algorithms

MDPI

Article

SepFree NMF: A Toolbox for Analyzing the Kinetics of Sequential Spectroscopic Data

Renata Sechi [1,2,3,*], Konstantin Fackeldey [2,4], Surahit Chewle [2,5] and Marcus Weber [2]

1 Furukawa Electric Institute of Technology, 1158 Budapest, Hungary
2 Zuse Institute Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany
3 Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, 1111 Budapest, Hungary
4 Institute for Mathematics, TU Berlin, Strasse des 17. Juni 135, 10623 Berlin, Germany
5 Federal Institute for Materials Research and Testing (BAM), Richard-Willstätter-Straße 11, 12489 Berlin, Germany
* Correspondence: renata.sechi@furukawaelectric.com

Abstract: This work addresses the problem of determining the number of components from sequential spectroscopic data analyzed by non-negative matrix factorization without separability assumption (SepFree NMF). These data are stored in a matrix M of dimension "measured times" versus "measured wavenumbers" and can be decomposed to obtain the spectral fingerprints of the states and their evolution over time. SepFree NMF assumes a memoryless (Markovian) process to underline the dynamics and decomposes M so that $M = WH$, with W representing the components' fingerprints and H their kinetics. However, the rank of this decomposition (i.e., the number of physical states in the process) has to be guessed from pre-existing knowledge on the observed process. We propose a measure for determining the number of components with the computation of the minimal memory effect resulting from the decomposition; by quantifying how much the obtained factorization is deviating from the Markovian property, we are able to score factorizations of a different number of components. In this way, we estimate the number of different entities which contribute to the observed system, and we can extract kinetic information without knowing the characteristic spectra of the single components. This manuscript provides the mathematical background as well as an analysis of computer generated and experimental sequentially measured Raman spectra.

Keywords: kinetics from experiments; separability assumption; sequential spectroscopic data

PACS: 02.50.Ga; 82.20.-w

Citation: Sechi, R.; Fackeldey, K.; Chewle, S.; Weber, M. SepFree NMF: A Toolbox for Analyzing the Kinetics of Sequential Spectroscopic Data. *Algorithms* **2022**, *15*, 297. https://doi.org/10.3390/a15090297

Academic Editors: Aneesha Bakharia and Khanh Luong

Received: 8 July 2022
Accepted: 15 August 2022
Published: 24 August 2022

1. Introduction

The laboratory conditions of sequential data from spectroscopy experiments are often chosen in such a way that the processes can be assumed to be autonomous. In autonomous processes, only the initial states of the entire system determine its future. The Markov property, the lack of a memory, relates to the microstates of the system in the processes examined, i.e., what happens in detail (metaphorically, behind the curtains) is Markovian. These details are often not searched for.

For instance, one could think of a toy-model reaction; in this reaction, a red solution gradually turns orange and then yellow with the time. By taking a picture of the solution with a certain time-step τ, one observes the solution changing color. However, the exact nuance of the solution, e.g., the solution is orange-red, is not relevant for the experiment. Rather, one is interested only if the solution is red, orange or yellow. From the perspective of the solution, the overall system breaks down into three states; the Markov process, which takes places within the continuum of the colors that the solution can assume, is projected onto a three-dimensional space. A vector h, $h \in \mathbb{R}^{3\times m}$, indicates the proportion occupied

by the colors red, orange and yellow in the overall system at each point in time of the measurement. This situation is displayed in Figure 1.

Figure 1. In the micro-system, the solution can assume a gradient of colors between red and yellow (continuum). In the macrosystem, the projection identifies only three states (red, orange, yellow).

In other words, the solution has a memory because its color cannot "switch" from red to yellow within the interval τ; if the solution is red, and it was dark red before, in the next time step it will be more likely orange-red, not yellow. Meanwhile, the particles in the solution can jump from red to yellow or to any color, regardless of the past particle color (Markov). If the Markov property were also met on the level of the projected macrosystem (the solution can be red-orange-yellow), one could find, e.g., an ordinary differential equation (a Markov process) matrix for the evolution of the concentration vectors. A solution of this system for fixed time-step τ could be represented by a matrix equation:

$$h_{t+\tau}^T = h_t^T K(\tau),$$

where the matrix $K(\tau)$ usually also depends on further quantities (e.g., temperature at which the reaction takes place), and $h^T(t)$ is the proportion of a state in the reaction. In the analysis of sequentially measured spectroscopic data, h is associated to a spectral fingerprint, as explained later in this work.

