Pure component spectral recovery and constrained matrix factorizations: Concepts and applications

Klaus Neymeyr[a], Mathias Sawall[a], Dieter Hess[b]

[a] *Universität Rostock, Institut für Mathematik, Universitätsplatz 1, 18055 Rostock, Germany*
[b] *Evonik Oxeno GmbH, Paul-Baumann Strasse 1, 45722 Marl, Germany*

Abstract

We present new ideas underlying a self-modelling factor analytical method which allows to extract pure component spectra and the associated concentration profiles from a set of spectroscopic measurements. The usefulness of the method is demonstrated and compared with established tools for model problems and for a system from catalytic hydroformylation by Rhodium complexes both with overlapping component spectra. Self-modelling methods tend to minimize the overlap of the recovered spectra, which can result in an unwanted distortion of the spectra and concentration profiles. For strongly overlapping spectra a penalty condition on a specific singular value of the absorptivity matrix factor and a global decomposition approach are appropriate tools to construct improved factorizations.

Key words: chemometrics, factor analysis, hydroformylation, pure component decomposition, spectral recovery.

1. Introduction

Chemometrics is the science of extracting from chemical data (like spectroscopic data) useful information on the explored chemical system by means of numerical methods. A typical and challenging problem is a pure component decomposition of a chemical multicomponent system in which a chemical reaction is monitored by a sequence of spectroscopic measurements; e.g. absorbance spectroscopy like infrared (IR, FT-IR) or ultraviolet (UV) spectroscopy.

The aim of this paper is to present a numerical method which allows to extract from spectroscopic measurements of a multi-component chemical system the spectra of the underlying pure components together with the associated concentration profiles.

The paper is organized as follows. In the remaining part of Sec. 1 the matrix formulation of the Lambert-Beer law is stated. In Sec. 2 the spectral recovery problem and its mathematical background are introduced. In Sec. 3 the Pure Component Decomposition algorithm is presented. In Sec. 4 several spectral recovery algorithms are applied to a model problem and FT-IR spectra from Rhodium-catalyzed hydroformylation.

1.1. The Lambert-Beer law in matrix form

In its most simple (scalar) form the Lambert-Beer law states that the absorbance of light depends on the product of the absorptivity, the concentration of the absorbing species and on the path length, which is assumed constant in the following. This idealized form of the Lambert-Beer law holds only in the absence of error sources like noise and nonlinearities. For any mixture of s absorbing species the total absorbance is (once again in an idealized form) the linear superposition of the absorbances of the participating species. Further one can consider a number of k spectra (in time) from a (reacting) chemical system; k can also be the total number of spectra gained from repeated experiments. These k spectra are given at n frequencies (or spectral channels) and can be inscribed on the rows of the absorbance matrix A. Then the matrix form of the Lambert-Beer law reads

$$A = C\hat{A}. \tag{1}$$

$A \in \mathbb{R}^{k\times n}$ Absorbance matrix; k measurements, n frequencies,

$C \in \mathbb{R}^{k\times s}$ Concentration matrix; the jth row gives the concentrations of the species $1,\ldots,s$ for the jth measurement,

$\hat{A} \in \mathbb{R}^{s\times n}$ Absorptivity matrix; the jth row contains the absorptivity vectors (spectra), $j = 1,\ldots,s$.

...

The bilinear model (1) is the basis for the subsequent factorization analysis in which we try to recover for given (measured) A the unknown factors C and $\hat{A}$.

2. Model-Free Analysis of Spectral Data

Spectral measurements can be used to fit a mechanistic model to the measurement. This is called a *model-based analysis*. Typically the degrees of freedom of the model (like kinetic constants or scaling constants if spectra from spectral libraries are fitted to the problem) are determined in a least-squares sense. This means that the Euclidean norm of the residual, which is the difference of the measured data and the fitted-model data, is minimized. For a survey on model-based chemometrical methods see Chapter 4 in [1] or [2].

In contrast to a model-based spectrum analysis we consider a *model-free analysis* here. The model-free approach is also called a *self-modelling curve resolution technique*, a name which goes back to the early work of Lawton and Sylvestre [3, 4]. Recent references on factor analytical and model-free methods are [1, 5].

2.1. The spectral recovery problem

The spectral recovery problem is a so-called *inverse problem*, namely to find for a given measurement the generating factorization (1) without any *a priori* knowledge of the components, pure component spectra and/or the concentration profiles. In chemometrics a solution tool for the spectral recovery problem is often called a self-modelling method. We start with the following matrix factorization problem.

