Why is the Peak Group Analysis so effective for IR spectra analysis?

Klaus Neymeyr[a,b], Christoph Kubis[a], Lukas Prestin[a], Robert Franke[c,d], Mathias Sawall[a]

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

Abstract

Peak Group Analysis (PGA), an MCR algorithm, has proven to be very successful in extracting pure component spectra and concentration profiles from IR and Raman spectral data when investigating carbonylation reactions with transition metal catalysts. In this field of study, mixture spectra typically exhibit high spectral selectivity, meaning certain peaks belong exclusively to a single chemical species. Under this condition, PGA can extract the associated pure component spectrum from the mixture spectra, and there is no factor ambiguity in these profiles. Here, we present a short mathematical proof of this remarkable PGA property.

Key words: Multivariate curve resolution, Peak group analysis, Nonnegative matrix factorization

1. Peak group analysis

Peak Group Analysis [1, 2] is a multivariate curve resolution technique used to recover pure component spectra from (time-)series of mixture spectra. Given a series of k mixture spectra, each with n channels, in the form of a $k \times n$ matrix D for a chemical system with s chemical species, the goal is to find the pure spectra $S \in \mathbb{R}^{n\times s}$ and the associated concentration profiles $C \in \mathbb{R}^{k\times s}$ such that

$$D = CS^T + E \tag{1}$$

according to the Lambert-Beer law. The matrix elements of the error matrix E are close to zero and represent the impact of noise, instrumental distortions, small deviations from strict bilinearity, and other perturbations. Many multivariate curve resolution techniques have been developed to solve the (approximate) pure component factorization problem (1), see the references [3, 4, 5, 6, 7] and many others. A major problem for any MCR technique is the factor ambiguity, or the fact that the factorization problem often has multiple solutions. These solutions can be grouped into sets of pairs of feasible factors (C, S) for which Equation (1) is satisfied. See [3, 4, 8, 9] for more information on the factor ambiguity, also known as rotational ambiguity. Factor ambiguity can be reduced when additional information about the chemical system is available, e.g., a kinetic model [10] or the partial knowledge of some pure component profiles.

1.1. Industrial process driven application of PGA

Here, we assume spectra with relatively sharp and partially isolated peaks, some of which are caused by only a single chemical species. This situation is commonly encountered in the in situ FTIR spectral analysis of rhodium-, cobalt-, and iridium-based catalysts for the homogeneously catalyzed hydroformylation of olefines. Hydroformylation is a multi-step catalytic process. In general, carbonylation processes are large-scale industrial processes, so a detailed understanding of them is important. We found that PGA is very useful for analyzing FTIR spectroscopic data gained from such chemical reactions because it can often determine true pure component spectra without any factor ambiguity. Since our initial work with PGA [1], the method has been further developed through collaboration between research and industrial chemists and numerical mathematicians. One goal of PGA is to extract pure component spectra of low-concentration, catalytically active species. The chemical background problem is the detailed analysis of the mechanistic and kinetic aspects of carbonylation reactions, particularly the hydroformylation process with transition metal complexes as catalysts [11, 12, 13, 14, 15, 16, 17]. The questions concern the formation of the catalyst, the various intermediates involved in the catalytic cycle, the dormant states of the catalyst, and the decomposition of the catalyst. For these types of reaction, IR spectroscopy using a high-pressure transmission IR flow cell is a well-established and powerful technique [18]. The resulting IR spectra are characterized by many more-or-less well separated relatively sharp peaks. See Figure 1 for a typical FTIR spectrum over a large spectral range together with the spectral window $[1960 - 2120]$ cm^{-1}.