The transition from a micro to a macrosystem can have a systematic source of error, because a projected Markov system loses its Markov property.

Mathematically, there are two possibilities to model this "non-Markovianity":

A The process is assumed to have a memory. Not only the current (macro)state of the system, but also past steps of the process development have to be taken into account in order to derive a "forecast" for the future.

B The clustering of microstates into coarser macrostates is assumed to be "fuzzy". Then Markoviantity can be preserved, but we lose control about what we will denote as a macrostate.

Chemically, this article develops option *B* and presents a tool to compute the minimal past-time dependency of compounds in a reaction spectrum. A compound is a system's state with an associated spectral fingerprint. The tool considers not only the compounds' dependency on the past, which is called *memory*, but also the dependency in the kinetic development between the compounds themselves. This memory effect has been investigated in the mathematical context (e.g., [1,2]) and it has been observed in chemistry, for instance in the context of protein folding [3,4].

The interpretation of the compounds' kinetics is more meaningful if the development of each kinetic curve is as independent as possible to the development of the other kinetic curves. A correlated development of kinetics causes an overlap of the kinetic curves. This is because they represent distinct physical system states and they are not chemical components.

The tool presented in the following measures the memory of the reaction compounds by scoring the overlap of the kinetic curves, following the theory of [5]. This analysis determines physical states corresponding to the basis vectors of an invariant subspace, as we will explain in the theory section. This way to compute the minimal memory effect is advantageous for spectral-data analysis, since it helps to determine the number of compounds in the reaction. If the number of compounds is too high, two or more compounds' kinetic curves will overlap strongly since they are assigned to the same basis vector, or a linear combination of them.

Knowing the number of reaction compounds is important not only when analyzing the datasets, but also when performing experiments. In fact, the memory measure gives insights into the quality of the sample and the noise levels of the experimental setup.

The estimation of the minimal memory effect is possible with the SepFree NMF (separability-assumption free non-negative matrix factorization), a novel method based on the *non-negative matrix factorization without separability assumption* framework [6]. The non-negative matrix factorization without separability assumption is a matrix-decomposition method already used in the analysis of sequentially measured spectroscopic data. The structure of the datasets is the spectrum, in a certain wavenumber interval, measured for different times after the excitation of the sample. Furthermore, it is applicable also to time-resolved spectra; these investigate the evolution of photo-induced phenomena, as they measure the change of the optical signal (emission, absorption, Raman scattering) as a function of time (e.g., [7–9]).

This work uses the SepFree NMF for the analysis of the sequentially measured spectra. The method examines the sequentially measured spectrum as a matrix M to decompose into a product of matrix H, describing concentration proportions of the compounds as function of time, and matrix W, which describes the peak intensities of the compounds as function of wavenumber. In particular, the SepFree NMF models the evolution of the concentrations' proportions of the compounds in H as a memoryless process, called the Markovian process [6].

A transition probability matrix matrix $K(\tau)$ governs the τ-evolution of the columns of H:

$$H_{t+\tau}^T = H_t^T K(\tau). \tag{1}$$

The main contribution of this work is to provide a measure for the overlap between the rows of the concentration–proportion matrix H.

The SepFree NMF identifies the *physical* states of the system, not necessarily the chemical ones. In order to understand the difference, think of a reaction in which all the molecules can react into two compounds, A and B, and then go back to the initial state. In a standard multivariate curve resolution analysis [10], the concentration profile of the molecules in A and B evolves as in Figure 2 left. These curves represent the concentrations of A and B in the system as a function of time; they are not between 1 and 0. By SepFree NMF of rank 2, the algorithm returns the amount of system that is either is state I or in state II. Their time development will look as in Figure 2 middle. The first state I is the initial *physical* state of the system at the beginning of the reaction; all the molecules are in this state at time $t = 0$, so its proportion will be (almost) 1. During the reaction, a second state II arises. This state II represents the amount of the system that is not in the initial condition I, that is, the amount of system molecules that is in either A or B. So, being in A or B is the second *physical* state of the system. By asking SepFree NMF for a decomposition into three physical compounds, the algorithm yields three states: the initial physical states, A', and B' (see Figure 2 right). Although the curves in Figure 2 right have a similar development to the concentration profiles in the left picture, their interpretation for the system is different. They represent the amount of the system that is in physical states A' and B', or in the initial state I, they are not concentration profiles.