Problem 1 (Factorization problem)**.** *For a given matrix $A \in \mathbb{R}^{k\times n}$ find an integer $s \le \min\{k,n\}$ and factors $C \in \mathbb{R}^{k\times s}$, $\hat{A} \in \mathbb{R}^{s\times n}$ so that*

$$f(C,\hat{A}) = \|A - C\hat{A}\|_F$$

is minimized. ($\|\cdot\|_F$ denotes the Frobenius norm [6].)

A solution of the factorization problem can always be constructed by means of the singular value decomposition, see Section 2.3. The matrix factors C, $\hat{A}$ of the solution are usually called *abstract factors*. According to (1) the columns of C should describe the concentration profiles and the rows of $\hat{A}$ should form the pure component spectra. Unfortunately these abstract factors most often do not comply with these requirements. It is not even guaranteed that the components of C are non-negative numbers.

Therefore we reformulate the factorization problem by adding several penalty terms. These weighted penalty terms allow us to impose various restrictions on the solution, e.g. componentwise non-negativeness of the matrix factors, smoothness of the absorptivity/concentration functions and further restrictions. For a discussion on the use of penalty functions versus hard constraints for self-modelling curve resolution techniques see Gemperline and Cash [9].

Problem 2 (Spectral recovery problem)**.** *For a given matrix $A \in \mathbb{R}^{k\times n}$ find an integer $s \le \min\{k,n\}$ and factors $C \in \mathbb{R}^{k\times s}$, $\hat{A} \in \mathbb{R}^{s\times n}$ so that*

$$f(C,\hat{A}) = \|A - C\hat{A}\|_F + \sum_{i=1}^{p} \gamma_i g_i(C,\hat{A}) \qquad (2)$$

is minimized.
The p penalty functions g_i are weighted by the (small) regularization parameters $\gamma_i \ge 0$.

The minimization of the spectral recovery problem balances the trade-off between the approximation error $\|A - C\hat{A}\|_F$ and the m constraints. Details on the penalty functions are given in Section 3.

2.2. A two-component model problem

Next a model problem with highly overlapping spectra is introduced. In Sec. 4 spectral recovery methods are applied to this (and further) problems.

For the two components we assume the following spectra (the frequency coordinate is denoted by ν; the frequency range has arbitrarily been set to $\nu \in [0, 500]$)

$$a_1(\nu) = 3\exp(-\frac{(\nu-200)^2}{100}) + \frac{3}{2}\exp(-\frac{(\nu-250)^2}{100}) + \frac{3}{2}\exp(-\frac{(\nu-150)^2}{100}),$$

$$a_2(\nu) = 2\exp(-\frac{(\nu-50)^2}{30000}) + 1.3\exp(-\frac{(\nu-200-\gamma)^2}{1000}),$$

with $\gamma \in [0, 50]$. The parameter γ controls the overlapping of the spectra and the hardness of the spectral reconstruction problem. See Fig. 1 with $\gamma = 20$. The correlation coefficient between $a_1(\nu)$ and $a_2(\nu)$ on the discrete grid $\nu = 1, 2, \ldots, 500$ is shown in Fig. 2; for $\gamma = 0$ the correlation takes its maximum being about 0.315.

We assume a chemical reaction which degrades the first component and forms the second component. The concentration profiles are assumed to be

$$c_1(t) = 1 - c_2(t),$$
$$c_2(t) = \frac{\exp(rt) - 1}{10 + \exp(rt)},$$

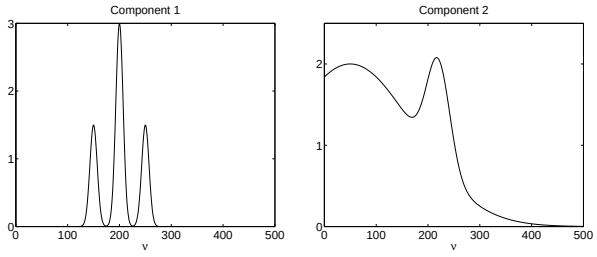

Figure 1: Absorbance spectra $a_1(\nu)$, $a_2(\nu)$.

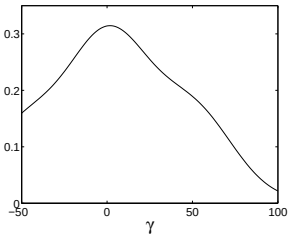

Figure 2: Correlation of $a_1(\nu)$ and $a_2(\nu, \gamma)$ for $\gamma \in [-50, 100]$.

