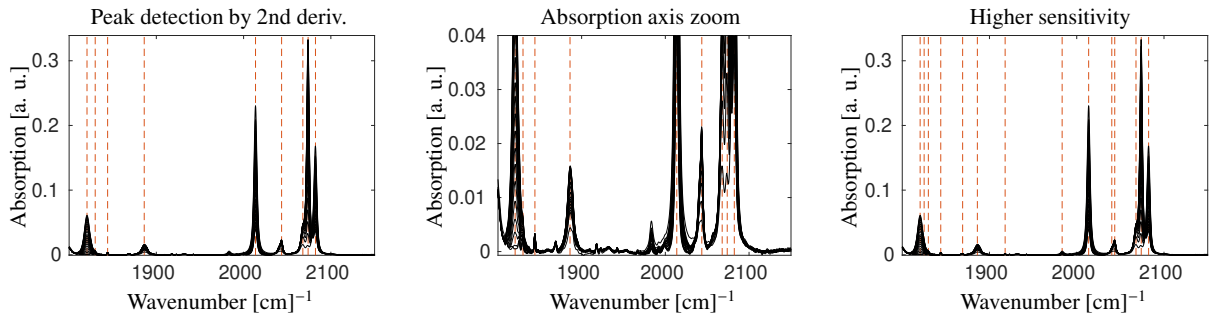

Figure 4: For the second order discrete derivative approach, the detected peaks are marked by red dashed vertical lines. Left and center: For the default sensitivity level, the algorithm detects 9 peaks. Right: At a user-defined higher sensitivity the algorithm detects 14 peaks.

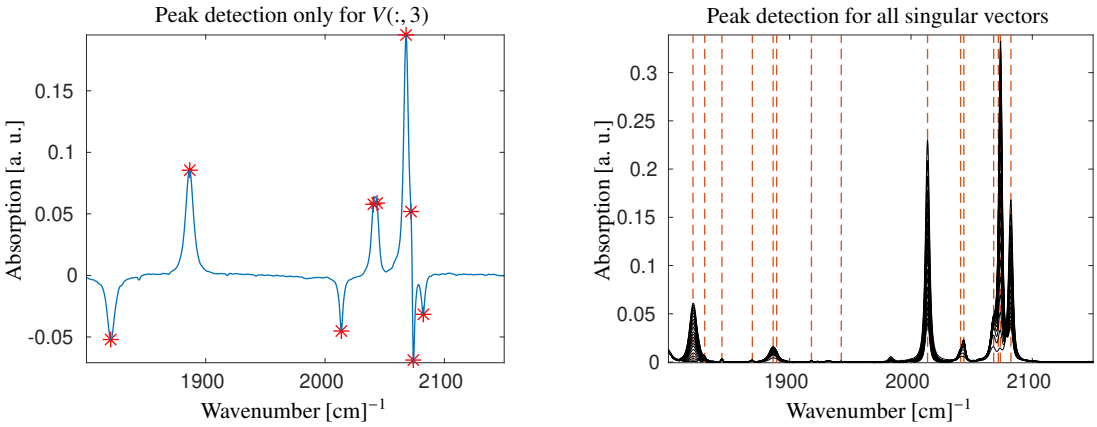

Figure 5: Detected peaks by finding minima and maxima of the leading right singular vectors. Left: the 3rd right singular vector and the detected peaks. Peak centers are indicated by the minima and the maxima of the 2nd order discrete derivative in frequency direction. In all other respects, the criterion is comparable to that explained in Sec. 3.2. Right: the detected 15 peaks (by red dashed vertical lines) by using the first $z = 8$ right singular vectors of the complete data set.

For noisy data, we consider z right singular vectors $V(:, i)$ with $i = 1, \ldots, z$, where z can be larger than the number of expected chemical species. Typically, the impact of noise on the singular vectors increases with increasing i, and $V(:, 1)$ is usually the smoothest singular vector. To compensate for this effect of noise, we first apply the detection algorithm to the leading z singular vectors and then multiply the indicator values for the singular vector $V(:, i)$ by the ith singular value σ_i. As i increases, the σ_i tend to zero. Fig. 5 illustrates the results for the third singular vector as well as for the complete data set.

3.4. Strategy 3: Changes of Spectra along Time Axis

This approach works with (time) series of spectra and analyzes changes in the absorption values along the time axis at fixed frequency coordinates. These changes can indicate the appearance or disappearance of a peak at just that frequency channel. The largest changes occur at the peak center, so the center of a peak can be distinguished from its flanks. Again, we first smooth the data and then compute the first-order finite differences

$$d_i^j = \frac{D_{i+1,j} - D_{i,j}}{\Delta t}, \quad i = 1, \ldots, k-1,$$

where Δt is the (time) step size. The time axis can be replaced by other parameter axes if the series of spectra is generated for the corresponding parameter changes. The only requirement is that the spectra continuously depend on the parameter change. Formally, we can use $\Delta t = 1$ in such cases. Summing up these changes in terms of the sum

$$\tau_j = 0.5\left|d_1^j\right| + \sum_{i=2}^{k-2} \left|d_i^j\right| + 0.5\left|d_{k-1}^j\right|, \tag{4}$$

then τ_j approximates the integral

$$\int_{t_0}^{t_1} \left|\frac{\partial d(t, \nu)}{\partial t}\right| \mathrm{d}t,$$

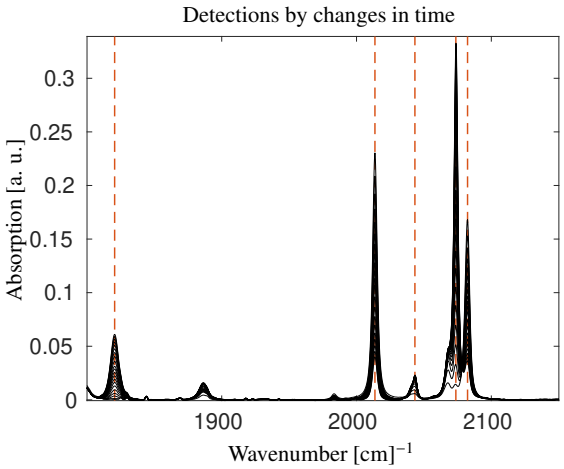

Figure 6: Detected peaks (red dashed vertical lines) for the complete data introduced in Sec. 1.3 by using the strategy 3. With the default parameter settings, the detection criterion finds 5 peaks. By increasing the sensitivity level, the algorithm still fails to detect peaks that do not show characteristic changes, namely time-invariant peaks.

which is the L_1 norm of the partial derivative of the continuous absorption function with respect to its time coordinate t. This criterion supplies indicator values for each wavenumber. Again, the application of a Savitzky-Golay smoothing filter typically improves the results. In this case, we use a least-squares fit of a first order polynomial passing through a number of 5 neighboring points for smoothing. High values of τ indicate peak centers. In practice, clustering the results for this criterion is more complicated than for the other techniques, i.e. to decide where a peak ends and a new one begins. Fig. 6 illustrates the application of this criterion to the given data set.