This article is divided into three main parts. The first one explains the background of the SepFree NMF (Section 2.1), and of the PCCA+ projection (Section 2.2). The second part shows how we propose to compute the memory from the NMF results (Section 3). In the third part, we present examples of sequentially measured Raman spectra: first for computer-generated data, then for an experiment measuring the Raman spectrum of the crystallization process of paracetamol in methanol (Section 4).

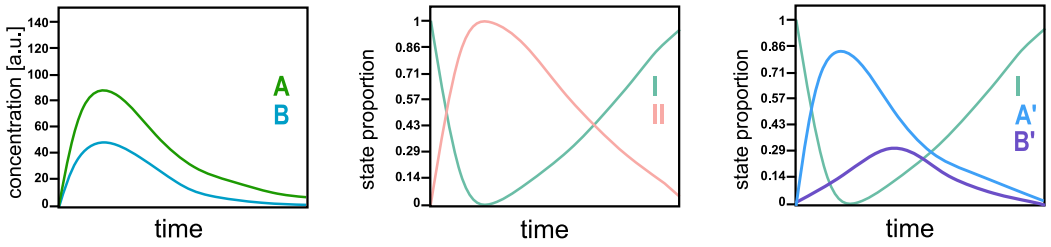

Figure 2. Qualitative comparison of kinetic profiles with MCR and SepFree NMF. MCR (**left**), SepFree NMF of rank 2 (**middle**), SepFree NMF of rank 3 (**right**).

2. Theoretical Foundations

Memory effects emerging from the projection of Markov micro processes to macro scales were discussed first with regard to mono- and multivalent binding events [11–13]. The resulting models provide a link for understanding macroscopic binding kinetics and to evaluate the strength and duration of the interaction of the ensemble molecules, facilitating the rational engineering of the binding dynamics.

It has been discussed whether and how these effects contribute to the binding rates of ligand–receptor systems [5]. The projection method to be applied is called non-negative matrix factorization (NMF).

2.1. SepFree NMF

SepFree NMF (separability-assumption free non-negative matrix factorization) is a novel method to decompose sequentially measured spectra. Decomposing the dataset into a matrix carrying the kinetic information and a matrix containing the peak intensities information is the first step for the computation of the memory effect of the compounds; this section briefly introduces the characteristics of the method.

The spectral dataset $M \in \mathbb{R}^{n\times m}$ consists of a spectrum measured at m time points and for n wavenumbers. Then, the matrix M contains the information about the dynamics of a number r of compounds. The SepFree NMF factorizes M into the multiplication of a matrix W, describing the peak intensities of the compounds, and a matrix H, describing their relative concentration proportions in the system as function of time [6]

$$M = WH \tag{2}$$

with the matrix $W \in \mathbb{R}^{n\times r}$ and the matrix $H \in \mathbb{R}_{+}^{r\times m}$. The matrix W represents r-component fingerprints as a function of the wavenumber.

Each of the r rows of the matrix H represents the proportion of the r-compounds as function of time t. Since H represents proportions of system's states, its entries are required to be positive and between $[0, 1]$. The column sum of H is a partition of unity.

Each of the r columns of the matrix W represents the intensity of a compound as a function of the wavenumber (or wavenumbers). With the standard ansatz, the matrix M has to be r-separable [14,15]. A matrix M is r-separable if there exists a factorization for which all r-columns of W are equal to a column of M [6]. This implies that a compound is present at least once as 100% at a time t in the process described by M, while all the other compounds have a concentration of 0% at t. However, in experiments, one measures always a mixture of compounds, and the forenamed situation in which a compound constitutes the entire system is not realistic. If only one compound is present at a specific time, then all the sample molecules would be in the same condition at that point in time.

Thus, the auxiliary of the separability assumption makes this algorithm particularly appropriate for the analysis of experimental data.