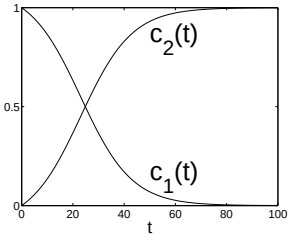

Figure 3: Concentration profiles of components 1 and 2.

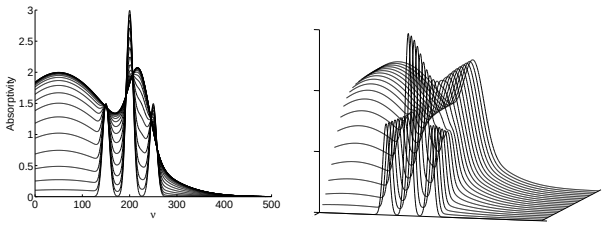

Figure 4: 2D/3D sequence of spectra.

with $t \in [0, 100]$ and $r = 0.1$; see Fig. 3.

According to Lambert-Beer's law the total absorption of the system at time t and at the wavelength ν is

$$A(t, \nu) = c_1(t)a_1(\nu) + c_2(t)a_2(\nu).$$

See Fig. 4 for a 2D/3D plot with $\gamma = 20$.

For a discrete time-frequency grid of k by n nodes the components of the absorbance matrix $A \in \mathbb{R}^{k \times n}$ are

$$\begin{aligned} A_{ij} &= c_1(t_i)a_1(\nu_j) + c_2(t_i)a_2(\nu_j) \\ t_i &= t_0 + i\delta_t, \qquad i = 1, \ldots, k, \\ \nu_j &= \nu_0 + j\delta_\nu, \qquad j = 1, \ldots, n, \end{aligned} \tag{3}$$

with δ_t, δ_ν being the time/frequency grid-widths.

2.3. The singular value decomposition and abstract factors

To solve the matrix factorization problem from Section 2.1 we start with a *singular value decomposition* [6] of $A \in \mathbb{R}^{k \times n}$. Without restriction of generality we assume that $k \leq n$ (the case $k > n$ can be treated analogously). The singular value decomposition of A reads

$$A = \bar{U}\bar{\Sigma}\bar{V}^T$$

with orthogonal matrices $\bar{U} \in \mathbb{R}^{k \times k}$, $\bar{V} \in \mathbb{R}^{n \times n}$ and a k by n diagonal matrix

$$\bar{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 & 0 & \ldots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_k & 0 & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{k \times n}.$$

The diagonal elements are the non-negative singular values with $\sigma_1 \geq \ldots \geq \sigma_k$. The number of non-zero singular values is the rank of A. Due to rounding errors the rank of A cannot be determined numerically in a stable way. Numerically one considers the ϵ-rank of a matrix, denoted by $\text{rank}_\epsilon(A)$, being the number of singular values greater or equal to ϵ.

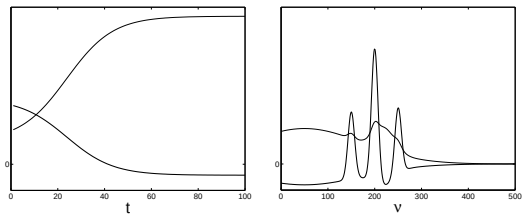

Figure 5: Abstract factors. Left: Concentration profiles. Right: Spectra.

2.3.1. Solution of the factorization problem

Next a solution of the factorization problem 1 (Section 2.1) is described. It is based on the singular value decomposition and a proper transformation of the matrix factors; for the underlying concepts of factor analysis and several references see Chapter 3 in Malinowski [5].

With $s = \text{rank}(A)$ let $u_i \in \mathbb{R}^k$ be the column vectors of $\overline{U} = (u_1, \ldots, u_k)$ and $\mathrm{v}_i \in \mathbb{R}^n$ the column vectors of $\overline{V} = (\mathrm{v}_1, \ldots, \mathrm{v}_n)$. Further let

$$U := [u_1, \ldots, u_s] \in \mathbb{R}^{k\times s},\ V := [\mathrm{v}_1, \ldots, \mathrm{v}_s] \in \mathbb{R}^{n\times s},$$
$$\Sigma := \text{diag}(\sigma_1, \ldots, \sigma_s).$$

Then $A = U\Sigma V^T$ is a *low-rank representation* of A and $C = U\Sigma$ and $\hat{A} = V^T$ solves the factorization problem.