3.5. Strategy 4: Windowed Sample Variance

Next, we discuss an approach to peak detection based on a local (statistical) variance analysis in both frequency and in time directions. Obviously, the time direction variance of the absorption values of a peak at a fixed frequency channel is much larger than when there is no peak. The same is true when the variance of the absorption values along the frequency axis is considered for a fixed spectrum.

Next we focus on a windowed variance analysis for a time series of spectra. The window is moved through the full data set and the current spectrum is supplemented by k_t leading and k_t trailing spectra. The mean value of the absorption data in this time window at the frequency channel index j is

$$\bar{D}_{i,j} = \frac{1}{2k_t + 1} \sum_{\ell=i-k_t}^{i+k_t} D_{\ell,j}$$

for the index $i = 1 + k_t, \ldots, k - k_t$ running through the windows. Then the unbiased sample variance is

$$s^2_{i,j} = \frac{1}{2k_t} \sum_{\ell=i-k_t}^{i+k_t} (D_{\ell,j} - \bar{D}_{i,j})^2. \tag{5}$$

The maximum over all windows gives the final peak indicator measure

$$\bar{s}^2_j = \max_i (s^2_{i,j}).$$

Fig. 7 shows the detected peaks for the data set on rhodium carbonyl clusters.

3.6. Combination of Peak Detection Techniques

The peak detection problem, which is essentially a pattern recognition problem, can also be solved using other techniques. All of these techniques have their strengths and weaknesses depending on the specific signal profile and the amount of overlap between adjacent peaks. A promising approach is to combine the individual techniques in order to construct an improved strong algorithm. Potentially, machine learning methods can help to find an optimized hybrid peak finding technique.

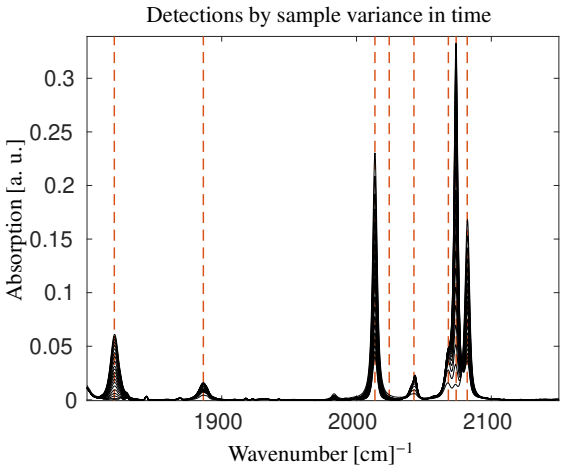

Figure 7: Peak detection based on the moving window sample variance in time direction. The default sensitivity level results in 8 peak detections (red dashed vertical lines).

3.7. Peak Acceptance Threshold

The four strategies as introduced above provide preliminary peak detection by individual indicator values for various frequency channels. The selected frequency channels tend to cluster at or near the true peak position. The next step is to group the clustered indicators into a single frequency channel. Optimally, each peak is then represented by only one frequency channel. This means that all but one indicator values of the cluster are set to zero so that only one frequency channel and its indicator value represent the peak found. The size of an indicator value typically correlates with the importance of the peak. The final step of the peak detection process is to define an acceptance threshold and to reject all potential peaks with an indicator value below this threshold. Three control parameters help to determine a proper acceptance threshold:

- The minimal number of peak detections $m_{\min}$,
- the maximal number of peak detections $m_{\max}$,
- and the sensitivity level α.

These parameters apply to each of the four peak finding strategies. All indicator values are sorted in descending order of their absolute values. We call the resulting function $L(i)$. We observed that $L(i)$ behaves roughly like an exponentially decaying function. Next, we apply a construction that is similar to the L-curve criterion for balancing weighting factors in constrained optimization [4, 11, 12]. For the following explanations see Fig. 8. First, draw a secant through the two points $(m_{\min}, L(m_{\min}))$ with $m_{\min} = 1$ and $(2\,m_{\max}, L(2\,m_{\max}))$, see the black straight line in Fig. 8. Then the point of greatest orthogonal distance of the secant from the curve $L(i)$ is determined. This point is marked with a red star. The associated ordinate value is multiplied by a sensitivity parameter α and gives the final acceptance threshold. Fig. 8 illustrates this procedure for $m_{\min} = 1$, $m_{\max} = 10$ and the sensitivity parameter $\alpha = 1$.

3.8. Correlation Analysis of the Profiles Resulting from PGA

The peak detection phase is followed by a second phase in which PGA is applied to all detected peaks. If m is the number of peaks found, then PGA results in m spectra $s^{(1)}, \ldots, s^{(m)}$. All of these spectra typically contain multiple peaks, not just the single peak that was the anchor point of the PGA. The set of m constructed spectra can be expected to contain repeated spectra, namely groups of similar spectra. The next step is to compute for the m spectra the $m \times m$ correlation matrix W with the matrix elements

$$w_{ij} = \frac{(s^{(i)},\, s^{(j)})}{\|s^{(i)}\|_2 \|s^{(j)}\|_2} = \frac{\sum_{\ell=1}^{n} s_\ell^{(i)} s_\ell^{(j)}}{\|s^{(i)}\|_2 \|s^{(j)}\|_2}$$

for $i, j = 1, \ldots, m$. Finally, we group all spectra whose correlation value is greater than δ with $0 \ll \delta \leq 1$ into a cluster, expecting that these spectra belong to the same chemical species.

A problem with this procedure is the fact that the relation (in the sense of mathematics) between two spectra of belonging to the same species is not necessarily transitive. This means that a sufficiently large correlation of the spectra in the pairs $(s^{(X)}, s^{(Y)})$ and $(s^{(Y)}, s^{(Z)})$ does not necessarily imply the same correlation level for $(s^{(X)}, s^{(Z)})$. In this sense, a spectrum $s^{(\ell)}$ is considered to be an independent spectrum if all non-diagonal entries in the ℓth row of W are less than δ.