The authors of [6] model the time evolution of the columns in H as a Markov process. The evolution of the compounds at time $t + \tau$ depends only on the conditions at time t, and it is given by (1), whereby $K \in \mathbb{R}^{r\times r}$ is a row-stochastic transition matrix. Its elements $K(\tau)_{ij}, i, j = 1, \ldots, r$, describe the transition probability from i to j after a time-step τ. In this

model, $K(\tau)$ is fixed for the time interval τ. The transition probability in the process does not explicitly depend on time, and the process is said to be autonomous [16].

The factorization of the matrix M as a product of W and H has not a unique solution [17]. One can find a set of solutions, but not all the solutions of this set will satisfy the necessary conditions the decomposition has to have to represent the system. To select the decomposition that better reconstructs the data matrix M from the set of solutions, a penalty function Ψ is defined [6]. This function weights the required structural properties of the decomposition; the optimal solution for the decomposition of M is obtained by minimizing the value of Ψ. The requirements the obtained matrices have to meet are five:

1. The entries of W are non-negative;
2. The entries of H are non-negative;
3. H is column-stochastic;
4. The entries of K are non-negative;
5. K is row-stochastic.

To simplify the reading, we drop the τ of $K(\tau)$. These five requirements are then incorporated in the objective function Ψ, defined as follows:

$$\Psi = \alpha\left(\min_{i,j} W_{ij}\right) + \beta\left(\min_{i,j} H_{ij}\right) + \gamma\left(\max_{j} \mid 1 - \sum_{i}^{r} H_{ij} \mid\right) + \delta\left(\min_{i,j} K_{ij}\right) + \mu\left(\max_{i} \mid 1 - \sum_{j}^{r} K_{ij} \mid\right), \quad (3)$$

where the coefficients $\alpha, \beta, \gamma, \delta, \mu$ before each addend allow to design an objective function that fits the data characteristics.

Regarding the requirements of the SepFree NMF method, the non-negativity condition of the matrix W holds or not, depending on the kind of analyzed spectrum. For example, in transition absorption spectra, the intensities in W can be also negative [18].

In this work, we compute the minimal memory effect from the kinetics matrix H, obtained with the SepFree NMF algorithm. The method of the SepFree NMF has been proposed by the authors of [6]. To the notation, for any matrix Y, Y_+ is the matrix Y without the first row and Y_- is the matrix Y without the last row. Let a data matrix M, $M \in \mathbb{R}^{n\times m}$ with $\text{rank}(M) = r$ be given. Then the following steps are done in SepFree NMF:

- Singular value decomposition (SVD) of M transposed: $M^T = U\Sigma V^T$.
- Define $\tilde{U}$, $\tilde{U} \in \mathbb{R}^{m\times r}$: the first column is the constant vector $(1, 1, \ldots 1)^T$, the other columns are the first $(r-1)$ columns of U.
- Use PCCA+ to find $\tilde{H} = (\tilde{U}A)^T$.
- Use the Penrose pseudoinverse to compute $\tilde{W} = M\tilde{H}^{-1}$ and $K = (\tilde{H}_-^{-1})^T \tilde{H}_+^T$.
- Minimize Ψ for the requirements in order to find the optimal A_{opt}.
- Reconstruct the concentrations and compounds intensities matrices with via A_{opt}: $H_{rec} = (\tilde{U}A_{opt})^T$, $W_{rec} = MH_{rec}^{-1}$, $K_{rec} = (H_{rec,-}^{-1})^T\, H_{rec,+}^T$.

The matrix K is computed as the autocorrelation matrix between the τ-time-shifted concentrations H_-, H_+. With this relation, the matrix K has the meaning of a Markovian transition matrix since it gives information on the τ-step development of the concentration proportions.

For simplicity, we will refer in the examples to the optimized quantities H_{rec}, W_{rec}, and K_{rec} as H, W, and K, respectively. This decomposition method allows to analyze experimental data with different structures because the parameters of the objective function Ψ can be adjusted to weigh more or less than a feature rather than another. Therefore, setting the parameters in one or another way influences the final results of the decomposition, or rather the algorithm finds the best decomposition for those parameters, which can be very different to another one found for other parameter sets in Ψ.

