Peak group analysis for the extraction of pure component spectra.

Mathias Sawall[a], Christoph Kubis[b], Enrico Barsch[b], Detlef Selent[b], Armin Börner[b], Klaus Neymeyr[a,b]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstrasse 69, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse e.V. an der Universität Rostock, Albert-Einstein-Strasse 29a, 18059 Rostock, Germany.*

Abstract

Structure elucidation for the reactive or catalytic species of a chemical reaction system can significantly be supported by spectroscopic measurements. If the spectroscopic data contains isolated signals or groups of partially separated peaks, then the identification of correlations between these peaks can help to determine the pure components by their functional groups.

A computational method is presented which constructs from a certain frequency window, which contains a single peak or a peak group, an associated pure component spectrum on the full frequency range. This global spectrum reproduces the spectrum in the local frequency window or, at least, reproduces the contribution from the dominant component in the local window. The method is called the Peak Group Analysis (PGA). The methodological background of the PGA are a multivariate curve resolution method and the solution of a minimization problem with weighted soft constraints. The method is tested for two experimental FT-IR data sets from investigations into equilibria of hydroformylation catalysts based on rhodium and iridium. An implementation of the PGA is presented as a part of the FACPACK software.

Key words: factor analysis, pure component decomposition, nonnegative matrix factorization, spectral recovery.

1. Introduction

Spectroscopic methods in combination with chemometric techniques are key tools for the determination of unknown chemical species in a reaction system. Here we consider a situation in which the course of a chemical reaction is recorded spectroscopically. If a series of k spectra is taken and each spectrum is a vector of n absorbance values, then the spectral data can be stored row-wise in a k-times-n matrix D. The Lambert-Beer law in matrix form $D = CA$ says that D can be factored in a product of two nonnegative matrices, namely in a concentration factor C and a spectral factor A. In the case of slightly perturbed, experimental data the aim is to determine at least an approximate factorization. If the reaction system contains a number of s independent components, then $C \in \mathbb{R}^{k\times s}$ contains columnwise the concentration profiles in time of the s pure components. The spectral factor $A \in \mathbb{R}^{s\times n}$ contains row-wise the pure component spectra. Such a factorization of D has first been conducted for a two-component system by Lawton and Sylvestre [1]. A fundamental problem of such factorizations is that the matrix factors are not unique due to the so-called rotational ambiguity [2, 3].

The four volumes of the book series *Comprehensive Chemometrics* [4] give a detailed overview on the wide range of chemometric methods and its mathematical analysis. Without claiming any completeness we would like to mention the evolving factor analysis [5], the window factor analysis [6], which each can be combined with Manne's theorems [7], as well as the target factor analysis [8] and, last but not least, the class of Multivariate Curve Resolution (MCR) methods with hard and soft constraints [9, 10].

1.1. The idea underlying the peak group analysis

In the present paper our objective is a bit different. We do not directly seek to compute a nonnegative factorization $D = CA$, but we try to extract single spectra and to identify related peak groups in order to determine some of the species by their functional groups. Our approach is driven by practical needs of the chemist who has some presumptions and assumptions on the species which might be found in the chemical reaction system. For instance in organo-metallic/catalytic chemistry the (IR-)spectra of the reactants and the spectra of the main reaction products are known, but the catalytic active species and the catalyst preformation process is sometimes unknown [11]. However, reasonable

assumptions on the catalytic active species can be made. Together with the molecular point group of the catalyst or its precursor, the character table provides insight into active and characteristic vibrations. Alternatively one can compute reliable approximations of the absorption spectrum by quantum mechanical SCF computations or DFT calculations. For these reasons that the chemist, after having identified a certain signal (like stretching vibrations of terminal carbonyl ligands in rhodium carbonyl complexes), would like to know if this signal is part of a pure component spectrum which also includes a characteristic signal that is associated with a further functional group [12, 13]. In this paper we present a numerical algorithm, called Peak Group Analysis (PGA), which aims at finding correlations between peaks and peak groups which are associated with the same pure component. Mathematically, the algorithm uses a specific target function which includes a weighted combination of soft constraints. The approach is based on a local rank-one reconstruction; its application to spectroscopic data with narrow and partially isolated peaks, which can typically be found in IR spectroscopic data, works very well.

1.2. Overview

In Section 2 a short introduction to the basics of multivariate curve resolution techniques is given and the idea of the PGA is explained. Section 3 presents the mathematical foundation of the PGA and its target function. Further the soft constraint functions are introduced. Some mathematical theorems on the construction of the concentration profiles are contained in Section 4. Section 6 refers to the FACPACK implementation of the PGA [14, 15]. Finally, the application of the PGA to experimental FT-IR data sets is discussed in Section 8.

2. Peak group analysis (PGA)

2.1. Aim of the PGA

The aim of the PGA is to identify those peaks or peak groups in a series of spectra taken from a chemical mixture which can be assigned to the same pure component. The starting point is the specification of a frequency window which contains a certain peak or peak group. Then the PGA intends to provide a pure component spectrum which in the given frequency window more or less reproduces the original signal. The PGA can be applied repeatedly in order to find all pure component spectra step-by-step.

2.2. The general MCR approach

The general approach to reconstruct the pure component factors C and A is based on the singular value decomposition (SVD) of D [1]. Let $U \in \mathbb{R}^{k\times k}$, $\Sigma \in \mathbb{R}^{k\times n}$ and $V \in \mathbb{R}^{n\times n}$ be the factors of the SVD, so that $D = U\Sigma V^T$ [16]. Furthermore let s be the number of independent components underlying the data D; for noise-free data the number s equals the rank of D. Then C and A can be reconstructed only by using the first s left- and right singular vectors [6, 8].

In the case of noisy data it is often advantageous to work with a number of $z \geq s$ singular vectors for the reconstruction of C and A [10]. Then the mathematical formulation reads

$$C = U\Sigma T^{+}, \qquad A = TV^T \tag{1}$$

with U, Σ and V^T containing the first z singular vectors and singular values. The associated transformation $T \in \mathbb{R}^{s\times z}$ is a rectangular matrix. Further $T^{+} \in \mathbb{R}^{z\times s}$ is the so-called Moore-Penrose pseudoinverse of T. This SVD based reconstruction approach reduces the number of degrees of freedom of the factorization problem to sz which is the number of matrix elements of T. Finally, the determination of feasible and chemically meaningful matrix factors C and A amounts to the computation of a proper matrix $T \in \mathbb{R}^{s\times z}$.

2.3. Construction of a single pure component spectrum

Equation (1) is a construction for the simultaneous formation of the spectra and concentration profiles for all factors. In contrast to this the PGA determines the pure component spectra, which are the rows of A, step-by-step. Mathematically a pure component spectrum $a \in \mathbb{R}^{1\times n}$ is a linear combination of the z right singular vectors which belong to the z largest singular vectors. The spectra are written as row vectors so that a has the form $a = tV(:, 1 : z)^T$. Our task is to determine the vector $t \in \mathbb{R}^{1\times z}$ of expansion coefficients. These z degrees of freedom can be reduced to $z - 1$ since any nonzero scaling of the spectrum is without relevance. The Perron-Frobenius theory on the spectrum of a nonnegative matrix [17] provides under mild assumptions on the spectral data matrix D, see Theorem 2.2 in [15], that the first coefficient t_1 is never equal to zero. The mathematical argumentation is as follows: the Perron-Frobenius theory guarantees that $V(:, 1)$ is a sign-constant vector. Without loss of generality it can be assumed as a component-wise nonnegative vector. Orthogonality of this vector to any linear combination of the remaining singular vectors $V(:, 2), \ldots, V(:, n)$ proves

that this linear combination must have positive *and* negative components. Since a feasible spectrum must have only nonnegative components, any feasible spectrum must have a contribution from $V(:,1)$.

All this justifies to use a scaling so that $t_1 = 1$, which for instance has been used in [18, 19, 14] and for the resolving factor analysis (RFA) [8]. Thus t can be written in the form $t = t_1(1, w)$ with $w \in \mathbb{R}^{1\times z-1}$ and $t_1 > 0$. Thus we get $a = t_1(1, w)V^T$.