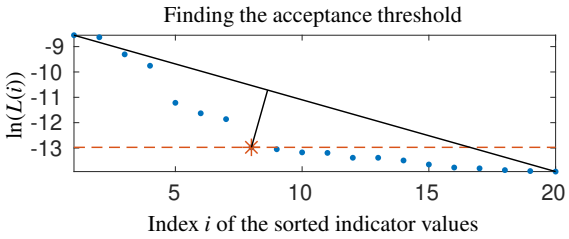

Figure 8: Determining the peak acceptance threshold (the red dashed line) here for the windowed sample variance. In a preliminary step, the clustered frequency channel indicators are grouped so that each peak found is represented by an associated channel index. The blue points are the natural logarithms of the 20 largest indicators forming the function $L(i)$. The solid black line is the secant through $(1, L(1))$ with $m_{\min} = 1$ and $(20, L(20))$ with $2m_{\max} = 10$. The sensitivity parameter is $\alpha = 1$ and the peak acceptance threshold (red dashed line) is defined by the point (red star) with the largest orthogonal distance to the secant. By changing the sensitivity parameter α it is possible to change the peak acceptance threshold and the number of peaks finally accepted.

	1821 cm^{-1}	1830 cm^{-1}	1844 cm^{-1}	1886 cm^{-1}	2014 cm^{-1}	2024 cm^{-1}	2038 cm^{-1}	2044 cm^{-1}	2068 cm^{-1}	2074 cm^{-1}	2082 cm^{-1}
Strategy 1: 2nd derivative	×	×	×	×	×			×	×	×	×
Strategy 2: singular vec.	×	×	×		×			×	×	×	×
Strategy 3: Time changes	×			×			×	×	×	×	
Strategy 4: Variance	×			×	×	×		×	×	×	×

Table 1: Detected peaks by the four strategies for the rhodium carbonyl complexes, see Sec. 1.3. The control parameters are $m_{\min} = 1$, $m_{\max} = 10$, $\alpha = 1$. The peak indicators at 1820 cm^{-1} and 1821 cm^{-1} are merged to 1820 cm^{-1} as strategy 3 detects both of them. The peaks at 1821 cm^{-1}, 2044 cm^{-1} and 2074 cm^{-1} belong to $Rh_6(CO)_{16}$, the peaks at 2014 cm^{-1} and 2082 cm^{-1} belong to $Rh(acac)(CO)_2$ and the peaks at 1886 cm^{-1}, 2043 cm^{-1}, 2068 cm^{-1} and 2073 cm^{-1} belong to $Rh_4(CO)_{12}$. The signal at the wave number 1844 cm^{-1} seems to belong to an artefact.

4. Numerical Results

4.1. Results for Rhodium Carbonyl Complexes Data

First, we apply the automatic peak detection strategies to the data set from Sec. 1.3. With the control parameters $m_{\min} = 1$, $m_{\max} = 10$ and $\alpha = 1$ we get the results as shown in Fig. 9 and in Table 1.

Combining the results of the four strategies we get 11 peak detections. This does not contradict $m_{\max} = 10$, since the limit of maximal 10 peak detections applies to each strategy individually. Instead, we have most of the peaks detected by more than one of the strategies. We apply the PGA to all these 11 frequency channels and compute a set of 11 potential pure component spectra. These 11 profiles belong to 4 independent and weakly correlated pure component spectra as explained in Sec. 3.8. Two strategies suggest a peak at 1844 cm^{-1}, but PGA cannot extract a pure component from this peak. Results based on previous in-situ spectroscopic and DFT computational studies are in accordance with the three observable chemical species in the mixture, namely $Rh(acac)(CO)_2$, $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$ [1, 15]. However, the spectrum of $Rh_4(CO)_{12}$ contains a weak signal at 2014 cm^{-1}, which is an artefact and does not belong to this species. Fig. 10 shows the results, namely the three reconstructed pure component spectra and a fourth spectrum indicating a failed numerical optimization.

4.2. Application to FT-IR Data from Equilibrium of Iridium Complexes

This section reports on an application of PGA to FT-IR data on a dynamic equilibrium mixture of three iridium complexes ($HIr(CO)_3(PPh_3)$, $HIr(CO)_2(PPh_3)_2$ and $H_3Ir(CO)(PPh_3)_2$) acting as catalysts in the hydroformylation of olefins [14]. The population of the hydrido iridium carbonyl complexes depend on the partial pressure of carbon monoxide and hydrogen. The data consists of a series of $k = 47$ FT-IR spectra for varying partial pressures of carbon monoxide between $4 \cdot 10^{-4}$ MPa and 3.9 MPa and a constant hydrogen partial pressure of 1.0 MPa. Each spectrum includes absorption measurements at $n = 913$ frequencies in the range 1900 cm^{-1} to 2150 cm^{-1}. Fig. 11 shows the data.

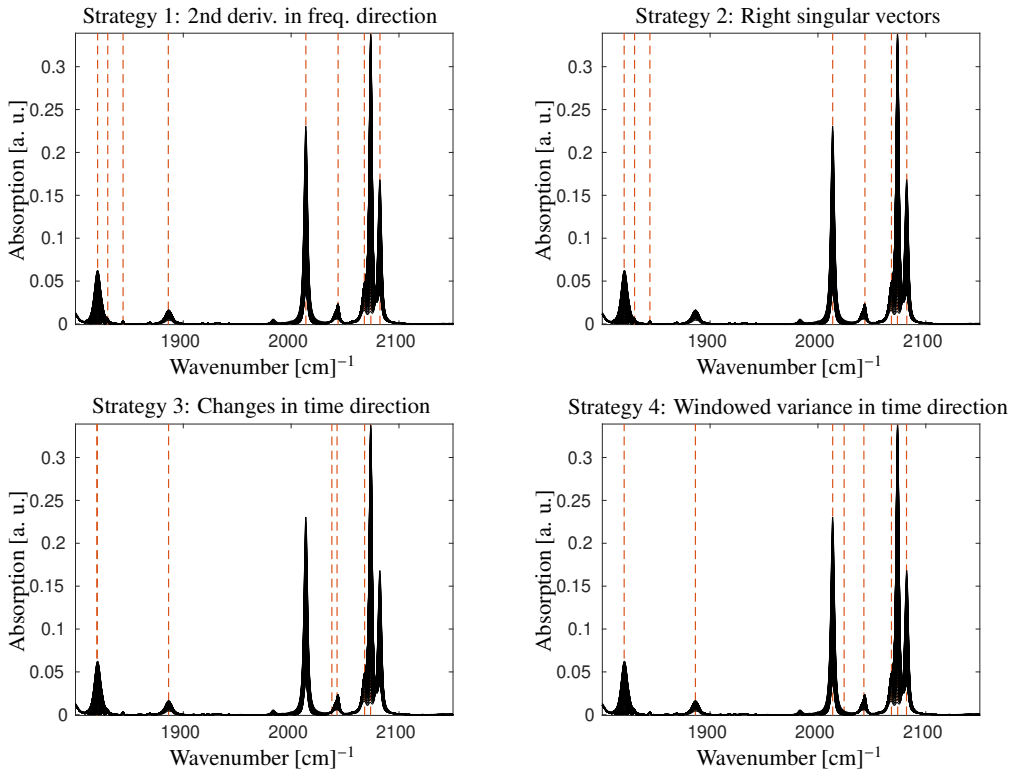

Figure 9: Automatic detection of peak centers (red dashed vertical lines) for FT-IR data from rhodium carbonyl complexes, see Sec. 1.3. Parameters are at least $m_{\min} = 1$ peak detections, a maximal number of $m_{\max} = 10$ detections and a sensitivity level $\alpha = 1$.