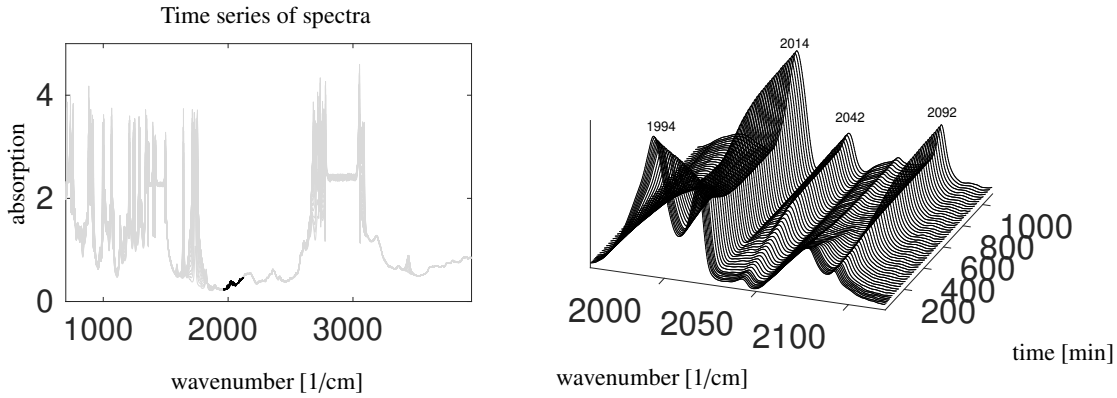

Figure 1: Time series in situ FTIR spectra (2D plot) on the rhodium catalyzed hydroformylation of 3,3-dimethyl-1-butene (producing 4,4-dimethylpentanal and 2,3,3-trimethylbutanal) based on the Rh(acac)$(CO)_2$/(P) catalyst system with P= TDTBPP. Left: full spectral range of the FTIR spectra. Right: the spectral window $[1960-2120]$ cm^{-1} after baseline correction that contains the absorbance signals of the catalyst species.

1.2. Spectral selectivity

In situ FTIR spectroscopic data of transition metal carbonyl complexes acquired during carbonylation reactions typically shows partially isolated peaks that are caused by only a single chemical species. This situation is referred to as *spectral selectivity* in chemometrics. For more information on this concept, see the work of Kvalheim and Liang [19] and Tauler, Smilde and Kowalski [20]. Selectivity is a well-known, very strong criterion for recovering the chemically correct profiles.

Selectivity is often found in the concentration space, namely zero-concentration windows can be identified for certain chemical species. This allows the spectral data matrix to be subdivided into submatrices. Then MCR is applied to these submatrices. The reduced number of chemical species makes the subsystem problems easier to solve. Numerical rank calculations for submatrices of the full spectral data matrix support the process of finding selective regions and to estimate the number of chemical species contributing to these regions. The submatrix rank is called the *local rank*. In an optimal case, if the rank of a submatrix equals 1, then the ambiguity of the concentration or spectral profile of this species is reduced to uniqueness. This is called a unique resolution.

Here, the spectroscopic data comes from in situ FTIR spectroscopic measurements on the reactions introduced in Section 1.1. This data often exhibits the strongest form of spectral selectivity, namely that for each of the catalyst species (e.g., precursor, active catalyst complexes with different ligands, intermediates, dormant or partially decomposed species) at least one single-peak spectral window exists to which (approximately) only this chemical species contributes. We call this extreme form of selectivity a *full spectral selectivity*. Given this strong assumption, PGA can recover the pure component spectrum of this species across the full frequency range of the spectral measurements. This property is proven in this work.

2. PGA algorithm

A brief description of PGA is provided below. In a first step, the current version of PGA, see [2], automatically detects the peaks in the series of mixture spectra. The input for this step includes the maximum number of peaks to be detected, a sensitivity level and a weighted selection of the criteria for the peak detection by using the discrete second derivative of the measured mixture spectra in the frequency direction, the first derivative in the time direction of the time series of spectra, discrete derivatives of the right singular vectors as well as the variances of the spectra in frequency and time direction. The second step of PGA is a constrained nonlinear optimization in which the detected peaks are locally approximated (even the strongest locality, with only one spectral channel, is possible) by linear combinations of the right singular vectors of the spectral data matrix D. Important constraints for minimizing the residual $\|D - CS^T\|$ are the (approximate) nonnegativity of the columns of C and S, a small sum of squares of the spectra to favor sharp peaks and a small sum of squares of the discrete second derivative (or second differences) of the spectra to favor smooth spectra. Thus, PGA determines an associated potential pure component spectrum for each detected peak. Since the number of peaks is typically much larger than the number of chemical species, some of the spectra appear multiple times in the list of found spectra. In a final step, PGA clusters spectra that were found multiple times using a correlation analysis.

Figure 2 shows an example of an FTIR dataset from a sequential dosage experiment for the identification of involved species of a rhodium-based hydroformylation reaction system [21]. PGA can extract nine pure component spectra.