2.4. Window selection and normalization

The starting point of the PGA is the selection of a channel window $[\nu_\ell, \nu_r]$ along the wavenumber/frequency axis. This window should contain a single peak or group of peaks whose affiliation to other peaks or peak groups of the same pure component within the series of spectra is to be analyzed. The channel window $[\nu_\ell, \nu_r]$ contains the discrete wavenumber values ν_i with respect to the given grid. The set $I \subset \{1, \ldots, n\}$ is the maximal set of indices so that

$$\nu_\ell \le \nu_i \le \nu_r, \qquad \text{for all } i \in I.$$

For the following construction of a pure component spectrum a, which reproduces more or less the selected signal in the window $[\nu_\ell, \nu_r]$, it is useful to normalize $a \in \mathbb{R}^n$ in I in a way that

$$\max_{i\in I} a_i = 1. \tag{2}$$

Together with the non-normalized representation $a = t_1(1, w)V^T$ and $a_i = t_1(1, w)V(i, :)^T$ we prefer to work in the following with the normalized form

$$a = a[w] := \frac{(1, w)\, V^T}{\max_{i\in I}((1, w)\, V(i, :)^T)}. \tag{3}$$

2.5. Stepwise extraction of the pure component spectra

The PGA can be applied repeatedly to a given series of spectra. In each cycle the spectrum of a single component can be extracted. The method works very well especially for IR spectroscopic data with its typically narrow and isolated peaks, see Section 8. If for a certain component the spectrum a (row vector) has been extracted and if for this component the concentration profile c (column vector) is accessible, then the contribution of this component to the spectral data matrix D can be removed by subtracting a proper multiple of the rank-1-matrix $ca \in \mathbb{R}^{k\times n}$ from D. The principles of such a rank-1-downdate of a nonnegative matrix in order to construct in the end a complete nonnegative matrix factorization has been analyzed in [20]. In every step the contribution of one component is removed from D and finally and ideally only the noise remains. The explanation is simple, but in practice such an approach has severe disadvantages due to the influence of noise. The decisive point is that such a stepwise rank-1-downdate of D is very sensible with respect to noise, since the errors of all previous rank-1-downdates accumulate in D. For instance the subtraction of a rank-1-matrix may result in small negative components in D or the subtraction of a certain peak of a slightly perturbed amplitude or frequency position could result in a small remaining peak with a somewhat shifted frequency position. All this adversely affects the accuracy of the subsequent rank-1-downdates.

Alternatively one can always work with the original spectral data matrix D without subtracting rank-1 matrices [21]. This reduces the impact of noise. If finally a series of s independent pure component spectra has been determined, then the associated concentration factor C can be computed by a "global" least-squares computation. Such a procedure appears to be more stable.

2.6. Application to IR data

By construction the PGA can be applied to spectral data which contains several narrow peaks and which also includes, at least for some time intervals, frequency ranges in which no absorption is observed. If, contrary to the foregoing, all pure component spectra show an absorption on the whole frequency range, then it would be difficult for the PGA to extract the contribution from a single component.

In particular the IR or Raman spectroscopy provide data with narrow peaks and several non-absorbing frequency ranges. Then a step-by-step extraction approach could successfully be applied; this has clearly been demonstrated by the BTEM software by Garland and his group [21]. A further technique which is based on a local analysis is SIMPLISMA [22, 23] which has been successfully applied to IR spectral data. In contrast to this the UV/Vis spectroscopy results in spectra which are rather unsuitable for an application of the PGA. Finally, the PGA can principally be applied to NMR data [24]; however the occurrence of the nuclear magnetic resonance chemical shifts necessitates a proper data preprocessing.

2.7. Relations to EFA, WFA and TFA techniques

The evolving factor analysis (EFA) [5] and the window factor analysis (WFA) [6] are powerful techniques for the analysis of spectroscopic data. EFA analyzes

the evolution of the rank of a series of growing submatrices of D. WFA computes the concentration profile of a certain component by using submatrices along the time axis for an evolutionary process; together with Manne's theorem pure component information can be extracted. There are some similarities between the WFA and the PGA, but there are also two differences. First, PGA uses windows along the frequency axis. Second, WFA and EFA are rather fixed computational procedures, whereas the PGA is based on an optimization process with a target function which includes several regularization terms, see Section 3.

Finally, the PGA is very different from the target factor analysis (TFA) [8] where a given factor (spectrum) is tested, whether it contributes to the spectral measurement or not. For the PGA no spectrum has to be known initially.

3. The target function for the PGA

Equation (3) shows the way how to compute a single pure component spectrum a by means of a row vector $w \in \mathbb{R}^{z-1}$. Next w is determined by the solution of a minimization problem for a target function which includes several weighted constraints. The choice of the constraint functions and their weight factors is a crucial step. Their suitable selection depends on the type of the spectroscopic data. The solution of constrained minimization problems for the computation of nonnegative matrix factorizations is a standard procedure in chemometrics, see e.g. [25, 26, 27, 21, 9, 2, 28, 10].

The PGA target function f is formed by a weighted mean of two functions f_1 and f_2 which is combined with several weighted soft constraints

$$f(w) = \omega_1 f_1(a[w]) + \omega_2 f_2(a[w]) + \sum_{i=1}^{q} \gamma_i^2 g_i(a[w]) \quad (4)$$

with $a[w]$ by (3); in the following we simply write a for $a[w]$. Therein $\omega_i \geq 0$ and $\gamma_i \geq 0$ are the weight factors. The functions f_1 and f_2 are:

1. Norm of the spectrum:

$$f_1 = \sum_{j=1}^{n} a_j^2.$$

   A spectrum with a small integral and narrow peaks is favored.

2. Norm of the discrete second derivative of the spectrum:

$$f_2 = \sum_{j=2}^{n-1} \left(\frac{a_{j-1} - 2a_j + a_{j+1}}{(\Delta\nu)^2} \right)^2$$

   with $\Delta\nu$ being the wavenumber increment along the equidistant wavenumber grid. The function f_2 is the sum of squares of the discrete second derivative of the spectrum a with respect to equidistant grid of wavenumber values. By f_2 a smooth spectrum is favored.

The constraint functions are introduced in Section 3.1.

3.1. Constraint functions

The construction of the soft constraints, also called regularization functions, and their weighting is decisive for the computation of meaningful pure component spectra. In order to construct a flexible curve resolution method, which can be applied to several series of spectra from different types of spectroscopic techniques with their different typical shapes, it is useful to have a stock of various regularization functions, see [21, 9, 10] for diverse examples.

For the PGA the following soft constraints are available (and can or cannot be used depending on the present conditions that, e.g., certain spectra are known or are not known):

- Nonnegativity.
- Local reconstruction error,
- Distance (by the sum of squares) to a given pure component spectrum. This constraint is similar to the target factor analysis [8],
- Correlation with other pure component spectra.

All these constraint functions have been written in a functional form depending on a. By (3) a depends on w so that these functions essentially depend on the $z-1$ components of w.

Within each step of the optimization the current approximation of a spectrum a can be used in order to compute a temporary concentration profile c with respect to the channel window I according to

$$c = U\Sigma v^* \quad \text{with} \quad v^* = \frac{V(I,:)^T\, a(I)^T}{\|a(I)\|_2^2}. \quad (5)$$

The resulting pair a and c allows an optimal reconstruction of D in the channel window I; see Section 4 for the mathematical analysis. It is important to note that c and a are only temporary approximations which are changed within each step of the optimization procedure. A further important point is that $a \geq 0$ and $D \geq 0$ imply that

$c \geq 0$; a proof of this fact is given in Corollary 4.2. The result that $c \geq 0$ shows that no further constraint function has to be added to f in order to guarantee the nonnegativity of c. At the end of the iterative minimization, Equation (5) can also be used to compute from the final spectrum a a final approximation of the concentration profile c.

The constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}_+$ read as follows:

1. Nonnegativity: The constraint function which is used to favor an almost nonnegative solution a is

$$g_1 = \sum_{j=1}^{n} \min\left(\frac{a_j}{||a||_\infty} + \varepsilon,\, 0\right)^2.$$

   Therein $\|\cdot\|_\infty$ denotes the maximum norm which is maximum of the absolute values of the components. A small value $\varepsilon \geq 0$ is used to allow slightly negative components for which the ratio $\min a_j/||a||_\infty$ is larger than $-\varepsilon$; accepting such small negative entries can be very helpful for finding a solution in the case of noisy data. In Corollary 4.2 it is shown that the associated concentration profile c is nonnegative, if a and D are nonnegative. Hence a constraint function on the nonnegativity of c is not needed.