If $\text{rank}_\epsilon(A) = s$, then $U\Sigma V^T$ is only a *low-rank approximation*. The error with respect to the spectral norm is

$$f(C, \hat{A}) = \|A - U\Sigma V^T\|_2 = \sigma_{s+1}.$$

The error σ_{s+1} is the smallest possible error for any $C \in \mathbb{R}^{k\times s}$ and $\hat{A} \in \mathbb{R}^{s\times n}$. As noted in Sec. 2.1 these abstract factors do not reconstruct the chemical data satisfyingly. For the model problem from Section 2.2 with $\gamma = 20$ the poor reconstruction is shown in Fig. 5. Both the concentrations and the absorptivities have negative components.

2.3.2. The spectral recovery problem

Assuming A to be formed by Lambert-Beer's law (1), i.e. $A = C\hat{A}$, then a reconstruction of C and $\hat{A}$ can be recovered only from A by means of the singular value decomposition. The key equation is

$$A = U\Sigma V^T = \underbrace{(U\Sigma S)}_{=C}\ \underbrace{(S^{-1}V^T)}_{=\hat{A}}. \qquad (4)$$

where a matrix pair $S, S^{-1} \in \mathbb{R}^{s\times s}$ is inserted. The existence of the matrix S can be shown as follows.

Lemma 2.1. *Let $A = C\hat{A}$ and let $A = U\Sigma V^T$ be the low rank representation of A with $s = \text{rank}(A)$. Then a regular matrix $S \in \mathbb{R}^{s\times s}$ exists so that*

$$C = U\Sigma S, \quad \hat{A} = S^{-1}V^T.$$

Proof. Since $A = C\hat{A}$ and $A = U\Sigma V^T$ the images of the mappings

$$U\Sigma : \mathbb{R}^s \to \mathbb{R}^k;\ x \mapsto U\Sigma x, \quad C : \ \mathbb{R}^s \to \mathbb{R}^k;\ y \mapsto Cy$$

coincide and have the dimension s, since $U\Sigma$ is of rank s. Hence a regular matrix $S \in \mathbb{R}^{s\times s}$ exists with $C = U\Sigma S$. This yields

$$0 = U\Sigma V^T - C\hat{A} = U\Sigma(V^T - S\hat{A}).$$

Further $U\Sigma \in \mathbb{R}^{k\times s}$ with $s \le k$ has the maximal rank s so that all columns of $V^T - S\hat{A}$ must be equal to zero. Thus $\hat{A} = S^{-1}V^T$. □

3. Pure Component Decomposition Algorithm

Next we describe a newly developed software for solving the spectral reconstruction problem (Section 2.1), called *Pure Component Decomposition* (PCD). In Sec. 4 numerical results for PCD are compared with the results which have been obtained by various software packages being available (in part commercially). The main characteristics of the PCD software (which makes this code to some extent different from formerly existing codes) are as follows: 1. PCD computes the matrix factors (the concentration and the absorptivity matrix) simultaneously in order to determine a stable solution with a well-balanced final residual. 2. Whenever possible (depending on the quality of the spectral data) PCD aims to compute global (non-band-targeted) factorizations. 3. PCD uses various coupled minimization tools (gradient iteration, quasi-Newton schemes and genetic algorithms) to solve the constrained minimization problem. 4. A measure for the linear independence of the computed spectra is taken as a penalty function for the optimization.

To introduce the PCD algorithm we start from a generalization of (4). Instead of (S, S^{-1}) a matrix pair (T, T^+) is used with T^+ being the pseudo-inverse [6] of T. This yields

$$A = U\Sigma T^+ T V^T$$

with $T \in \mathbb{R}^{s\times z}$ and $s \le z \le \min\{k, n\}$. Then $C = U\Sigma T^+$ and $\hat{A} = TV^T$. This amounts to a reconstruction of the concentration profiles by z left singular vectors.

An even more general approach is to consider

$$A = U\Sigma RTV^T$$

with $R \in \mathbb{R}^{z\times s}$ and $T \in \mathbb{R}^{s\times z}$. The components of R and T are taken as the degrees of freedom for the minimization of the Lagrange function of the spectral recovery problem underlying PCD

$$f(R,T) = \|A - \underbrace{U\Sigma R}_{C}\underbrace{TV^T}_{\hat{A}}\|_F + \sum_{i=1}^{p}\gamma_i g_i(C,\hat{A}).$$

Numerical minimization of f is achieved by means of the adaptive nonlinear least-squares algorithm NL2SOL [7, 8]. The penalty conditions are activated stepwise in the course of the minimization procedure and the positive weights γ_i have been adapted to each problem class (like UV-VIS, IR).