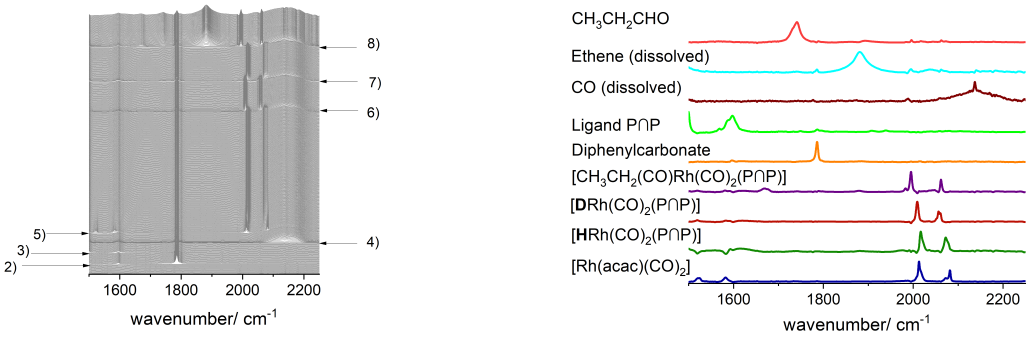

Figure 2: . In situ FTIR data set from a sequential dosage experiment under hydroformylation conditions based on the $Rh(acac)(CO)_2$/(P∩P) catalyst system with P∩P= BiPhePhos. Left: time resolved spectral series from a stepwise dosage of chemical components/complexes in the in situ transmission FTIR reactor system. 1) Solvent (Cyclohexane), 2) IR-standard (Diphenylcarbonate), 3) Ligand (BiPhePhos), 4) Gas 1 (CO/H_2), 5) Rh-precursor ($Rh(acac)(CO)_2$), 6) Gas 2 (CO/D_2), 7) Gas 1→3 (CO/H_2 →CO), 8)Gas 4 (Ethene). PGA can extract nine pure component spectra comprising organometallic complexes, organic components and dissolved gases.

3. PGA for data with full spectral selectivity

3.1. Case of ideal, non-perturbed data

The starting point is the spectral data matrix $D \in \mathbb{R}^{k\times n}$. We ignore noise or other perturbations in the following analysis. However, the numerical PGA algorithm can handle noise and other perturbations, see Section 3.2. The spectral data matrix is assumed component-wise nonnegative, i.e. $D \geq 0$, and that its rank number equals the number of chemical species s. The singular value decomposition of D is

$$D = U\Sigma V^T$$

with rectangular orthonormal matrices $U \in \mathbb{R}^{k\times s}$ and $V \in \mathbb{R}^{n\times s}$. All singular values $\sigma_1, \ldots, \sigma_s$ are strictly positive resulting in the regular diagonal matrix $\Sigma \in \mathbb{R}^{s\times s}$ of singular values. The desired pure component factorization reads

$$D = CS^T$$

with $C \in \mathbb{R}^{k\times s}$ and $S \in \mathbb{R}^{n\times s}$. The true concentration profiles and pure component spectra are the columns of these matrices.

The key steps of a PGA application to this ideal, non-perturbed data matrix D are as follows:

1. Find all peaks in the series of spectra that is stored in the rows of D.
   For the found peaks $i = 1, \ldots, \#\text{MaxNumberOfPeaks}$ let the set
   $$\mathcal{I} = \{\ell_i : \ \ell_i \text{ is the channel index of the maximum for the } i\text{th peak}\} \subset \{1, 2, \ldots, n\}$$
   collect the frequency channel indexes at the peak maxima.
2. For each $\ell \in \mathcal{I}$ find a nonnegative linear combination $Vx \geq 0$ so that Vx reproduces the peak in a way that $(Vx)_\ell = 1$ (peak height normalized to 1) and where x is of smallest Euclidean length. Orthogonality of V implies that $\|x\|_2 = \|Vx\|_2$ so that also the spectrum Vx has a smallest sum of squares; compare this with the similar optimization in [22]. Hence, the vector x of smallest Euclidean norm is the solution to the minimization problem
   $$x = \arg\min_z \{\|z\|_2 : \ Vz \geq 0, \ \ (Vz)_\ell = 1\}. \tag{2}$$

The next theorem proves that full spectral selectivity is sufficient for Vx, x by Eq. (2), to be a true pure component spectrum.