2. Local reconstruction: With υ^* by (5) and $c = U\Sigma\upsilon^*$ the local reconstruction error is

$$g_2 = \left\| \Sigma \left(V(I,:)^T - V(I,:)^T \frac{a(I)^T a(I)}{||a(I)||_2^2} \right) \right\|_F^2.$$

   Therein $a(I)$ and $V(I,:)$ are the vector resp. matrix which are reduced to the indices contained in the index set I. Further $\|\cdot\|_F$ is the Frobenius norm, which is the square root of the sum of squares of its argument.

3. Distance to a given spectrum $\hat{a} \in \mathbb{R}^{1\times n}$: This constraint function measures the distance of the optimally scaled $\hat{a}$ to a

$$g_3 = \sum_{j=1}^{n} (\alpha\hat{a}_j - a_j)^2 \quad \text{with} \quad \alpha = \frac{\hat{a}a^T}{||\hat{a}||_2^2}$$
$$= ||a||^2 - \frac{\hat{a}a^T}{||\hat{a}||^2}.$$

4. Correlation with other spectra: This constraint function favors a solution with a small correlation with other pure component spectra $A(i,:)$, $i = 1, \ldots, s_0$,

$$g_4 = \sum_{i=1}^{s_0} \sum_{j=1}^{n} (A(i,j)\, a_j)^2.$$

3.2. Numerical solution of the minimization problem

The target function f in (4) defines a nonlinear least squares problem with $z-1$ free variables. In our FACPACK implementation of the PGA, see Section 6 we use a combination of genetic algorithm and of the ACM software NL2SOL [29] written in FORTRAN.

A careful choice of the weight parameters, especially the choice of ω_1 and ω_2, is very important for computing reliable and meaningful spectra. If a channel window is selected which contains peaks originating from more than one component, then the local reconstruction cannot be successful and γ_2 should be relatively small.

4. Concentration profiles by local reconstruction

A well-known approach to construct the concentration profile c which fits bests to a certain pure component spectrum a is to compute

$$c = Da^+$$

with the pseudo-inverse a^+ of a, see e.g. Equation (5) of [21]. With the representation $a = tV^T$ of a it holds that

$$c = Da^+ = U\Sigma V^T V t^+ = U\Sigma t^+.$$

The pseudo-inverse of t reads

$$t^+ = \frac{t^T}{||t||_2^2} = \frac{V^T a^T}{||aV||_2^2} = \frac{V^T a^T}{||a||_2^2}$$

so that

$$c = U\Sigma\upsilon \quad \text{with } \upsilon = \frac{V^T\, a^T}{||a||_2^2}. \tag{6}$$

This representation of c is similar to the "windowed" representation by Equation (5) which has the form

$$c = U\Sigma\upsilon^* \quad \text{with } \upsilon^* = \frac{V(I,:)^T\, a(I)^T}{||a(I)||_2^2}. \tag{7}$$

In this section a proof is given that the windowed representation (7) is a suitable generalization of (6) which has optimal reconstruction properties with respect to the channel window I.

The central ideas for the reconstruction of c are as follows: A single spectrum $a \in \mathbb{R}^{1\times n}$ should be associated with a single concentration profile $c \in \mathbb{R}^{k\times 1}$ so

that the rank-1 matrix ca is a best approximation of the spectral data D within the channel window which is determined by the index set I. The restriction $D|_I$ of D to the channel window I is given by

$$D\Big|_I = U\Sigma V(I,:)^T$$

and the rank-1 matrix ca with $a = tV^T$ reads

$$ca = \underbrace{U\Sigma\upsilon}_{c}\ \underbrace{tV^T}_{a}.$$

The restriction of ca to the channel window I is

$$ca\Big|_I = U\Sigma\upsilon\ tV(I,:)^T.$$

Under the assumption that the spectral signal in the channel window I is determined only by a single component, the difference function

$$f(\upsilon) = \left\|D\Big|_I - ca\Big|_I\right\|_F^2 = \left\|U\Sigma V(I,:)^T - U\Sigma\upsilon\ tV(I,:)^T\right\|_F^2$$

is to be minimized.

In Theorem 4.1 and in Corollary 4.2 a minimizer υ^* is determined. It is shown that the concentration profile $c = U\Sigma\upsilon^*$ is nonnegative if a and $D(:,I)$ are nonnegative. Finally, Lemma 4.3 presents an error estimation for $C(:,i) - c$.

Theorem 4.1. *Let $D \in \mathbb{R}^{k\times n}$ with $\operatorname{rank}(D) \geq z$ be given and let $U \in \mathbb{R}^{k\times z}$, $\Sigma \in \mathbb{R}^{z\times z}$ and $V \in \mathbb{R}^{n\times z}$ be the factors of a truncated singular value decomposition of D. Furthermore let $I \subset \{1,\ldots,n\}$ be the index set of the channel window. The vector $t \in \mathbb{R}^{1\times z}$ determines a nonzero spectrum $a = tV^T$ and it is assumed that the restriction of a to I, which is denoted by $a\big|_I =: a(I)$ does not vanish, i.e. $\|a(I)\| > 0$.*

Then the local reconstruction error

$$f(\upsilon) = \left\|U\Sigma V(I,:)^T - U\Sigma\upsilon tV(I,:)^T\right\|_F^2 \tag{8}$$

attains its minimum in $\upsilon^* \in \mathbb{R}^{1\times z}$ with

$$\upsilon^* = \frac{V(I,:)^T\,V(I,:)\,t^T}{\|V(I,:)\,t^T\|_2^2} = \frac{V(I,:)^T\,a(I)^T}{\|a(I)\|_2^2}. \tag{9}$$

With the associated optimal concentration profile $c = U\Sigma\upsilon^*$ the rank-1 reconstruction ca results in a rank-1 downgrade of D

$$\operatorname{rank}(D - ca) = \operatorname{rank}(D) - 1. \tag{10}$$

Proof. Due to the orthogonal invariance of the Frobenius-norm [16] the distance $f(\upsilon)$ can be simplified

$$\begin{aligned} f(\upsilon) &= \left\|\Sigma V(I,:)^T - \Sigma\upsilon tV(I,:)^T\right\|_F^2 \\ &= \sum_{j=1}^{z}\sum_{i\in I}\sigma_j^2 V_{ij}^2 - 2\sum_{j=1}^{z}\sigma_j^2\upsilon_j\sum_{i\in I}V_{ij}\sum_{l=1}^{z}t_l V_{il} \\ &\quad + \sum_{j=1}^{z}\sigma_j^2\upsilon_j^2\sum_{i\in I}\left(\sum_{l=1}^{z}t_l V_{il}\right)^2. \end{aligned}$$

The gradient vector has the form

$$\nabla f(\upsilon) = 0 - 2\Sigma^2 V(I,:)^T V(I,:)\,t^T + 2\|V(I,:)\,t^T\|_2^2\Sigma^2\upsilon$$

and $\nabla f(\upsilon^*) = 0$ as a necessary condition for an extremum results in (7). The Hessian of $f(\upsilon)$ reads

$$H_f(\upsilon) = 2\|V(I,:)\,t^T\|_2\,\Sigma^2 = 2\|a(I)\|_2\,\Sigma^2.$$

Since $\operatorname{rank}(D) \geq z$ the truncated matrix Σ has the maximal rank and together with $\|a(I)\| > 0$ the Hessian $H_f(\upsilon^*)$ is a positive definite matrix. Hence f attains its minimum in υ^*.

In order to prove (10) we need the following relation for the $n\times n$ identity matrix I and column vectors $a, b \in \mathbb{R}^n$: The matrix $I - ab^T$ is regular if and only if $a^Tb \neq 1$. This relation follows from the more general Sherman-Morrison-Woodbury formula [16]. For the given special case the proof is simple. If $I - ab^T$ is regular, then $(I - ab^T)a = a - a(b^Ta) \neq 0$ only if $b^Ta \neq 1$. To prove the other direction one has to check that the inverse of $I - ab^T$ is $(I + ab^T/(1 - b^a))$.