The penalty functions g_i are constructed to minimize the following (unwanted) traits of the solution:

1. Negative components of $\hat{A}$ and C.
2. Non-smooth solutions (Tychonoff regularization by the Euclidean norms of the first and second discrete derivatives of $\hat{A}$ and C).
3. A large discrete integral of the spectra (columns of $\hat{A}$).
4. The symmetric Kullback-Leibler divergence of $\hat{A}$, see [10].
5. The Shannon entropy of $\hat{A}$, see [11].
6. The inverse $1/\sigma_s(\hat{A})$ of the sth singular value of $\hat{A}$.

The constraint on the inverse of the sth singular value $\sigma_s(\hat{A})$ is a measure for the linear independence of the rows of $\hat{A}$. This constraint improves the separability of mixture spectra with highly overlapping bands, cf. Sec. 4.

Data preprocessing is an essential topic. Especially for the case of FT-IR spectra, the subtraction of the background yields distorted negative spectra. To handle those spectra the algorithm is combined with a (polynomial) baseline correction. Band targeting as used in [12] is of minor importance in PCD; whenever possible global decompositions are used in order to gain reliable spectral factorizations.

4. Numerical results

Various software tools for spectral recovery have been applied to the two-component model problem introduced in Sec. 2.2. Numerical results are given for the *Pure Component Decomposition* (PCD) code presented in Sec. 3, for the *Band Target Entropy Minimization* (BTEM) code of Garland et. al. [11, 13, 14, 15] using the implementation [12] and the *Nonnegative matrix factorization/Alternating Nonnegativity Constrained Least Squares* (NMF/ANLS) algorithm of Kim and Park [16]. Further computations are carried out with the SPECFIT software [17, 18] and the *Spectral Isolation Factor Analysis* (SPEXFA) Matlab Toolbox for the chemical factor analysis by Malinowski et. al. [5, 19].

γ		PCD	BTEM	NMF/ANLS	SPECFIT	SPEXFA
0	ε_1	$8.3\cdot10^{-6}$	$3.7\cdot10^{-6}$	$5.0\cdot10^{-1}$	$4.3\cdot10^{-7}$	$2.4\cdot10^{-15}$
	ε_2	$3.4\cdot10^{-4}$	$2.9\cdot10^{-2}$	$6.6\cdot10^{-4}$	$3.3\cdot10^{-4}$	$3.3\cdot10^{-4}$
10	ε_1	$1.0\cdot10^{-5}$	$3.5\cdot10^{-6}$	$7.0\cdot10^{-1}$	$4.0\cdot10^{-7}$	$2.3\cdot10^{-15}$
	ε_2	$3.1\cdot10^{-4}$	$8.0\cdot10^{-3}$	$1.0\cdot10^{-3}$	$3.3\cdot10^{-4}$	$3.3\cdot10^{-4}$
20	ε_1	$2.0\cdot10^{-5}$	$2.8\cdot10^{-6}$	$8.7\cdot10^{-1}$	$4.3\cdot10^{-7}$	$1.6\cdot10^{-15}$
	ε_2	$4.8\cdot10^{-4}$	$1.2\cdot10^{-2}$	$4.0\cdot10^{-4}$	$3.4\cdot10^{-4}$	$3.4\cdot10^{-4}$
30	ε_1	$6.1\cdot10^{-6}$	$3.6\cdot10^{-6}$	$7.3\cdot10^{-1}$	$3.4\cdot10^{-7}$	$1.3\cdot10^{-15}$
	ε_2	$5.1\cdot10^{-4}$	$6.6\cdot10^{-3}$	$4.2\cdot10^{-4}$	$3.3\cdot10^{-4}$	$3.3\cdot10^{-4}$
40	ε_1	$6.3\cdot10^{-6}$	$3.7\cdot10^{-6}$	$6.0\cdot10^{-1}$	$3.0\cdot10^{-7}$	$2.4\cdot10^{-15}$
	ε_2	$5.1\cdot10^{-4}$	$1.7\cdot10^{-3}$	$4.0\cdot10^{-4}$	$3.3\cdot10^{-4}$	$3.3\cdot10^{-4}$
50	ε_1	$6.8\cdot10^{-6}$	$4.7\cdot10^{-6}$	$5.3\cdot10^{-1}$	$3.2\cdot10^{-7}$	$2.0\cdot10^{-15}$
	ε_2	$1.0\cdot10^{-2}$	$3.1\cdot10^{-2}$	$4.0\cdot10^{-4}$	$3.3\cdot10^{-4}$	$3.3\cdot10^{-4}$

Table 1: Reconstruction errors of the absorptivities $\hat{A}_{i,:}^{(\text{METHOD})}$.