Theorem 3.1. *For the given spectral data matrix D and for the set $\mathcal{I}$ of peak positions let* full spectral selectivity *be fulfilled. This means that for each chemical species $p = 1, \ldots, s$, there is a frequency channel ℓ_p with*

$$\begin{aligned} S(\ell_p, p) &> 0 \qquad \text{nonzero absorption by the species } p \text{ at channel } \ell_p, \\ S(\ell_p, j) &= 0 \quad \forall j \neq p \qquad \text{no interfering absorption by other species.} \end{aligned} \tag{3}$$

(In words, the pure component spectrum $S(:, p)$ of the pth chemical species has at least one frequency channel ℓ_p where only this species has nonzero absorption among all species.)

Then, from given D, all pure component spectra $S(:, p)$, $p = 1, \ldots, s$, can be recovered (up to multiplicative constants). The concentration factor is also uniquely determined (columnwise up to multiplicative constants).

Proof. If the peak position ℓ_p satisfies (3) for the chemical species p and if Vx according to (2) fulfills $(Vx)_{\ell_p} = 1$, then it holds that

$$\begin{aligned}
Vx =& V \arg\min_z \{\|Vz\|_2 :\ Vz \geq 0,\ \ (Vz)_{\ell_p} = 1\} && \text{by (2) with } \|z\|_2 = \|Vz\|_2 \\
=& D^T U\Sigma^{-1} \arg\min_z \{\|D^T U\Sigma^{-1} z\|_2 :\ D^T U\Sigma^{-1} z \geq 0,\ \ (D^T U\Sigma^{-1} z)_{\ell_p} = 1\} && \text{by } V = D^T U\Sigma^{-1} \\
=& D^T \arg\min_y \{\|D^T y\|_2 :\ D^T y \geq 0,\ \ (D^T y)_{\ell_p} = 1\} && \text{by setting } y = U\Sigma^{-1} z \\
=& SC^T \arg\min_y \{\|SC^T y\|_2 :\ SC^T y \geq 0,\ \ (SC^T y)_{\ell_p} = 1\} && \text{by } D^T = SC^T \\
=& S \arg\min_w \{\|Sw\|_2 :\ Sw \geq 0,\ \ (Sw)_{\ell_p} = 1\} && \text{by setting } w = C^T y \\
=& S \arg\min_w \{\|\sum_{\substack{i=1 \\ i\neq p}}^{s} w_i S(:,i) + w_p S(:,p)\|_2 :\ Sw \geq 0,\ \ (Sw)_{\ell_p} = 1\}.
\end{aligned}$$

Assuming $w_m < 0$ for an $m \in \{1, \ldots, s\}$, then $S(\ell_m, m) > 0$ by (3) implies that $w_m S(\ell_m, m) < 0$. Hence with $S(\ell_m, j) = 0$ for all $j \neq m$ it holds that $(Sw)_{\ell_m} < 0$. Such a vector w is not feasible for the minimization because it breaks the constraint $Sw \geq 0$. Hence $w_m \geq 0$ for all m. Then all summands in the norm $\|\sum_{i\neq p} w_i S(:,i) + w_p S(:,p)\|_2$ are nonnegative and the minimum with respect to w can only be attained if $w_i = 0$ for all $i \neq p$, but for w_p the normalization constraint holds. Hence

$$\begin{aligned}
Vx =& S \arg\min_w \{\|w_p S(:,p)\|_2 :\ Sw \geq 0,\ \ (Sw)_{\ell_p} = 1\} \\
=& S(:,p)/S(\ell,p)),
\end{aligned}$$

where in the last step the minimizer is $w = e_p/S(\ell_p, p)$. Therein, e_p is the standard basis vector (pth column of the identity matrix).

The resulting equation $Vx = S(:,p)/S(\ell_p, p))$ says the desired spectrum $S(:,p)$ of the pth chemical species equals Vx times the (nonzero) scaling constant $S(\ell_p, p)$. The concentration factor is given by $C = D(S^T)^+$ with the pseudoinverse of the found factor S^T. □

The proof shows that under the assumption of full spectral selectivity, the true factors C and S^T can be recovered (up to the trivial and principally unavoidable scaling ambiguity).