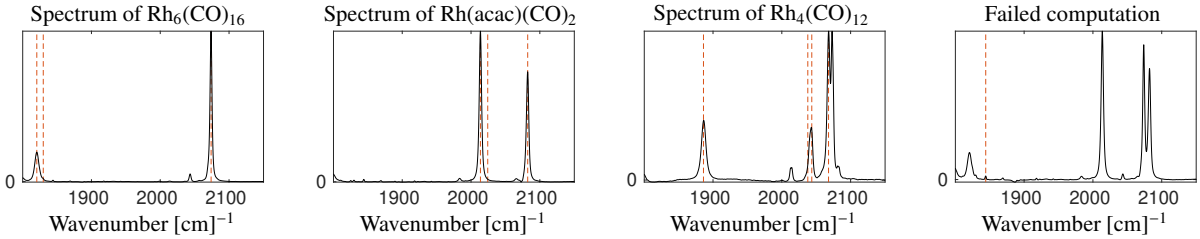

Figure 10: The results of 11 PGA runs using 11 detected peaks for the data from Sec. 1.3. The correlation analysis finds 4 uncorrelated spectra. The spectral channels with a detected peak that have led PGA to the spectrum are marked with red vertical lines. The profile on the right may refer to a combination of the spectra of $Rh_6(CO)_{16}$ and $Rh(acac)(CO)_2$ or may be due to a failed optimization (local minimum). In the PGA graphical user interface, such a spectral profile can be rejected manually. The spectrum of $Rh_4(CO)_{12}$ contains a weak signal at 2014 cm^{-1} which represents an artefact (possibly related to $Rh(acac)(CO)_2$).

We apply the PGA analysis with parameter settings $m_{\min} = 1$ and $m_{\max} = 10$ on the minimal and maximal number of peak detections. The sensitivity level is $\alpha = 1$. The results of the automatic peak detection are shown in Fig. 12. The first peak finding strategy using the second derivative in the frequency direction results in a number of $m = 9$ detections. For these peaks, PGA computes 9 partially correlated pure component spectra. The correlation analysis results in three clusters representing three potential pure component spectra. Fig. 13 shows these spectra along with the original peak positions (indicated by the dashed red lines). These results are consistent with the outcomes presented in [25] except for the scaling.

5. Matlab GUI for PGA

PGA is implemented in the free MatLab-software package FACPACK, which is a toolbox containing various programs for the pure component analysis of spectroscopic data sets. The new PGA module contains the automatic peak detection and correlation analysis. This section explains how to use the software.

5.1. Organization of the GUI Communication Window

The GUI communication window is organized from top left to the bottom right as follows. The starting point is file selection and loading. The GUI informs about the data dimensions and allows to choose the peak detection strategy.

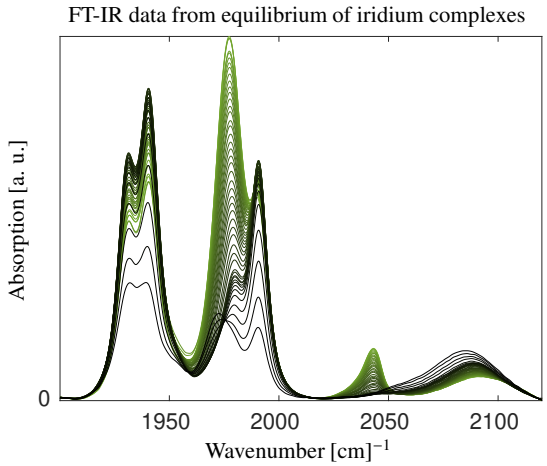

Figure 11: The series of FT-IR spectra on an equilibrium of iridium complexes data as introduced in Sec. 4.2. The color gradient goes from black with $p(\mathrm{CO}) = 3.9\mathrm{MPa}$ to green with $p(\mathrm{CO}) = 4 \cdot 10^{-4}\mathrm{MPa}$.

The top right subwindow allows the user to inspect the left and right singular vectors and singular values. The data display can be limited to a user-defined subset of singular vectors. The user can also specify the number of singular vectors z to be used for signal reconstruction. The lower left subwindow allows the user to construct the objective function for the constrained optimization and to set the weighting factors, see Sec. 2 for the details on the optimization problem. This subwindow allows the user to apply PGA to individually marked peaks (by manual selection) or to sets of automatically detected peaks. The lower right subwindow is used to display, manage and save the results of the calculation. Highlighted buttons or fields in red indicate the next function to be activated or where manual input is required. The following subsections provide a step-by-step guide through a PGA calculation.

5.2. Top left: Data Management Window

The button *Load data* allows loading data from a *.mat file containing the $k \times n$ spectral data matrix D together with an optional n-dimensional vector x of the wavenumber values and an optional k-dimensional vector t containing the time coordinates of the measured spectra. The vectors x and t are used only for axis labeling in the respective plots. Figure 14 shows a screenshot after loading the data.

In the second step, the user can either select a specific frequency window for a PGA application or start automatic peak detection. Manual and automatic peak detection can also be mixed. Peak detection can be started multiple times, with check-boxes allowing activation of single or multiple peak detection strategies. The user can optionally change the maximal number of detections $m_{\max}$ and the sensitivity level α by dragging a slider. The default position of the sensitivity slider is the middle position, which represents $\alpha = 1$. The minimal number of peaks is internally set to $m_{\min} = 1$ to reduce the number of parameters. The plot at the top right of this window shows the series of spectra as well as the selected peaks. If the user does not wish to use the automatic detection, spectral intervals containing a peak can be selected by pressing the left mouse button while moving the cursor over the peak.

5.3. Top right: SVD Window

An SVD of the spectral data matrix is automatically computed immediately after the data is loaded. This subwindow of the GUI is used to display the left singular vectors (columns of factor U), the right singular vectors (columns of factor V) and the singular values in a semilogarithmic plot. After an evaluation of the SVD, the user is required to specify the number z of right singular vectors to be used for the reconstruction of the spectra. The number z should be greater than or equal to the number of active chemical species. In most cases z is equal to the number of singular values of the dominant (clearly nonzero) singular values. In some cases it is advantageous to choose z slightly larger. Click on the zth singular value in order to activate a computation using z-dimensional subspaces of singular vectors.