We take $k = 100$ single spectra each with $n = 501$ frequencies and $s = z = 2$ for BTEM and PCD. SPECFIT uses $z = 2$ and $s = 3$. (The parameters z and s are described in Sec. 3). For each value of γ the BTEM code has been started twice to find the best possible results for the two components. This yields an accurate absorbance spectrum of component 1 (the triplet signal). For $\gamma = 0, 10$ the band targets were 195–205 (2nd derivative constraint) for the first component and 195–205 (integrated area and 2nd derivative constraints) for the second component. For $\gamma \geq 20$ we substituted the band targets for the first component by 48–54. No band targets are required for PCD; a global decomposition is possible.

The quality of the factorizations is measured by the maximum norm [6] of the difference vector of the (normalized) exact solutions from Sec. 2.2 and the (normalized) computational results gained by the methods PCD, BTEM, NMF/ANLS, SPECFIT and SPEXFA.

$$\varepsilon_i^{(\text{METHOD})} = \left\|\hat{A}_{i,:}^{(\text{METHOD})} - \hat{A}_{i,:}^{(\text{orig.})}\right\|_\infty, \quad i = 1,2,$$
$$\delta_i^{(\text{METHOD})} = \left\|C_{i,:}^{(\text{METHOD})} - C_{i,:}^{(\text{orig.})}\right\|_\infty, \quad i = 1,2.$$

The index $i = 1, 2$ denotes the two components of the model problem. The reconstruction errors of the pure component spectra are listed in Table 1. SPECFIT, SPEXFA, BTEM and PCD show the best results. The NMF/ANLS algorithm correctly finds a non-negative factorization with a residual being in the range of the machine precision. Hence the NMF/ANLS algorithm

works successfully. Nevertheless, the found factors are only coarse approximations of the original data.

An analogous result holds for the concentration profiles, see Table 2. SPECFIT and SPEXFA clearly profit from the structure of the model problem, namely that the first and the last spectrum of the spectra series are very good approximations of the pure component spectra. (The SPECFIT code uses the *Evolving Factor Analysis* (EFA), cf. [20].) To avoid such an unwanted access to the pure component data we select a subset of 31 spectra (indexes $10, \ldots, 40$) from the original sequence of $k = 100$ spectra, see Table 3. Then the PCD decompositions appear to provide the best approximations. A nearly correct factorization succeeds even for the critical overlap parameters $\gamma = 0$ and $\gamma = 50$.

γ		PCD	NMF/ANLS	SPECFIT	SPEXFA
0	δ_1	$1.0 \cdot 10^{-3}$	$2.5 \cdot 10^{-1}$	$1.1 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$
	δ_2	$8.5 \cdot 10^{-4}$	1.3	$1.1 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$
10	δ_1	$1.0 \cdot 10^{-3}$	$1.7 \cdot 10^{-1}$	$1.0822 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$
	δ_2	$7.0 \cdot 10^{-4}$	1.5	$1.1 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$
20	δ_1	$1.0 \cdot 10^{-3}$	$4.7 \cdot 10^{-2}$	$1.0 \cdot 10^{-3}$	$9.1 \cdot 10^{-4}$
	δ_2	$5.9 \cdot 10^{-5}$	1.8	$9.1 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$
30	δ_1	$1.0 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$7.4 \cdot 10^{-4}$
	δ_2	$1.7 \cdot 10^{-5}$	1.5	$7.4 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$
40	δ_1	$1.0 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$6.0 \cdot 10^{-4}$
	δ_2	$1.7 \cdot 10^{-5}$	1.2	$6.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$
50	δ_1	$2.1 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$5.2 \cdot 10^{-4}$
	δ_2	$2.7 \cdot 10^{-5}$	1.1	$5.2 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$

Table 2: Reconstruction errors of the concentration profiles $C_{i,:}^{(\text{METHOD})}$.

	PCD	NMF/ANLS	SPECFIT	SPEXFA
ε_1	$1.6 \cdot 10^{-5}$	$5.0 \cdot 10^{-1}$	$1.3 \cdot 10^{-1}$	$1.3 \cdot 10^{-1}$
ε_2	$1.6 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-1}$	$1.3 \cdot 10^{-1}$
δ_1	$4.0 \cdot 10^{-2}$	$2.6 \cdot 10^{-1}$	$4.1 \cdot 10^{-1}$	$4.0 \cdot 10^{-1}$
δ_2	$3.6 \cdot 10^{-2}$	1.1	$4.0 \cdot 10^{-1}$	$4.1 \cdot 10^{-1}$

Table 3: Reconstruction errors for a interior subset of 31 spectra and $\gamma = 0$.