3.2. PGA analysis of experimental data

Experimental spectral data can only approximately fulfill the full spectral selectivity assumption made by Thm. 3.1. One reason is noise, which increases the number of nonzero singular values. However, at a low noise level, the dominant nonzero singular values belonging to the essential chemical information are well separated from the other singular values close to zero. In the numerical PGA algorithm, the truncated SVD ignores the near-zero singular values. Furthermore, a baseline can violate the assumption (3) even after a baseline correction step. These and other data perturbations must be successfully addressed. The numerical PGA algorithm [2] uses penalized nonlinear optimization to construct local peak approximations with respect to the basis of the right singular vectors belonging to the dominant singular values. This stabilizes the computation and steers the numerical calculations in the desired direction, yielding results similar to those attainable with ideal, non-perturbed data. See references [1, 2] for details on the constrained minimization problem and its numerical solution.

4. Conclusion

Full spectral selectivity is the key to the effectiveness of PGA in MCR analyses of in situ IR and Raman spectroscopic data collected during homogeneously catalyzed carbonylation reactions. This demonstrates the usefulness of the selectivity criterion in MCR analyses once again. Our future goal is to develop PGA as a tool for the fast online/real-time data analysis of the spectral data stream produced in chemical laboratory and industrial processes.

Acknowledgement

Funding by the DFG Research Training Group 2943 "SPECTRE" (project number 507189291) is gratefully acknowledged.

References

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

[2] M. Sawall, C. Kubis, B. N. Leidecker, L. Prestin, T. Andersons, M. Beese, J. Hellwig, R. Franke, A. Börner, and K. Neymeyr. An automated Peak Group Analysis for vibrational spectra analysis. *Chemom. Intell. Lab. Syst.*, 254:105234, 2024.

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

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

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

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

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

[8] H. Abdollahi and R. Tauler. Uniqueness and rotation ambiguities in multivariate curve resolution methods. *Chemom. Intell. Lab. Syst.*, 108(2):100–111, 2011.

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

[10] M. Sawall, A. Börner, C. Kubis, D. Selent, R. Ludwig, and K. Neymeyr. Model-free multivariate curve resolution combined with model-based kinetics: Algorithm and applications. *J. Chemom.*, 26:538–548, 2012.

[11] C. Kubis, R. Ludwig, M. Sawall, K. Neymeyr, A. Börner, K.-D. Wiese, D. Hess, R. Franke, and D. Selent. A comparative in situ HP-FTIR spectroscopic study of bi- and monodentate phosphite-modified hydroformylation. *ChemCatChem*, 2:287–295, 2010.

[12] C. Kubis, M. Sawall, A. Block, K. Neymeyr, R. Ludwig, A. Börner, and D. Selent. An operando FTIR spectroscopic and kinetic study of carbon monoxide pressure influence on rhodium-catalyzed olefin hydroformylation. *Chem.-Eur. J.*, 20(37):11921–11931, 2014.

[13] R. Franke, D. Selent, and A. Börner. Applied hydroformylation. *Chem. Rev.*, 112:5675–5732, 2012.

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

[15] 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.

[16] B.N. Leidecker, D. Peña-Fuentes, C. Wei, M. Sawall, K. Neymeyr, R. Franke, A. Börner, and C. Kubis. In situ FTIR spectroscopic investigations on rhodium carbonyl complexes in the absence of phosphorus ligands under hydroformylation conditions. *New J. Chem.*, 48(43):18365–18375, 2024.

[17] C. Wei, B.N. Leidecker, D. Peña-Fuentes, H. Schröder, M. Sawall, K. Neymeyr, E.V. Kondratenko, A. Börner, R. Franke, and C. Kubis. Impact of the P-ligand concentration on the formation of hydroformylation catalysts: An in situ FTIR spectroscopic study. *Chem. Ing. Tech.*, 96(12):1657–1667, 2024.

[18] M. Garland. Combining operando spectroscopy with experimental design, signal processing and advanced chemometrics. State of the art and a glimpse of the future. *Catal. Today*, 155:266–270, 2010.

[19] O. M. Kvalheim and Y.-Z. Liang. Heuristic evolving latent projections: resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution. *Anal. Chem.*, 64(8):936–946, 1992.

[20] R. Tauler, A. Smilde, and B. Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.*, 9(1):31–58, 1995.

[21] B.N. Leidecker, D. Peña Fuentes, M. König, J. Liu, W. Baumann, M. Sawall, K. Neymeyr, H. Jiao, R. Franke, A. Börner, and C. Kubis. In situ spectroscopic investigations on BiPhePhos modified rhodium complexes in alkene hydroformylation. *Catal. Sci. Technol.*, 14(14):3966–3983, 2024.

[22] 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.