2.2. Interpreting H as a Clustering

PCCA+ [19] is a clustering algorithm that projects the microstates with similar behavior to fewer dominant or macrostates (metastable states). Macrostates are collections of microstates grouped together by a similar feature.

From a mathematical point of view, macrostates/metastable states correspond to conformations $\tilde{U}$ that keep their structure upon application of the transition matrix $\tilde{K}(\tau)$:

$$\tilde{U} \approx \tilde{K}(\tau)\tilde{U}. \tag{4}$$

This means that j metastable states are eigenvectors to the eigenvalues $\ell_j \approx 1$, which are the dominant eigenvalues of $\tilde{K}(\tau)$. Note that transition matrices have eigenvalues $\ell_i \in [0,1], i = 1,\ldots,n$ and that the first eigenvalue (Perron eigenvalue) $\ell_1 = 1$. Now PCCA+ uses $\tilde{U}$ is to assign to which degree a microstate belongs to/is a member of each one of the macrostates. The vectors in $\tilde{U}$ in Equation (4) are not membership functions yet, so the problem now is to compute the membership functions from these dominant eigenvectors. Consider a transition matrix $\tilde{K}(\tau) \in \mathbb{R}^{n\times n}$ with eigenvalues $\ell_i, i = 1,\ldots,n$ and eigenvectors $\mathcal{U} \in \mathbb{R}^{n\times n}$. Solving the eigenvalue problem, i.e., $\tilde{K}(\tau)\mathcal{U} = \tilde{\Lambda}\mathcal{U}$, $\tilde{\Lambda} = diag(\ell_1,\ldots\ell_n)$, r dominant eigenvalues are identified. Then the matrix of the dominant eigenvectors, $\tilde{U} \in \mathbb{R}^{n\times r}$, is the input of the PCCA+. To find the membership functions χ, the algorithm has to project the matrix $\tilde{U}$ such that the entries of χ are not negative and form a partition of unity. That means finding a matrix $\mathcal{A} \in \mathbb{R}^{r\times r}$ such that with

$$\chi = \tilde{U}\mathcal{A} \tag{5}$$

the membership functions matrix χ satisfies

$$\chi_j(i) \in [0,1], i \in 1,\ldots,n; \quad \sum_{j=1}^{r} \chi_j(i) = 1. \tag{6}$$

Equation (6) tells that $\chi_j(i)$ gives information about how much the i-th microconformation belongs to the j-th macroconformation [19].

With the membership functions in χ, the transition matrix of the macroconformations, $K(\tau)$, can be computed with

$$K(\tau) = \langle \chi, \chi \rangle_\pi^{-1} \langle \chi, \tilde{K}(\tau)\chi \rangle_\pi, \tag{7}$$

with π being the density distribution (e.g., uniform, stationary distribution). We remark that $\tilde{K}(\tau) \in \mathbb{R}^{n\times n}$ and $K(\tau) \in \mathbb{R}^{r\times r}$ with $r \ll n$.

3. Minimal Memory Effect in Sequentially Measured Spectroscopic Data

This section introduces a method for the computation of the minimal memory effect carried by the elements of the decomposition analysis of a two-dimensional spectrum. In sequentially measured spectroscopic datasets, signals arise and decay gradually over time. Abrupt dropping and rising in the intensity over time are not common, and/or originated by noise in the experiment. The intensity of the signal at time t is conditioned by the intensity of the signal at time $t - \tau$; in this way, the development of spectra includes short-time memory effects. In other words, the spectral signal "remembers" the near-past condition. With the method in [6], it is possible to estimate a transition matrix $K(\tau)$, which describes the transition probability between the r compounds after a time-step τ. The memory effect can be noticed prima facia from the transition matrix $K(\tau)$, in which it emerges as negative entries [20].

3.1. Relation between EDMD and SepFree NMF

For a better understanding of the meaning that the minimal memory effect has for the decomposition and the studied system, a few considerations on the involved mathematical objects should be made.

An observable is "something that can be measured", so it is a function on the state space. Spectra are optical signals, and so they are observables themselves; sequentially measured spectra are observables depending on the wavenumber and delay time. Considering only the relative-concentration matrix H, each relative concentration H^i, $H^i \in \mathbb{R}^m$, is a row-vector which contains m delay-time evaluations of the ith observable, so of one of the r compounds. We can think about the row-vector H^i as a specific function for the ith observable that depends on time. The whole matrix H, $H \in \mathbb{R}^{r \times m}$, contains a collection of time developments of all the considered r-observables [21].