For t and υ^* it holds that

$$t\upsilon^* = \frac{t\,V(I,:)^T\,V(I,:)\,t^T}{\|V(I,:)\,t^T\|_2^2} = 1$$

so that

$$D - ca = D - U\Sigma\upsilon^* tV^T = U\Sigma(I - \upsilon^* t)V^T.$$

Since $I - \upsilon^* t$ is a singular matrix with the rank $z - 1$ the same holds for $D - ca$. □

The concentration profile c by Theorem 4.1 is nonnegative if D and a are nonnegative. This is proved next.

Corollary 4.2. *Let the assumptions of Theorem 4.1 be satisfied and let $D(:,I) \geq 0$ as well as $a = tV^T \geq 0$.*

Then the associated concentration $c = U\Sigma\upsilon^*$ is also nonnegative.

Proof. Since $D(:, I) \geq 0$ and $a(I) \geq 0$, it holds that

$$c = U\Sigma v^* = U\Sigma \frac{V^T(I,:)\, a(I)}{\|a(I)\|_2^2} = \frac{D(:,I)\, a(I)^T}{\|a(I)\|_2^2} \geq 0. \tag{11}$$

□

The factors a and c arise by construction from a local factorization within the channel window I. If $D = CA$ is the correct global factorization of D, then a and c should be (aside from scaling) recoverable as a certain row of A and as the associated column of C so that $a = A(i,:)$ and $c = C(:,i)$ for a suitable index i.

The difference between c (by the local construction) and the associated column $C(:,i)$ (as determined by the global factorization) depends on the computed a and on the absorption by the other species in the window I. An error-estimation for $c - C(:,i)$ is given in the following lemma.

Lemma 4.3. *Let the assumptions of Theorem 4.1 be satisfied, let m be the number of channel indices in I and let $a(I) = tV(I,:)^T$ and c as given in (7). Further let $D = CA$ be a feasible nonnegative factorization so that a equals $A(1,:)$ on the interval I, which is $a(I) = A(1,I)$. Finally let*

$$E = \sum_{i=2}^{s} C(:,i)A(i,I)$$

be the sum of all absorptions by all other components in the channel window I.

Then the error $\|c - C(:,1)\|_2$ is bounded from above as

$$\|c - C(:,1)\|_2 \leq \frac{\|E\|_2}{\|A(1,:)\|_2}.$$

Therein $\|E\|_2 = \max_{a \neq 0} \|Ea\|_2 / \|a\|_2$ is the spectral operator norm.

Proof. With $c = D(:,I)\, a(I)^T / \|a(I)\|_2^2$ by (11) and $a(I) = A(1:,I)$ it holds that

$$\begin{aligned} c &= \left(\sum_{i=1}^{s} C(:,i)\, A(i,I) \right) \frac{a(I)^T}{\|a(I)\|_2^2} \\ &= C(:,1) + E\, \frac{a(I)^T}{\|a(I)\|_2^2}. \end{aligned}$$

With respect to the Euclidean norm and the associated spectral norm one gets

$$\|c - C(:,1)\|_2 = \frac{\|E\, a(I)^T\|_2}{\|a(I)\|_2^2} \leq \frac{\|E\|_2}{\|a(I)\|_2}. \tag{12}$$

□

A useful consequence of (12) is the following: If the inner products of a with all other pure component spectra within the window I are equal to zero, then $c = C(:,1)$ and the pure component concentration profile is reproduced exactly.

In Remark 4.4 an estimate for the error $c - C(:,1)$ for the case that c is constructed by the non-windowed pseudo-inverse of t in the form $c = U\Sigma t^+$. According to our experience the windowed reconstruction of c by 4.1 provides the better results for the concentration profiles compared to the construction with the pseudo-inverse of t. The different approaches are compared in Section 8.2 for experimental data.

Remark 4.4. *Let the assumptions of Lemma 4.3 be satisfied. If the concentration profile c is computed by $c = U\Sigma t^+$ with the pseudo-inverse t^+ of t and $A(1,:) = a = tV^T$, then the error vector $c - C(:,1)$ reads*

$$c - C(:,1) = \frac{1}{\|A(1,:)\|_2^2} \sum_{i=2}^{s} C(:,i)A(i,:)A(1,:)^T. \tag{13}$$

Its Euclidean norm is bounded from below and above in the form

$$\alpha \leq \|c - C(:,1)\|_2 \leq \beta$$

with

$$\alpha = \frac{\max_{i=2,\ldots,s} \|C(:,i)\|_2 A(i,:)A(1,:)^T}{\|A(1,:)\|^2}, \tag{14}$$

$$\beta = \frac{\sum_{i=2}^{s} \|C(:,i)\|_2 A(i,:)A(1,:)^T}{\|A(1,:)\|^2}. \tag{15}$$

Proof. With $a = A(1,:) = tV^T$ and $a^+ = (tV^T)^T / \|tV^T\|_2^2$ it holds that

$$\begin{aligned} c &= U\Sigma t^+ = U\Sigma t^T / \|t\|_2^2 = U\Sigma V^T V t^T / \|tV^T\|_2^2 \\ &= D(tV^T)^+ = Da^+ = DA(1,:)^+ \\ &= \left(\sum_{i=1}^{s} C(:,i)A(i,:) \right) A(1,:)^+ \\ &= C(:,1) + \left(\sum_{i=2}^{s} C(:,i)A(i,:) \right) A(1,:)^+. \end{aligned}$$

This together with $A(1,:)^+ = A(1,:)^T / \|A(1,:)\|_2^2$ results in (13). The bounds (14) and (15) can easily be proved. □

The difference (13) vanishes if and only if $A(i,:)A(1,:)^T = 0$ for all spectra $i = 2, \ldots, s$. This condition is more restrictive compared to (12) where only a local condition with respect to the channel window I has to hold. Thus the error for $c - C(:,1)$ is expected to be larger compared to the estimate from Lemma 4.3.

5. Delimitation to alternative MCR methods

The PGA method is one among many alternative and well-established MCR methods. Explicitly we would like to mention the famous MCR-ALS toolbox [9], the resolving factor analysis (RFA) [30], the SIMPLISMA algorithm [22], the band target entropy minimization (BTEM) [21] and the pure component decomposition (PCD) [10].

All these methods focus on the computation of a single nonnegative factorization $D = CA$. However, due to the rotational ambiguity there are nearly always continua of nonnegative factorizations which can be computed by global MCR-methods and which can be represented by the so-called *Area of Feasible Solutions* (AFS). Computationally the AFS can be determined, e.g., by the FACPACK software [31, 32]. The strategy of the first-mentioned MCR methods is to filter out only a single factorization by means of soft constraints. A reliable MCR method should provide a good approximation of the chemically correct solution. The general approach of all these MCR-methods is to construct a cost function which depends on the transformation matrix T, see (1), and whose feasible matrix elements are associated with the AFS. However, the minimization of the soft-constrained cost function results in the desired matrix factors C and A. With this approach the computed solution strongly depends on the choice of the constraints. The numerical computation of the global minimum of the cost function is sometimes a difficult numerical problem as the numerical minimization may get stuck a local minimum.

5.1. Benefits of the PGA algorithm

A characteristic trait of the PGA algorithm is that the pure component spectra are computed step-by-step, whereas other methods like RFA or PCD compute all spectra (and concentration profiles) simultaneously. According to our experience, PGA is a robust algorithm for medium-to-strong perturbed spectral data. The method also works very well in case of systematic perturbations for example from a suboptimal baseline correction. Whenever MCR methods like RFA, MCR-ALS or PCD show difficulties in the simultaneous computation of the spectra and concentration profiles, the PGA algorithm can be applied in order to try a stepwise decomposition. Then PGA may extract single pure components and can be able to uncover correlations within highly overlapping peak groups. The technical reason for these abilities are the window analysis of the spectra and the reduced number of variables of the cost function.

The strategy underlying the PGA method shows some resemblance to the BTEM algorithm [21]. BTEM also allows a step-by-step extraction of the pure component spectra. However, the computational procedures of PGA and BTEM are different. Compared to BTEM the number of unknown variables is reduced from z to $z - 1$. Further, PGA works with a different construction of the associated concentration profile by using a windowed pseudo-inverse construction. In BTEM the pseudo-inverse is used for a global fit, whereas for PGA the potentially more stable local approximation from (5) is applied. See Lemma 4.3 and Remark 4.4.