5.4. Bottom left: Optimization Control Window

This subwindow of the GUI allows the user to set the parameters for the optimization underlying PGA. The user can construct the objective function by assigning non-zero weighting factors to the constraints. The relative size of these factors is responsible for the balanced attainment of the constraints. However, over-weighting a particular constraint will result in a solution that weakly satisfies the other constraints. The column of *contribution* values informs the user how much each constraint contributes to the computed optimized solution; this is the numerical evaluation of each constraint function for the found optimal solution. The *Manual PGA* button is used to start a single

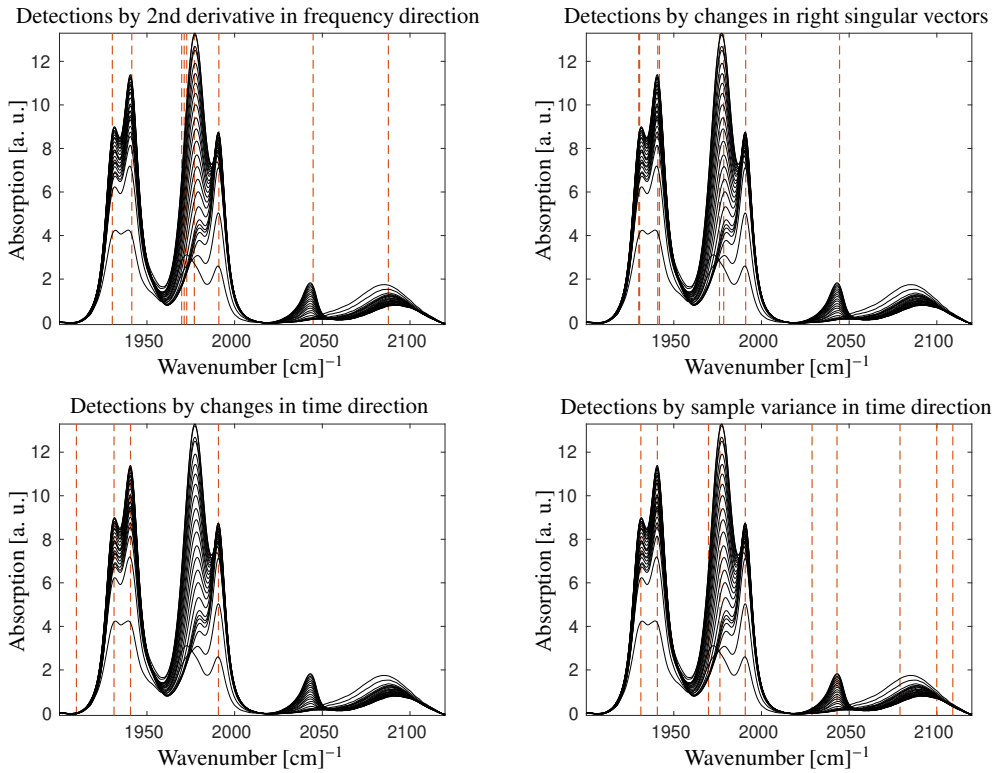

Figure 12: Automatic detection of peaks (red dashed vertical lines) for the data on an equilibrium of iridium complexes, see Sec. 4.2, using the four introduced approaches.

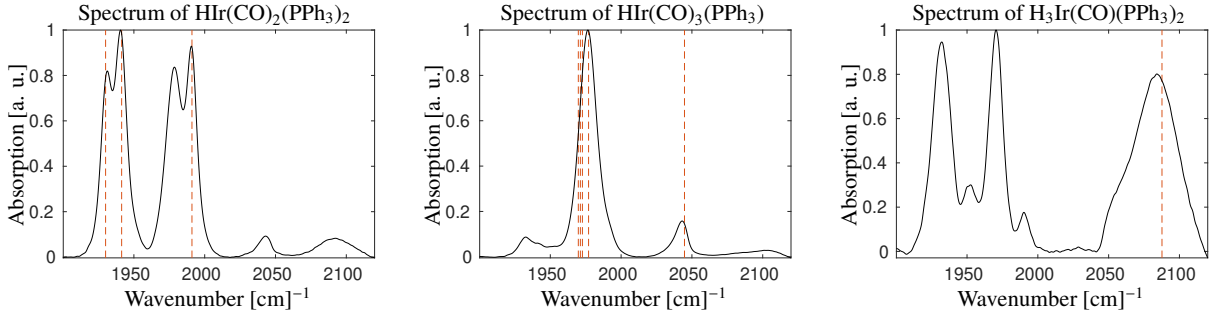

Figure 13: The PGA results using the detected peaks for the data on equilibrium of iridium complexes, see Sec. 4.2. We compute 9 profiles and cluster them into 3 groups of different pure component spectra. The red dashed lines are the frequency channels that lead to the profiles.

PGA run on a selected frequency window. On the other hand, the *Auto PGA* button is used to start multiple PGA runs for the larger number of automatically detected peaks. The buttons are only active after all required manual inputs have been completed. As long as a manual input is missing, the corresponding field is highlighted in red and the associated field *reconstruction details* displays *missing*. In case of a single PGA run, the result is displayed in the solution window (bottom right). In the case of multiple PGA runs, a new window opens showing the series of spectra as computed by PGA for all detected peaks, see Fig. 16. This window allows the user to accept or reject the calculated spectra one by one. A default selection of these spectra is shown in this window, where the default selection is based on the correlation analysis as described in Sec. 3.8. This step can be controlled by additional check boxes.

5.5. Bottom right: Solution Window

The results of the analysis of the complete data set are collected in the lower right window. For a single PGA run, the result is displayed immediately after it is computed. The results of multiple PGA runs are displayed after closing the additional management window. Clicking the *accept* button locks the temporary result and allows the user to continue with other windows or frequency channels. Otherwise, if a new peak window is manually selected or new peaks are automatically detected, the last results are discarded.

The GUI offers three options for saving and exporting data. Option 1 saves only the current pure component spectrum in a *.mat file, along with the wavenumber vector and some information on the PGA run. Option 2 saves