By analyzing the data via the NMF without separability assumption, the transition matrix is given by

$$(H_t^T)^{-1} H_{t+\tau}^T = K(\tau). \tag{8}$$

Equation (8) means that the transition probability is given by the multiplication of the following:

1. $H(t)$, a set of time-dependent functions of the compounds at t;
2. $H(t+\tau)$, a set of time-dependent functions of the compounds at $t+\tau$, so evolved in time.

For the decomposition of sequential spectroscopic data, the collection of the concentrations of the r compounds contains the set of the time-dependent functions. The formulation in Equation (8) is the extended dynamic mode decomposition (EDMD) formulation for the projection of the Koopman operator [21,22].

In the SepFree NMF, the matrix H is not obtained as a function of the observable; rather, it results from the optimization of χ, the membership-functions matrix of PCCA+, for the components in the system. The discretized Koopman operator is obtained in Equation (8) in a EDMD manner. The PCCA+ is a specific kind of Galerkin projection (see Section 2.2) [5,19]. The corresponding transition matrix (the matrix representation of a discretized Koopman operator) is given by

$$K = \mathcal{S}^{-1} T, \tag{9}$$

with $\mathcal{S}, T \in \mathbb{R}^{r \times r}$ with r being the number of identified components [19]. Details about the derivation, the discretization and the computation of the objects in Equation (9) can be found in [19].

The expression in Equation (9) is the starting point for the derivation of the minimal memory effect [5].

EDMD computes finite-dimensional approximations of the Koopman operator; for a certain kind of decomposition of the state space, it results in Ulam's method, and thus a Markov state model (MSM) [23]. The MSM is underlying SepFree, thus the relation to EDMD and SepFree NMF gives the foundation to introduce how to compute the minimal memory effect by SepFree NMF. The relation between the two representations of the transition matrix with EDMD and SepFree NMF are analyzed in [18]. The concept of the *minimal memory effect* is based on the theory of the rebinding effect of [5].

3.2. Computation of the Minimal Memory Effect

The minimal memory effect is computed starting with the Equation (9). In (9), the matrix T represents the "pure" transition probability between compounds, i.e., T_{ij} is the probability of going from compound i to compound j after a time-step τ [5].

If the transition matrix $K(\tau)$ is identical to the "pure" transition probability matrix T, then the process is totally Markovian; the compounds' kinetics do not depend on the past. In this case, the matrix $\mathcal{S}$ is the identity matrix.

As in the expression in (9), the matrix $\mathcal{S}$ relates the PCCA+ projection $K(\tau)$ to the "pure" transition matrix T; one can think about $\mathcal{S}$ as a measure of the difference between T and K.

In [5], the authors explain that the minimal memory effect (that they call rebinding) is proportional to the trace of the rate matrix Q related to the transition matrix as

$$K(\tau) = \exp \tau Q. \tag{10}$$

The entries of Q describe the transition rates between the reaction compounds.

It has been shown in [5] that the determinant of the rate matrix in Equation (10) is given as

$$F = \tau^{-1}[(\ln \det(\mathcal{S})) - \ln \det(T))]. \tag{11}$$

Both $\mathcal{S}$ and T are stochastic matrices, and their determinant cannot be larger than one. If the difference in the compounds' kinetic is not very sharp, the matrix T has a lower-valued determinant, which means a high negative logarithm. The matrix $\mathcal{S}$ indicates the overlap between kinetic curves; in the theory of PCCA+, these are the membership functions of the dominant conformations (χ). If the overlap is large, the determinant of $\mathcal{S}$ reduces F. That means the more the membership function overlap, the more the system is stable, or the lower $\det(\mathcal{S})$, the higher the memory effect. The determinant of $\mathcal{S}$, $\det(\mathcal{S})$, is an indicator of the memory effect. Then we can compute the memory effect as

$$\mathcal{S} = \frac{\langle H^T, H^T \rangle_\pi}{\langle H^T, e_m \rangle_\pi}, \tag{12}$$

where $e_m = (1, \ldots, 1)^T$, $e_m \in \mathbb{R}^m$, is a constant vector. It is always possible to compute the minimal memory effect, because in the PCCA+ projection, there exists a rate matrix Q [24].