5.2. Limitations of a step-by-step factorization

The higher robustness of PGA with respect to noise by means of a windowed step-by-step decomposition has to be paid with a partial loss of insight to the global correlations between the factors. In other words, step-by-step computed concentration profiles and spectra must not necessarily result in factors C and A with a small reconstruction error $D - CA$. Whenever the data does not include systematic perturbations or a considerable portion of noise one should first apply MCR methods like RFA, MCR-ALS or PCD.

6. PGA and the FACPACK software

The FACPACK software is a program package for the computation of multi-component factorizations and the area of feasible solutions for two- and three-component systems by means of the polygon inflation method [14, 15]. This software also allows the simultaneous representation of C and A and a step-by-step construction of the pure component factors by using arguments from the complementarity theory [31].

The peak group analysis (PGA) is a separate module of the forthcoming revision 1.2 of FACPACK, which is planned to be published in the second half of the year 2015. The current revision of the software can be downloaded from the web page

http://www.math.uni-rostock.de/facpack/

7. Analysis of a model problem

Next the PGA method is applied to a three-component model problem with a number of $k = 201$ spectra each with $n = 501$ channels. The simulated concentration profiles, the columns of C, and the pure component spectra, the rows of A, are shown in Figure

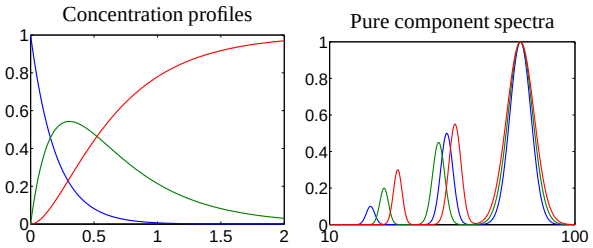

Figure 1: The concentration profiles, the columns of C, are shown together with the pure component spectra, the rows of A, for the model problem from Section 7. In a second step noise is added to the product $D = CA$.

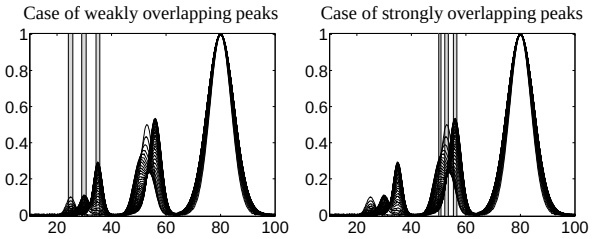

Figure 2: Window selection in the series of spectra with normal distributed noise. Left: Three windows positioned on the well-separated peaks, Right: Three windows positioned on the three strongly overlapping peaks.

1. The spectral data matrix is $D = CA$. Noise of different types, namely random noise and systematic noise or bias, are added to the data in order to demonstrate the capabilities of the PGA method. The perturbed data sets are shown in Figure 2 (case of random noise) and in the left subplot of Figure 5 (systematic noise). More information is given below.

7.1. Overlapping peaks

The pure component spectra are constructed in a way that each spectrum contains three peaks and each peak is a Gaussian. One peak for each spectrum is relatively isolated; these peaks are centered at $x_0 = 25$, $x_0 = 30$ and $x_0 = 35$. A second peak of each spectrum strongly overlaps with one peak of the other pure component spectra; the three peaks are centered at $x_0 = 50$, $x_0 = 53$ and $x_0 = 56$. A third peak of each spectrum is centered at $x_0 = 80$ and has the amplitude 1. Thus this peak completely overlaps with the other spectra. However, the three peak widths are slightly different. See the right subplot of Figure 1 for these three spectra.

7.2. PGA applied to windows on well-separated peaks

First PGA is applied to the data matrix D plus random noise (normal distributed noise with a standard

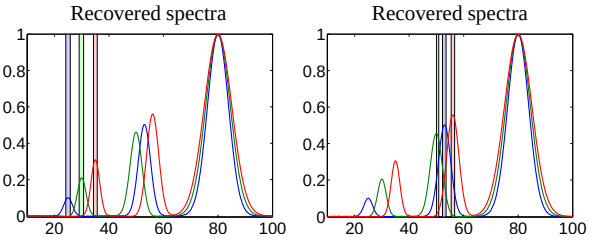

Figure 3: PGA results on the pure component spectra for data including random normal distributed noise. Left: Recovered pure component spectra for the case of isolated peaks, see left subplot in Figure 2. Right: Recovered pure component spectra for the case of strongly overlapping peaks, see right subplot in Figure 2. The original pure component spectra are plotted in black lines.

deviation of 0.002 and mean 0). As the model problem includes three independent components, three separate windows are selected for the calculation of the three pure component spectra. First, small windows are placed at the three isolated peaks at $x_0 = 25$, $x_0 = 30$ and $x_0 = 35$. These three narrow windows in gray color are shown in the left subplot of Figure 2. These windows are the basis for the identification of the associated pure component spectra.

The weights factors ω_i and γ_i, see Equation (4), are set to $\omega_1 = 0.1$, $\gamma_1 = 10$, $\gamma_2 = 1.5$ and $\omega_2 = \gamma_3 = \gamma_4 = 0$. The reconstruction is based on a number of $z = 3$ singular vectors. The computed pure component spectra are shown in Figure 3. The relative reconstruction error

$$e_i = \frac{\|A^{(\mathrm{orig})}(i,:) - a^{(\mathrm{PGA})}(i,:)\|_2}{\|A^{(\mathrm{orig})}(i,:)\|_2}, \qquad i = 1,2,3, \quad (16)$$

allows to compare the results with the original solutions. We get $e = (4.7 \cdot 10^{-3},\ 1.5 \cdot 10^{-2},\ 1.1 \cdot 10^{-2})$. Thus PGA works very well.

7.3. PGA for strongly overlapping peaks

Next we demonstrate the capability of the PGA to identify peak correlations and peak groups for the case of overlapping peaks. Therefore the three windows are placed around the three centers $x_0 = 50$, $x_0 = 53$ and $x_0 = 56$ which belong to strongly overlapping peaks. Once again the three pure component spectra are reconstructed from the noisy data, see Section 7.2 on the noise intensity. The window selection is plotted by gray bars in the right subplot of Figure 2.

For the PGA we used the weight factors $\omega_1 = 0.1$, $\gamma_1 = 10$, $\gamma_2 = 0.5$ and $\omega_2 = \gamma_3 = \gamma_4 = 0$. The number of singular vectors is $z = 3$. The solutions are presented in the right subplot of Figure 3. Even for these strongly overlapping signals PGA works very well and the correct single pure component spectra have been identified

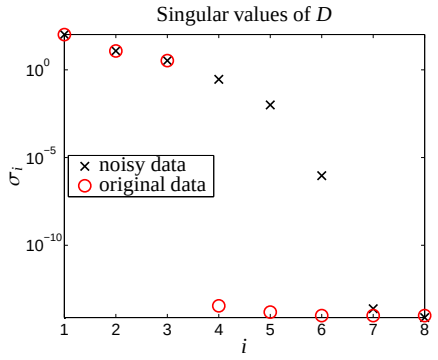

Figure 4: The singular values of the data matrix D including systematic noise ($\times$) and the singular values of the non-perturbed data ($\circ$). The non relatively large singular values σ_4, σ_5, σ_6 are a clear indicator of the presence of systematic noise (bias).

with errors less than one percent. The error vector (16) reads $e = (3.7 \cdot 10^{-3},\ 7.9 \cdot 10^{-3},\ 6.8 \cdot 10^{-3})$.

7.4. PGA and systematic noise

Next PGA is applied to the data including systematic noise. Therefore D is computed as

$$D_{i,j} = \sum_{\ell=1}^{3} C_{i,\ell} A_{\ell,j} + 0.005\left(2 - \frac{(x_j - 50 - \frac{i}{10})^2}{500}\right) \quad (17)$$

for $i = 1, \ldots, 201$ and $j = 1, \ldots, 501$. The series of spectra including these perturbations is shown in the left subplot of Figure 5. The first eight singular values of the data matrix D are shown in Figure 4. Due to the perturbation, see (17), the singular values σ_4, σ_5 and σ_6 are very different from the non-perturbed data. For non-perturbed data the matrix has the rank 3. Hence these singular values are equal to zero aside from rounding errors.

For the PGA computation the windows are located on the strongly overlapping peaks. A number of $z = 4$ singular vectors is used. The selected windows and the computed pure component spectra are presented in Figure 5.