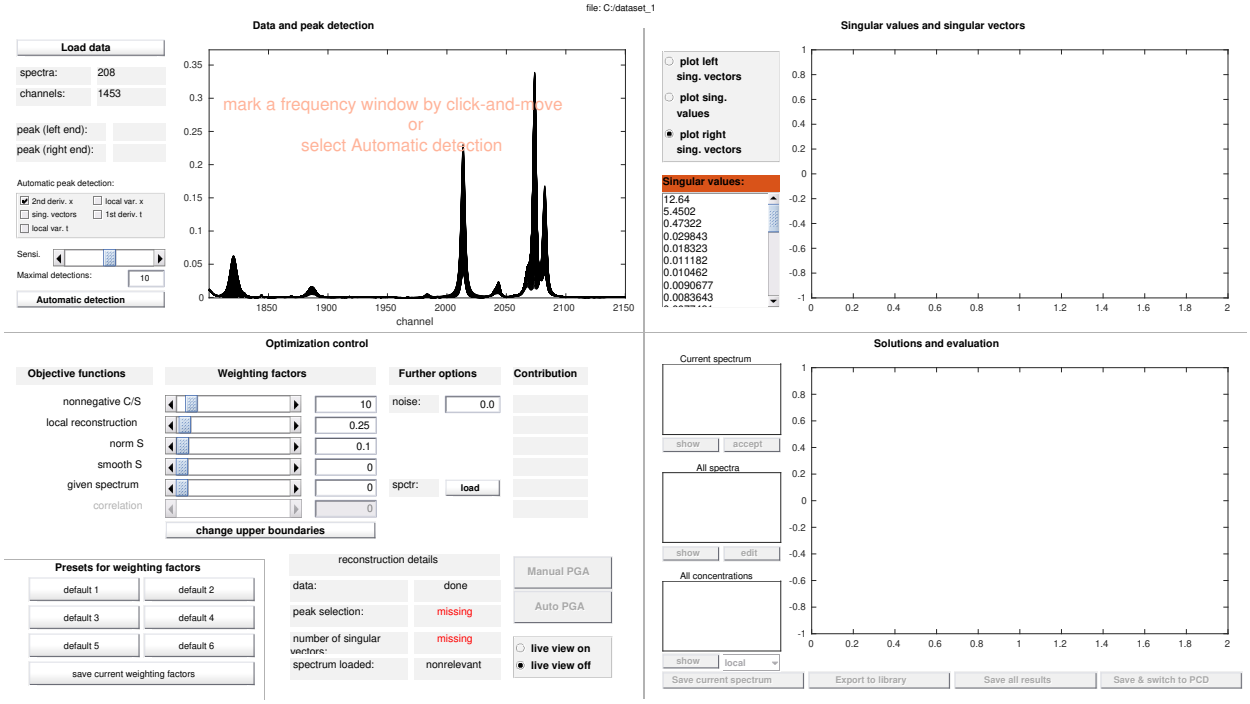

Figure 14: A screenshot of the PGA GUI after loading the data set on rhodium carbonyl complexes as introduced in Sec. 1.3. The next steps are to select the number of singular values (*using vectors*) for the profile reconstruction (this field is highlighted in red), and to mark a frequency window to investigate (move the mouse cursor over the peak while holding down the left mouse button) or to use the automatic detection of multiple peaks (*Automatic detection*).

all results generated by PGA when selecting *save all*. Option 3 exports all results and switches to the FACPACK module PCD, which stands for the Pure Component Decomposition algorithm. This module allows the user to refine the spectra and concentration profiles of pure components. The goal is to determine a pure component factorization of the spectral data matrix D, represented as $D \approx CS^T$. This refined analysis enables a subsequent quantitative analysis. Additional information, such as mass balances or given concentration values at certain times, can be optionally included to improve the results.

5.6. How to use PGA effectively

The PGA software module is designed to be easy to use and adaptable. To obtain all underlying pure component spectra, it is recommended to first use automatic peak detection and then compute the associated potential pure component profiles using automatic PGA for the detected peaks. In a second cycle, the user can manually select and mark certain small peaks to complete the set of pure profiles if any anticipated profiles are still missing.

Due to the design principles, PGA computes the pure component profiles separately without regard to their inherent connections. Hence, PGA cannot benefit from the full and simultaneous factorization approaches underlying many other MCR algorithms. This disadvantage can, however, be cured if PCD is applied as described in Sec. 5.5. Then both the pure component spectra and the associated concentration profiles are made available. But even PGA can compute the associated concentration profiles. The buttons *local* and *global* can be pressed to compute a factor C based on the marked peak positions and the associated spectral profiles.

Finally, the live-view mode of the PGA GUI allows the user to interactively change the weighting factors underlying the objective function. The computational results are automatically visualized as the weighting factors are changed. This functionality is only available within a manually controlled PGA run. This option is enabled by checking the *live view on* radio button. When a chemically meaningful result has been calculated, then select *live view off* and accept the results.

6. Conclusion

It is truly remarkable that the special structure of FT-IR and Raman spectra (with their typical pattern of partially isolated peaks representing specific vibrational modes of the molecules involved) allows PGA to extract the associated pure component spectra from only single isolated peaks. In contrast, a similar approach will never work for instance for UV/Vis spectra with their highly overlapping broad peaks. PGA involves repetitive steps that can and should be

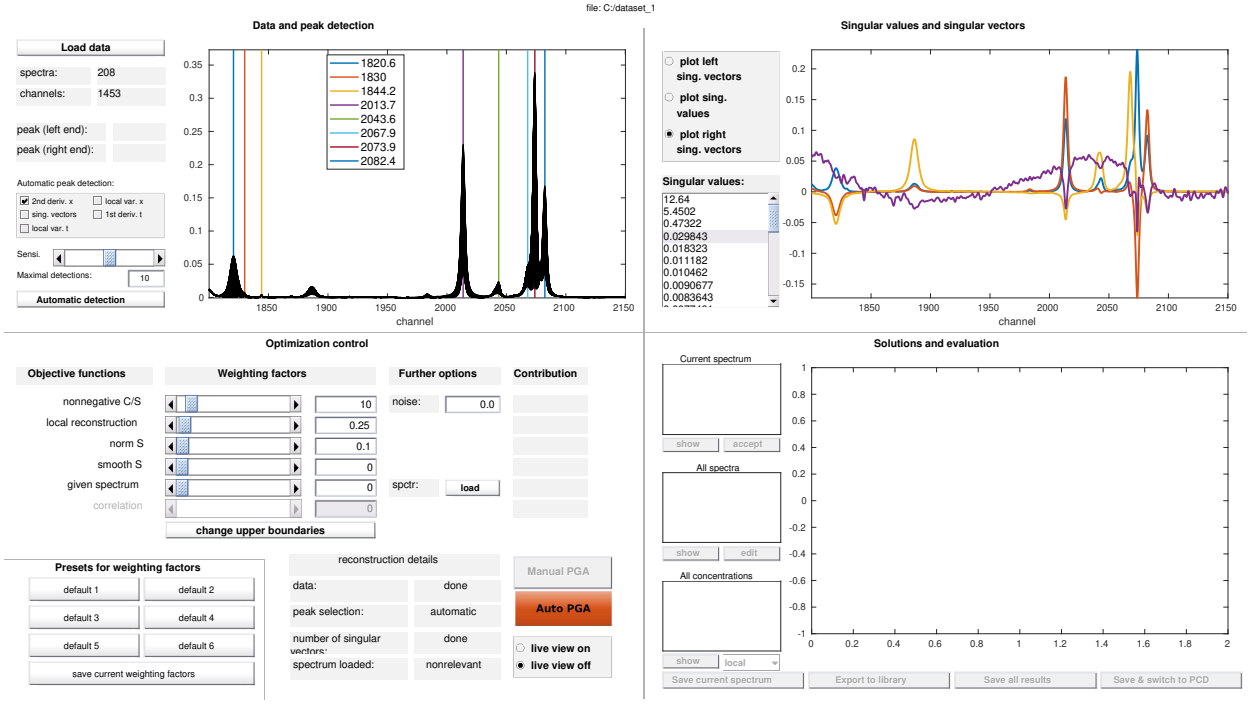

Figure 15: A screenshot of the PGA module after an automatic peak detection. The algorithm detects 8 peaks for the further investigation. The number of right singular vectors is set to $z = 4$ (*using vectors*) and the next step is to compute the spectral profiles. Therefore the button *Auto PGA* is highlighted in red.