4.1. Application to the Rhodium-catalyzed hydroformylation

For the Rhodium-catalyzed hydroformylation the catalyst formation is monitored by FT-IR spectroscopy. Spectral recovery is used to get insight into the reactive system. Eleven experiments with varying (temperature/concentrations) conditions were carried out resulting in a total number of 165 spectra each with 624 spectral channels. This gives rise to a 165×624 absorbance matrix. By means of the PCD algorithm 6 independent components have been recovered from the the mixture spectra. This finding correlates with the largest singular values of the absorbance matrix; the first 10 singular values are 7.5890, 4.0335, 3.3468, 2.1408, 0.9938, 0.8716, 0.3753, 0.1833, 0.1293, 0.1000. The right singular vectors 1–5 and 17 (example of an oscillatory and chemically not meaningful) are drawn in Fig. 6. The component spectra are proper linear combinations of these singular vectors. PCD with $s = 6$ has determined the following component spectra, see Fig. 7. Very similar results have been obtained by the BTEM code, see Fig. 8. Slightly oscillatory or negative components can be controlled by proper choice of the weighting factors γ_i.

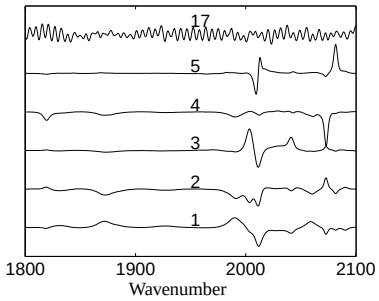

Figure 6: Right singular vectors 1–5 and 17.

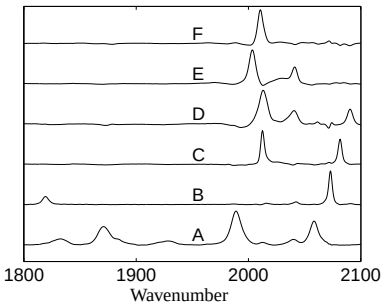

Figure 7: Spectra of PCD recovered pure components A-F.

These 6 components have been identified to be:

A Cyclohexane (the solvent).
B $Rh_6(CO)_{16}$.
C $Rh(acac)(CO)_2$.
D $HRh(CO)_3$(I) , (I) see Fig. 9.
E $HRh(CO)_2$(II), (II) see Fig. 9.
F Complex formed by each one molecule of C and D.

We note that this problem shows overlapping bands (C 2012.5cm^{-1}, D 2013cm^{-1}, E 2003.5cm^{-1}, F 2011cm^{-1}); For the components C, D and E these band are correlated with further peaks in the range 2020–

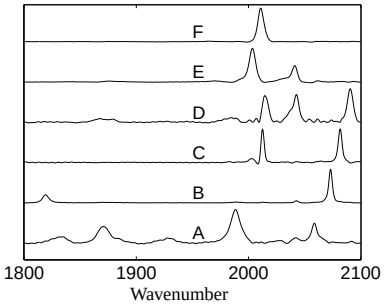

Figure 8: Spectra of BTEM recovered pure components A-F.

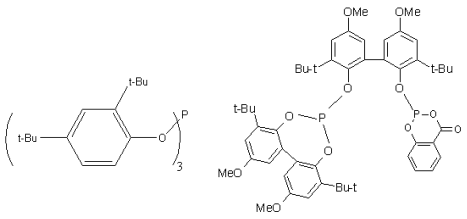

Figure 9: Left: Ligand I. Right: Ligand II.

2100cm^{-1} which supports the reliable work of spectral recovery algorithms.

Next two of the underlying experiments are described in more detail. The associated concentration profiles (disjoint participating components!) are shown in Fig. 10.

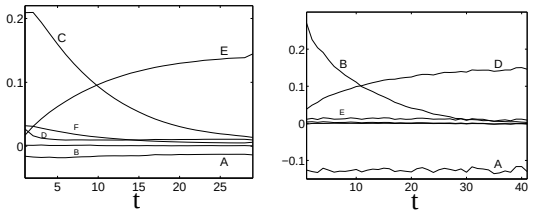

Figure 10: Left: 1st experiment. Right: 2nd experiment.