Even in presence of systematic noise (bias) and for the strongly overlapping peaks it is possible to identify the correct pure component spectra with the PGA. The error vector reads $e = (2.8 \cdot 10^{-2},\ 3.7 \cdot 10^{-2},\ 5.1 \cdot 10^{-2})$.

8. Application to IR data from experimental studies of equilibria of rhodium and iridium hydroformylation catalysts

In this section the PGA is applied to FT-IR spectroscopic data on the formation of rhodium-carbonyl clusters and on the catalyst formation for the iridium-catalyzed hydroformylation; see [33, 10, 12] for experimental details. Further an application of the PGA to weak signals is presented in Section 8.4. In this section the frequency axis for all spectra is given in wavenumbers.

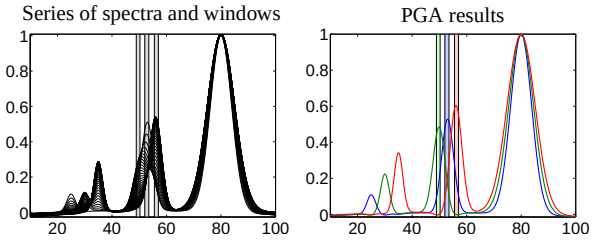

Figure 5: Left: The three PGA windows positioned on the strongly overlapping peaks. Right: The PGA results of the pure component spectra. Even in presence of systematic noise PGA can correctly identify the pure component spectra.

8.1. Rhodium-carbonyl complex formation from Rh(acac)(CO)$_2$

The displacement of the organic ligand acetylacetonate in Rh(acac)(CO)$_2$ by CO is an unwanted side reaction in the formation of a catalytically active rhodium-hydridocarbonyl complex. In this experiment at 303 K and 20 bar synthesis gas pressure (CO:H$_2$ = 1:1) the initial concentration of Rh(acac)(CO)$_2$ is $6.6 \cdot 10^{-4}\,\mathrm{mol\,L^{-1}}$ in the solvent cyclohexane. Under these conditions the "unwanted" rhodium carbonyl clusters $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$ together with Rh(acac)(CO)$_2$ are the dominating absorbing components. The spectrum of the solvent cyclohexane has been subtracted from the FT-IR spectroscopic data. A number of $k = 292$ spectra has been taken; each spectrum has $n = 1479$ channels. A subset of the series of spectra is shown in Figure 6.

In order to compute the pure component spectra of this system with $s = 3$ dominating components we have selected three separate channel windows for the application of the PGA. Each of these three windows $[2005.5,\ 2020.1]\,\mathrm{cm^{-1}}$, $[1875.1,\ 1893.9]\,\mathrm{cm^{-1}}$ and $[1812.7,\ 1823.4]\,\mathrm{cm^{-1}}$ contains only a single peak. The PGA has been applied three times with $z = 4$ and the weights $\omega_1 = 0.15$ (first run to determine the first component), $\omega_1 = 0.3$ (second run to determine the second component) and $\omega_1 = 0.15$ (third run for the third component). Further $\omega_2 = 0$ in all program runs. In all these case f_1 and f_2 according to Equation (4) have been used. Further, only the constraint function g_1 and g_2 have been used with the weight factors $\gamma_1 = 10$ and $\gamma_2 = 1$. This parameter selection has been used for all

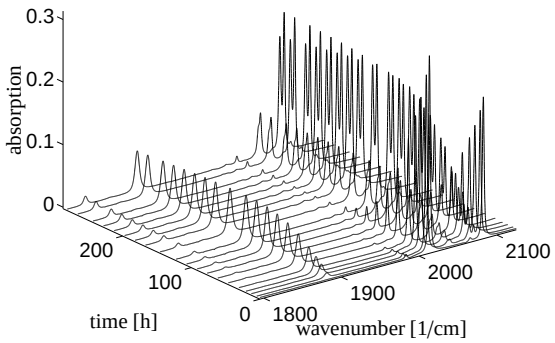

Figure 6: A subset of the $k = 292$ spectra for the formation of rhodium-carbonyl complexes from $Rh(acac)(CO)_2$.

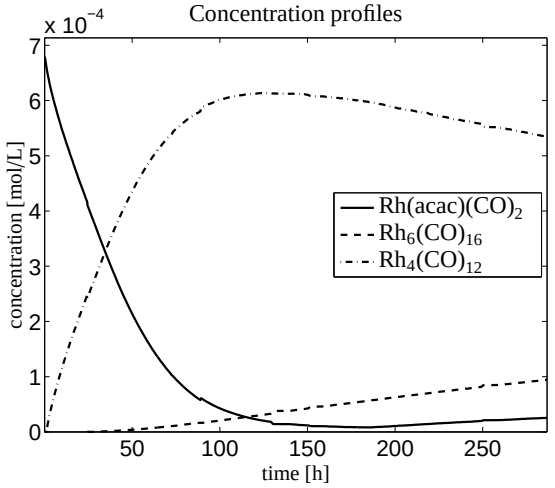

Figure 8: The concentration profiles (absolute values) for the $s = 3$ components by a global least-squares fit.

three PGA computations. The results, namely the pure component spectra of $Rh(acac)(CO)_2$, $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$ are shown in Figure 7.

The concentration profiles of these three components can be computed by the techniques introduced in Section 4; the use of Equation (9) is recommended. However, in the present situation the three spectra of the three-component system are known. Hence it is more stable to form the matrix A and then to solve a global least-squares problem on the full wavenumber interval in order to compute C. Thus all three concentration profiles are computed simultaneously. These profiles are shown in Figure 8.

8.2. Comparison of the different approaches to the computation of C

The most common way to compute a concentration profile which is associated with a single pure component spectrum a is to use the pseudo-inverse a^+ so that $c = Da^+$ or $C(:,i) = U\Sigma t^+$ if the ith concentration profile is considered. The benefit of computing $C(:,i)$ via v^* according to Theorem 4.1 is explained in Section 4. Next $C(:,i)$ is computed for the data set from Section 8.1 in three different ways:

1. Column-wise computation of $C(:,i) = U\Sigma v(i)^*$ with $v(i)^*$ by (9) for $i = 1,2,3$.
2. Column-wise computation of $C(:,i) = U\Sigma t(i)^+$ with the pseudo-inverses $t(i)^+$ for $i = 1,2,3$.
3. By a global (simultaneous) least squares fit in order to find C for known three pure component spectra given row-wise in A.

Figure 9 shows the results. The results for the approaches 1. and 3. are very similar, whereas the difference between the second and the third approach is only small for $Rh(acac)(CO)_2$. With respect to the Euclidean vector norm the distances

$$\varepsilon_i = \left\|C(:,i) - C^{(\text{global})}\right\|_2 / \left\|C^{(\text{global})}\right\|_2, \qquad i = 1,2,3,$$
$$\overline{\varepsilon}_i = \left\|\widetilde{C}(:,i) - C^{(\text{global})}\right\|_2 / \left\|C^{(\text{global})}\right\|_2, \qquad i = 1,2,3$$

with $C = U\Sigma v^*$ and $\widetilde{C} = U\Sigma t^+$ are as follows

$$\varepsilon = (0.0034,\ 0.0908,\ 0.0080),$$
$$\overline{\varepsilon} = (0.0856,\ 4.5919,\ 0.0812).$$

8.3. Equilibrium of iridium complexes

A detailed analysis of the equilibrium of iridium complexes for the hydroformylation of olefins has recently been published [12]. The equilibria of various hydridocarbonyltriphenylphosphine-iridium catalysts are analyzed at 373 K for varying partial pressure of carbon monoxide between $p(CO) = 10^{-2}$ to 3.9 MPa at a constant hydrogen partial pressure of $p(H_2) =$ 1.0 MPa. The hydrido complexes were formed from Ir(COD)(acac), with $c(Ir) = 5.0 \cdot 10^{-3}$mol L^{-1}, and 10 equivalents of PPh_3 under 2.0 MPa of synthesis gas [12]. Here we consider a sequence of $k = 47$ FT-IR spectra, and each spectrum is taken at $n = 913$ frequencies values in the interval $[1900, 2150]$cm^{-1}. The sequence of spectra is shown in Figure 10.