automated in order to assist the chemist in the data analysis. In this sense, the automated PGA assists the user in extracting potential pure component spectra from the series of mixture spectra. Finally, chemical expertise is required to evaluate the automatically generated potential pure component spectra, to filter out the true and meaningful spectra, and to refine the computational results by searching for more chemical information, for example, low concentration chemical species within the given spectral data.

We, as a group of authors from chemical research, chemical industry, and numerical mathematics, have developed PGA as a very useful tool that has proven its function and usefulness in various applications in homogeneous catalysis in recent years.

References

[1] A.D. Allian, Y. Wang, M. Saeys, G.M. Kuramshina, and M. Garland. The combination of deconvolution and density functional theory for the mid-infrared vibrational spectra of stable and unstable rhodium carbonyl clusters. *Vib. Spectrosc.*, 41(1):101–111, 2006.

[2] O.S. Borgen and B.R. Kowalski. An extension of the multivariate component-resolution method to three components. *Anal. Chim. Acta*, 174:1–26, 1985.

[3] R. Bro and A. K. Smilde. Principal component analysis. *Anal. Methods*, 6(9):2812–2831, 2014.

[4] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari. Tikhonov regularization and the L-curve for large discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, 123(1-2):423–446, 2000.

[5] A. de Juan, M. Maeder, M. Martínez, and R. Tauler. Combining hard and soft-modelling to solve kinetic problems. *Chemom. Intell. Lab. Syst.*, 54:123–141, 2000.

[6] J. Dennis, D. Gay, and R. Welsch. Algorithm 573: An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7:369–383, 1981.

[7] P. Du, W.A. Kibbe, and S.M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 07 2006.

[8] Siong F.S. and Woei T. An automated approach for analysis of fourier transform infrared (ftir) spectra of edible oils. *Talanta*, 88:537–543, 2012.

[9] H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbühler. Calculation of equilibrium constants from multiwavelength spectroscopic data IV: Model-free least-squares refinement by use of evolving factor analysis. *Talanta*, 33(12):943–951, 1986.

[10] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall, and R. Tauler. A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data. *Anal. Chim. Acta*, 911:1–13, 2016.

[11] P. C. Hansen. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Review*, 34(4):561–580, 1992.

[12] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific and Statistical Computing*, 14(6):1487–1503, 1993.

[13] J. Jaumot, A. de Juan, and R. Tauler. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.*, 140:1–12, 2015.

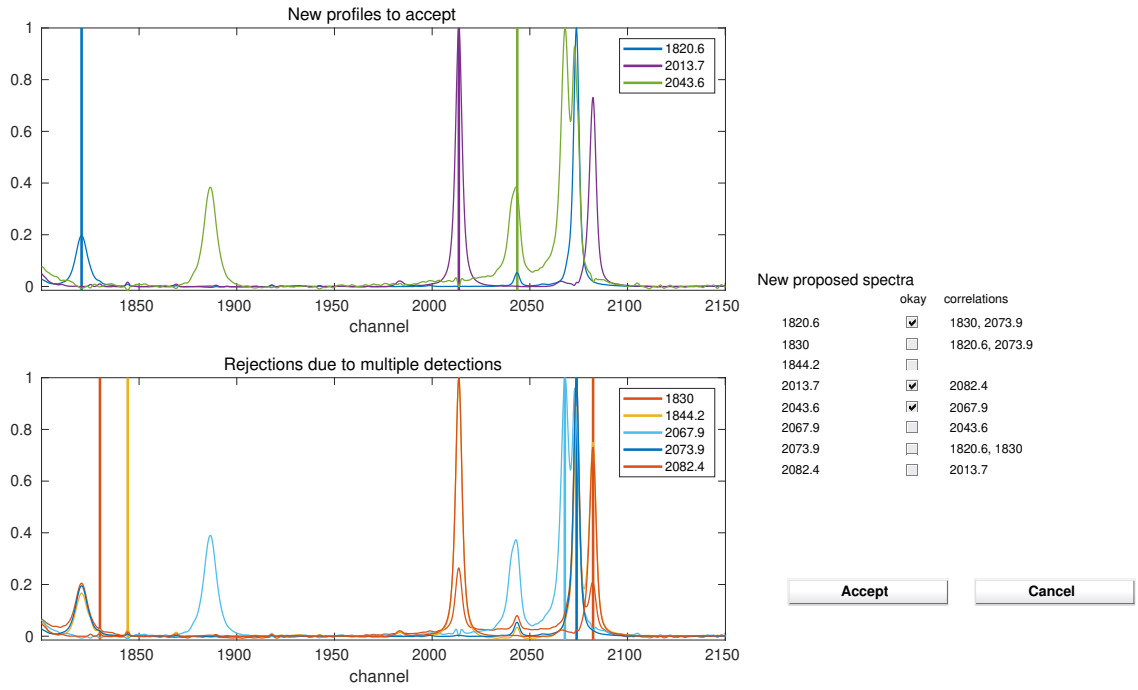

Figure 16: A screenshot of a separately opened window after an automatic computation of multiple pure component spectra. This additional window allows the user to manage and cluster the results. Since each pure component spectrum typically includes more than one peak, multiple a peak detection within a pure component spectrum results in finding the same pure component spectrum multiple times. A correlation analysis of all computed spectra is used to sort out repeated spectra. A manual decision confirms 3 spectra by checking the *okay* boxes and rejects 5 others.

[14] C. Kubis, W. Baumann, E. Barsch, D. Selent, M. Sawall, R. Ludwig, K. Neymeyr, D. Hess, R. Franke, and A. Börner. Investigation into the equilibrium of iridium catalysts for the hydroformylation of olefins by combining in situ high-pressure FTIR- and NMR-spectroscopy. *ACS Catal.*, 4:2097–2108, 2014.