First experiment: (see Fig. 10 left) A solution of 0.0128 g C and 0.0903 g acylphosphite (II) in 48.877 g cyclohexane was prepared in a 100 ml Parr autoclave. The molar surplus of the ligand relative to Rh is 2.79. A hydridocarbonyl complex E is then formed at 303 K and 10 bar synthesis gas pressure ($CO:H_2$ = 1:1). The formation of the active, modified hydridocarbonyl complex is completed after 90 min. The spectra were taken at an interval of 3 minutes. There is no formation of $Rh_6(CO)_{16}$. The concentration of the solvent cyclohexane is almost constant. Component D is not present in this experiment.

Second experiment: (see Fig. 10 right) A solution consisting of 0.0091 g $Rh_4(CO)_{12}$ in 42.7505 g cyclohexane was prepared in a 100 ml Parr autoclave. $Rh_4(CO)_{12}$ is quantitatively converted to $Rh_6(CO)_{16}$ (B) by heating the solution at 50 bar synthesis gas to 393 K.

A Rh-hydridocarbonyl complex D with one molecule of monophosphite (I) is then prepared from the component B at 393 K and 50 bar synthesis gas pressure ($CO:H_2$ = 1:1) by addition of 0.6286 g of component (I) in 6.2435 g cyclohexane. The molar surplus of the ligand relative to Rh is 21.1.

Spectra were recorded at a time interval of 30 seconds. The formation of D is completed after 15 minutes. E is not present in this experiment.

For the BTEM spectra the associated concentration profiles can be computed by means of the least-squares procedure; we observed in some part negative concentration profiles. The SPEXFA tool has difficulties to treat the negative absorption data (which were given after background subtraction); the found spectra are not very well separated. The NMF/ANLS tool was not able to find six independent components; negative absorption values cannot be treated.

5. Conclusion

The chemometric problem of pure component spectral recovery has been presented as a constrained matrix factorization problem. Without any a priori information reliable estimates for the pure component spectra and the associated concentration profiles can be computed by means of the PCD software. The results were compared with the powerful software packages BTEM, NMF/ANLS, SPECFIT and SPEXFA. Especially for overlapping spectra and in the case of truncated sequences of spectra the BTEM and the PCD algorithms have proved as reliable algorithms and as valuable tools for revealing reactants and unknown intermediates in chemical reactions.

References

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

[2] R. Kramer. *Chemometric techniques for quantitative analysis*. CRC Press, Boca Raton, 1998.

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

[4] E.A. Sylvestre, W.H. Lawton, and M.S. Maggio. Curve resolution using a postulated chemical reaction. *Technometrics*, 16:353–368, 1974.

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

[6] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

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

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

[9] 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.

[10] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, 52:155–173, 2007.

[11] Y. Zeng and M. Garland. An improved algorithm for estimating pure component spectra in exploratory chemometric studies based on entropy minimization. *Analytica Chimica Acta*, 359:303–310, 1998.

[12] M. Garland et. al. Band target entropy minimization. Graphical user interface software. Institute of chemical engineering sciences (ICES), Singapore, July 2006.

[13] 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.

[14] E. Widjaja, C. Li, and M. Garland. Algebraic system identification for a homogeneous catalyzed reaction: Application to the Rhodium-catalyzed hydroformylation of alkenes using in situ FTIR spectroscopy. *J. Catal.*, 223:278–289, 2004.

[15] L. Chen, W. Chew, and M. Garland. Spectral pattern recognition of in situ FT-IR spectroscopic reaction data using minimization of entropy and spectral similarity (MESS); Application to the homogeneous Rhodium catalyzed hydroformylation of isoprene. *Appl. Spectrosc.*, 57:491–498, 2003.

[16] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix. Anal. Appl.*, 30:713–730, 2008.

[17] R.A. Binstead, B. Jung, and A.D. Zuberbühler. SPECFIT/32 Global analysis system. Spectrum Software Associates, Chapel Hill, NC, USA, 2000.

[18] H. Gampp, M. Maeder, C.J. Meyer, and D. Zuberbühler. Calculation equilibrium constants from multi-wavelength spectroscopy data I, Mathematical considerations. *Talanta*, 32:95–101, 1985.

[19] E. Malinowski. Factor analysis toolbox for Matlab. Applied Chemometrics, Inc., PO Box 100, Sharon, MA 02067, USA.

[20] M. Maeder and A.D. Zuberbühler. Nonlinear least-squares fitting of multivariate absorption data. *Anal. Chem.*, 62:2220–2224, 1990.