The PGA has been applied three times in order to extract the three dominant components. A particular challenge of this problem is that all pure component spectra are highly overlapping. The reconstruction is based on

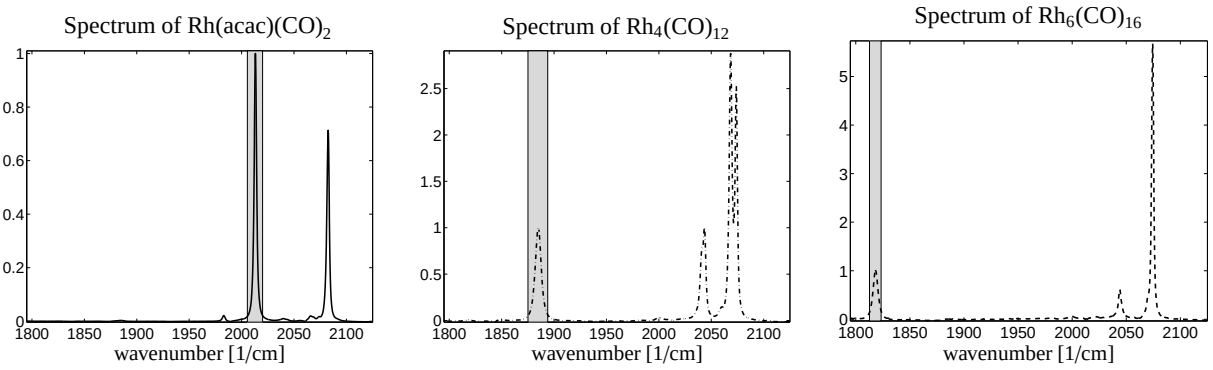

Figure 7: Results of a spectrum-by-spectrum computation with the PGA. The three selected channel windows are [2005.5, 2019.2] cm^{-1}, [1813.1, 1822.5] cm^{-1} and [1872.1, 1894.4] cm^{-1}. These channel windows are marked by a gray background. All spectra are normalized so that the maximum in the window equals 1.

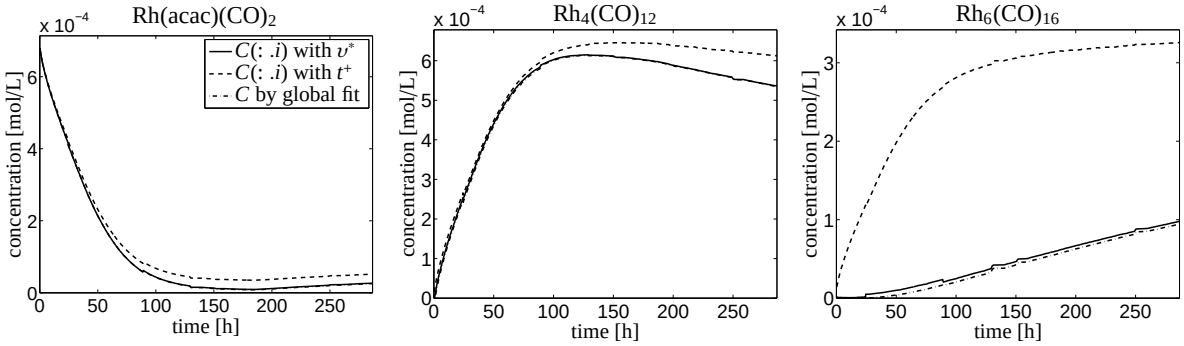

Figure 9: Comparison of the three different approaches to compute C for the three pure components of the data set from Section 8.1. The global fit provides the best results (by the solution of the highest dimensional least squares problem). The local (windowed) reconstruction with (9) gives very similar results which shows that the PGA approach works very well. These two concentration profiles are much better than $\widetilde{C} = U\Sigma t^{+}$ by means of the pseudo-inverse t^{+}. The latter approach which is associated with the solution of a least squares problem in a one-dimensional space results in a relatively poor approximation.

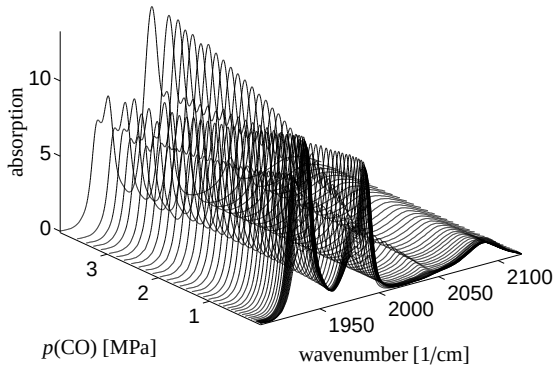

Figure 10: Series of $k = 47$ FT-IR spectra for an equilibrium mixture of iridium hydrido complexes registered during a variation of the carbon monoxide partial pressure, see Section 8.3.

the first $z = 3$ left and right singular vectors. The channel windows and the weight factors are as follows (for the third window $\omega_2 = 50$ appears to be large, but all spectra are relatively smooth so that the contribution to the target function is small)

window	ω_1	ω_2	γ_1	γ_2	γ_3	γ_4
[2038.7, 2045.2]	0.05	0	10	0.03	0	0
[1934.7, 1942.9]	0.05	0	10	0.03	0	0.12
[2081.3, 2088.1]	0.005	50	2	0.03	0	0

Three hydrido complexes have been identified, namely $HIr(CO)_3(PPh_3)$, $HIr(CO)_2(PPh_3)_2$ and $H_3Ir(CO)(PPh_3)_2$. The spectra for these three pure components are shown in Figure 11. Once again, the concentration profiles have been computed by solving a global least-squares problem on the full wavenumber interval since all three pure component spectra are available. The concentration profiles are shown in Figure 12.

8.4. PGA for weak peaks

In order to demonstrate the local amplification of weak peaks in the PGA the spectral data set from Section 8.1 is resumed. Now the wavenumber window $[1980, 1985]\,\mathrm{cm}^{-1}$ is taken which is associated with the index interval $I = [641, 649]$, see Figure 13. The peak in this window is very small compared to the maximal absorption since

$$\frac{\max_{i\in I} D(:,i)}{\max D} = 0.02.$$

Once again a number of $z = 4$ singular vectors are used. For f_1 and f_2 the weights are set to $\omega_1 = 0.05$ and $\omega_2 = 0$. For the optimization the constraint functions g_1 and g_2 with $\gamma_1 = 50$ and $\gamma_2 = 10$ are active. The weight factors have a very different size due to the fact that the normalization (2) affects only g_1 but not f_1 and f_2. What is finally found is the spectrum of the pure component $Rh(acac)(CO)_2$. The relative error of this solution called a_{weak} compared to the spectrum a_1, see Section 8.1 and the leftmost sub-figure in Figure 7, is about 1.2% since

$$\frac{\|\alpha a_{\mathrm{weak}} - a_1\|_2}{\|a_1\|_2} = 0.012$$

with an optimized scaling parameter $\alpha = 0.022$.

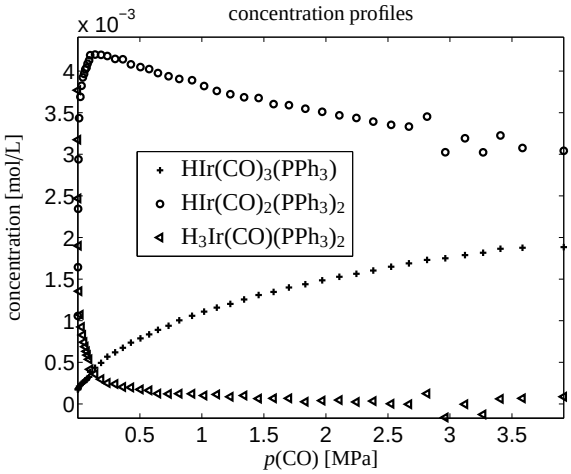

Figure 12: The pure concentration profiles for the three iridium complexes as computed by a global least-squares fit.

9. Conclusion

The Peak Group Analysis (PGA) has been presented as a numerical algorithm which allows a step-by-step computation of the pure component spectra from the initial spectral data set for the chemical mixture. A crucial requirement for a successful application of the PGA is that certain single peaks or isolated peak groups can be identified whose spectral profile is dominated by a single pure component. Then this peak or peak group is the starting point for a local optimization procedure which results in a global spectrum of a pure component. This global spectrum more or less reproduces the initial peak or peak group. The mathematical algorithm of the PGA is based on minimization of a target function to which various weighted soft constraints are added. We have also shown for experimental spectral data from the rhodium- and iridium-catalyzed hydroformylation process that the PGA is a useful tool for structure elucidation.