[15] C. Kubis, M. König, B.N. Leidecker, D. Selent, H. Schröder, M. Sawall, W. Baumann, A. Spannenberg, A. Brächer, K. Neymeyr, R. Franke, and A Börner. Interplay between catalyst complexes and dormant states: In situ spectroscopic investigations on a catalyst cystem for alkene hydroformylation. *ACS Catal.*, 13(8):5245–5263, 2023.

[16] W.H. Lawton and E.A. Sylvestre. Self modelling curve resolution. *Technometrics*, 13:617–633, 1971.

[17] M. Maeder and Y.M. Neuhold. *Practical data analysis in chemistry*. Elsevier, Amsterdam, 2007.

[18] M. Maeder and A. Zilian. Evolving factor analysis, a new multivariate technique in chromatography. *Chemom. Intell. Lab. Syst.*, 3(3):205–213, 1988.

[19] M. Maeder and A. D. Zuberbühler. The resolution of overlapping chromatographic peaks by evolving factor analysis. *Anal. Chim. Acta*, 181(0):287–291, 1986.

[20] E. Malinowski. *Factor analysis in chemistry*. Wiley, New York, 2002.

[21] K. Neymeyr, M. Beese, H. Abdollahi, and M. Sawall. Can angle measures be useful in MCR analyses? *preprint*, 2023.

[22] K. Neymeyr, M. Sawall, and D. Hess. Pure component spectral recovery and constrained matrix factorizations: Concepts and applications. *J. Chemom.*, 24:67–74, 2010.

[23] C. Ruckebusch and L. Blanchet. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal. Chim. Acta*, 765:28–36, 2013.

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

[25] M. Sawall, C. Kubis, E. Barsch, D. Selent, A. Börner, and K. Neymeyr. Peak group analysis for the extraction of pure component spectra. *J. Iran. Chem. Soc.*, 13(2):191–205, 2016.

[26] M. Sawall, H. Schröder, D. Meinhardt, and K. Neymeyr. On the ambiguity underlying multivariate curve resolution methods. In S. Brown, R. Tauler, and B. Walczak, editors, *In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, pages 199–231. Elsevier, 2020.

[27] H. Schröder, M. Sawall, C. Kubis, A. Jürß, D. Selent, A. Brächer, A. Börner, R. Franke, and K. Neymeyr. Comparative multivariate curve resolution study in the area of feasible solutions. *Chemom. Intell. Lab. Syst.*, 163:55–63, 2017.

[28] S.-T. Tan, H. Zhu, and W. Chew. Self-modeling curve resolution of multi-component vibrational spectroscopic data using automatic band-target entropy minimization (AutoBTEM). *Anal. Chim. Acta*, 639(1-2):29–41, 2009.

[29] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. van Nederkassel, Y. Vander Heyden, and D.L. Massart. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part i: Peak detection. *Journal of Chromatography A*, 1096(1):133–145, 2005. Chemical Separations and Chemometrics.

[30] E. Widjaja, C. Li, W. Chew, and M. Garland. Band target entropy minimization. A robust algorithm for pure component spectral recovery. Application to complex randomized mixtures of six components. *Anal. Chem.*, 75:4499–4507, 2003.

[31] E. Widjaja, C. Li, and M. Garland. Semi-batch homogeneous catalytic in-situ spectroscopic aata. FTIR spectral reconstructions using Band-Target Entropy Minimization (BTEM) without spectral preconditioning. *Organometallics*, 21:1991–1997, 2002.

[32] Yong-Jie Yu, Qiao-Ling Xia, Sheng Wang, Bing Wang, Fu-Wei Xie, Xiao-Bing Zhang, Yun-Ming Ma, and Hai-Long Wu. Chemometric strategy for automatic chromatographic peak detection and background drift correction in chromatographic data. *J. Chromatogr. A*, 1359:262–

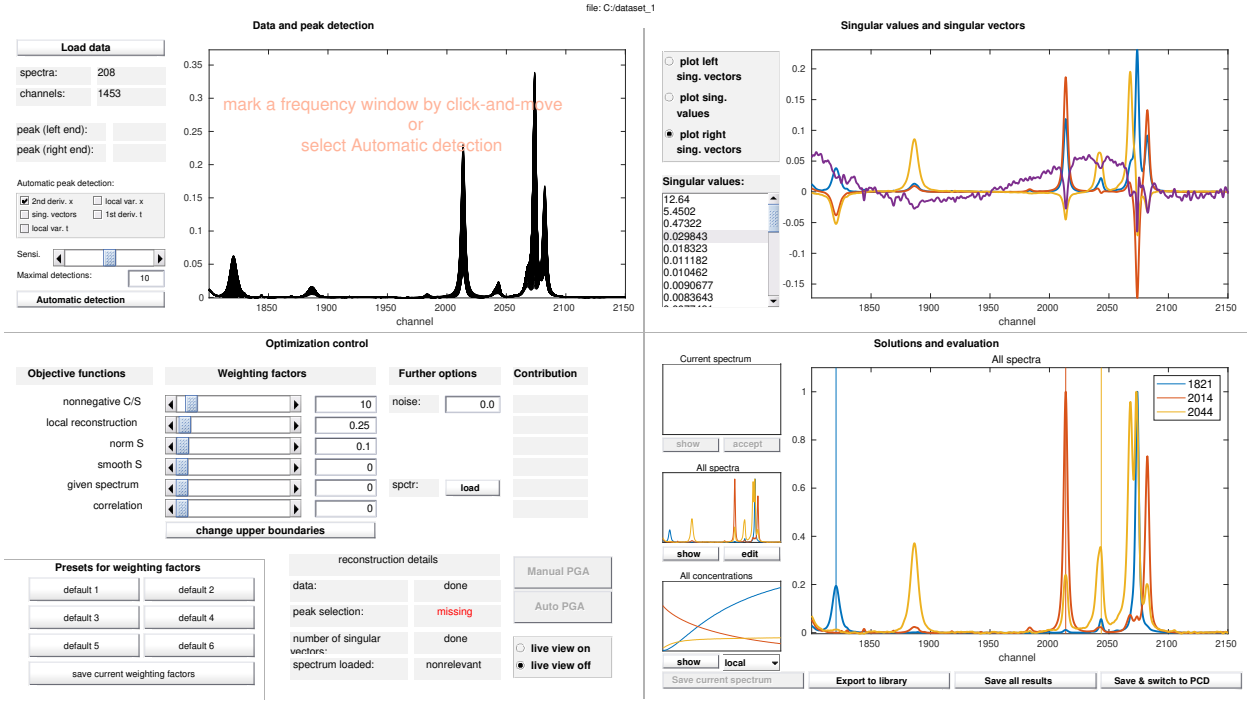

Figure 17: A screenshot of the PGA GUI after accepting the 3 detected and manually confirmed spectra. The PGA analysis is now complete. Optionally, the results can be saved or passed to a quantitative analysis if some additional concentration information is available.

270, 2014.