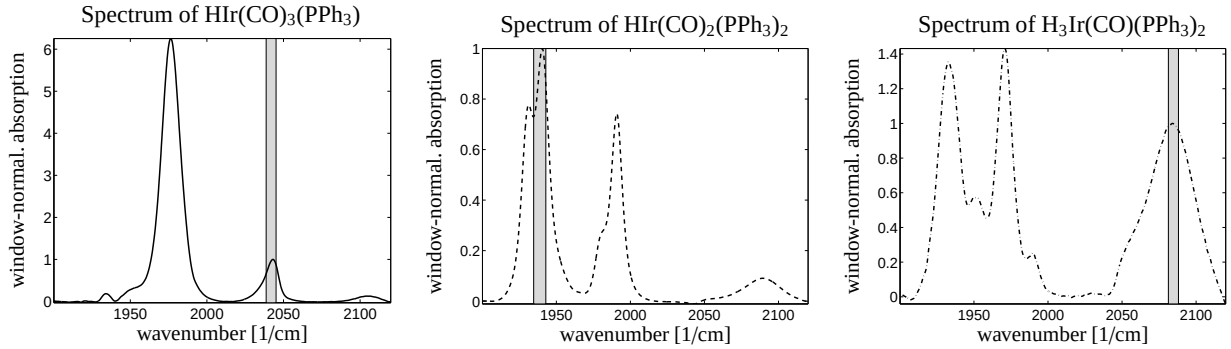

Figure 11: Results of a spectrum-by-spectrum computation with the PGA for the analysis of equilibria of iridium complexes. The three intervals [2039, 2045] cm^{-1}, [1935, 1943] cm^{-1} and [2081, 2088] cm^{-1} have been selected for three runs of PGA. The channel windows are marked by a gray background. Each spectrum has been normalized so that the maximum in the window equals 1. The band at 1933cm^{-1} of the rightmost spectrum does not belong to the trihydride complex. It is perhaps an artifact due to low CO partial pressures.

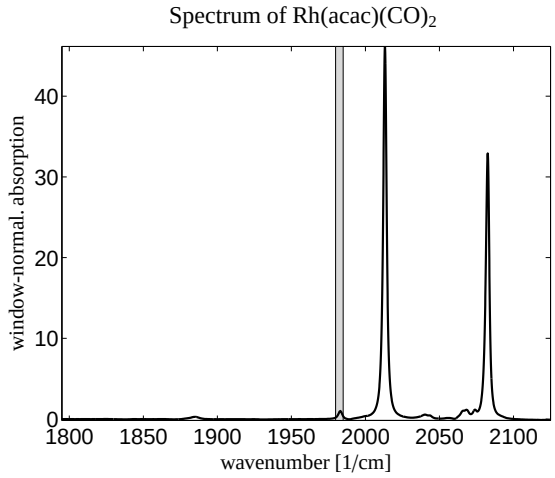

Figure 13: Application of the PGA to a weak signal (gray bar). The resulting spectrum of $Rh(acac)(CO)_2$ reproduces the spectrum which is shown leftmost in Figure 7 with a relative error of about 4%.

For the authors of the present paper the PGA results from of a multi-annual interdisciplinary research cooperation of catalytic chemists, from the Leibniz institute for Catalysis, with numerical mathematicians. The algorithm of the PGA has grown out of the desire to have a reliable computational method which can identify the correlation of single peaks (within the series of spectra of a multicomponent system) to the remaining part of the spectrum of a certain pure component.

References

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

[2] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, and M. Maeder. On rotational ambiguity in model-free analyses of multivariate data. *J. Chemom.*, 20(6-7):302–310, 2006.

[3] R. Rajkó. Computation of the range (band boundaries) of feasible solutions and measure of the rotational ambiguity in self-modeling/multivariate curve resolution. *Anal. Chim. Acta*, 645(1–2):18–24, 2009.

[4] S.D. Brown, R. Tauler, and B. Walczak. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Vol. 1-4*. Elsevier Science, 2009.

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

[6] E.R. Malinowski. Window factor analysis: Theoretical derivation and application to flow injection analysis data. *J. Chemom.*, 6(1):29–40, 1992.

[7] R. Manne. On the resolution problem in hyphenated chromatography. *Chemom. Intell. Lab. Syst.*, 27(1):89–94, 1995.

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

[9] J. Jaumot, R. Gargallo, A. de Juan, and R. Tauler. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in {MATLAB}. *Chemom. Intell. Lab. Syst.*, 76(1):101–110, 2005.

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

[11] B. Heaton. *Mechanisms in Homogeneous Catalysis: A Spectroscopic Approach*. John Wiley & Sons, 2006.

[12] 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.

[13] 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. *Vibrational spectroscopy*, 41(1):101–111, 2006.

[14] M. Sawall, C. Kubis, D. Selent, A. Börner, and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications. *J. Chemom.*, 27:106–116, 2013.

[15] M. Sawall and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and FAC-PACK implementation. *J. Chemom.*, 28:633–644, 2014.

[16] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2012.

[17] H. Minc. *Nonnegative matrices*. John Wiley & Sons, New York, 1988.

[18] H. Abdollahi and R. Tauler. Uniqueness and rotation ambiguities in Multivariate Curve Resolution methods. *Chemom. Intell. Lab. Syst.*, 108(2):100–111, 2011.

[19] A. Golshan, H. Abdollahi, and M. Maeder. Resolution of Rotational Ambiguity for Three-Component Systems. *Anal. Chem.*, 83(3):836–841, 2011.

[20] M. Biggs, A. Ghodsi, and S. Vavasis. Nonnegative Matrix Factorization via Rank-one Downdate. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 64–71, New York, NY, USA, 2008. ACM.

[21] 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.

[22] W. Windig and D. A. Stephenson. Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach. *Anal. Chem.*, 64(22):2735–2742, 1992.

[23] A. Bogomolov and M. Hachey. Application of {SIMPLISMA} purity function for variable selection in multivariate regression analysis: A case study of protein secondary structure determination from infrared spectra. *Chemom. Intell. Lab. Syst.*, 88(1):132–142, 2007. Special Issue: WSC-5 2006 Special Issue: WSC-5 2006.

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

[25] K. Sasaki, S. Kawata, and S. Minami. Constrained nonlinear method for estimating component spectra from multicomponent mixtures. *Applied Optics*, 22(22):3599–3603, 1983.

[26] W. Chew, E. Widjaja, and M. Garland. Band-target entropy minimization (BTEM): An advanced method for recovering unknown pure component spectra. Application to the FT-IR spectra of unstable organometallic mixtures. *Organometallics*, 21(9):1982–1990, 2002.

[27] P.J. Gemperline and E. Cash. Advantages of soft versus hard constraints in self-modeling curve resolution problems. Alternating least squares with penalty functions. *Anal. Chem.*, 75:4236–4243, 2003.

[28] H. Kim, H. Park, and L. Eldén. Non-negative tensor factorization based on alternating large-scale non-negativity-constrained least squares. *Proceedings of the 7th IEEE international conference on Bioinformatics & Bioengineering (IEEE BIBE 2007)*, 2:1147–1151, 2007.

[29] J. Dennis, D. Gay, and R. Welsch. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7:348–368, 1981.

[30] C. Mason, M. Maeder, and A. Whitson. Resolving Factor Analysis. *Anal. Chem.*, 73(7):1587–1594, 2001.

[31] M. Sawall and K. Neymeyr. On the area of feasible solutions and its reduction by the complementarity theorem. *Anal. Chim. Acta*, 828:17–26, 2014.

[32] A. Jürß, M. Sawall, and K. Neymeyr. On generalized Borgen plots. I: From convex to affine combinations and applications to spectral dataSpectra. *Journal of Chemometrics*, 29(7):420–433, 2015.

[33] C. Kubis, D. Selent, M. Sawall, R. Ludwig, K. Neymeyr, W. Baumann, R. Franke, and A. Börner. Exploring between the extremes: Conversion dependent kinetics of phosphite-modified hydroformylation catalysis. *Chem. Eur. J.*, 18(28):8780–8794, 2012.