Summarizing, the minimal memory effect is high if $\det(\mathcal{S})$ is close to zero; it is low if $\det(\mathcal{S})$ is close to 1. The higher $\det(\mathcal{S})$, the better the decomposition.

4. Numerical Examples

These numerical examples illustrate how the minimal memory effect can be calculated from the SepFree NMF. The examined data are sequentially measured Raman spectra. One spectrum is computer generated (Section 4.1), while the other one is an experimental spectrum (Section 4.2).

The analyzed synthetic dataset is the same one proposed in [14], by applying the algorithm in [6] for the decomposition. For the experimental dataset, the algorithm is applied to a sequentially measured Raman spectrum of the crystallization process of paracetamol in methanol.

After the decomposition, every compound is labeled with a letter (A, B, C, etc.). These letters do not have any chemical meaning, but only serve as a naming convention for the data analysis routine. The labeling does not refer to the chemical meaning, and in each decomposition we can have a different order of the physical states. We will point it out in the analysis when such a situation is displayed in the figures.

Both the Python version of the algorithm to generate the synthetic data, as well as the algorithm in [6] are available in the supplementary materials. Furthermore, in the supplementary material section, a tutorial explains how to use the hereby presented method for the analysis of a dataset. This version of the SepFree NMF algorithm was developed in [18].

4.1. Synthetic Dataset

This first-order kinetics process describes the reaction of five compounds, called A, B , C, D, and E. The rate matrix Q that generates the kinetics is

$$Q = \begin{pmatrix} -0.53 & 0.53 & 0.00 & 0.00 & 0.00 \\ 0.02 & -0.66 & 0.43 & 0.21 & 0.00 \\ 0.00 & 0.25 & -0.36 & 0.00 & 0.11 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.10 & 0.00 & -0.10 \end{pmatrix}. \tag{13}$$

The concentration proportions of the chemical moieties evolve in time following the rule, $H(t+\tau)^T = H(t)^T \exp(\tau Q)$.

The generated peak intensities consist of Lorentzian curves in the matrix W. The generated dataset is then given by

$$M = WH, \tag{14}$$

whereby no artificial noise is added, and the peaks of the matrix W are well separated. Figure 3 represents the generated spectral compounds as a function of the wavenumber ν (right) and their concentration proportions as a function of delay time t (left).

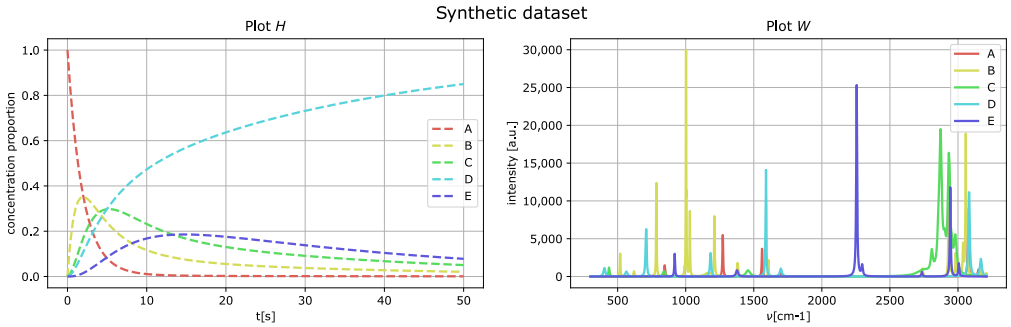

Figure 3. First-order-kinetics reaction of five spectral compounds (A, B, C, D, and E) generating the dataset in Section 4.1. The (**left figure**) shows the time development of the concentration proportions, and the (**right figure**) displays the spectral fingerprints of the compounds.

The outcome of the application of the algorithm in Section 2.1 is presented in Figure 4, and the parameters used for the optimization of the function Ψ are

$$\alpha = -0.001, \quad \beta = -10, \quad \gamma = 1, \quad \delta = 0, \quad \mu = 0.$$

We set δ and μ to zero, since they are related to the computation of the Koopman transition probability matrix K. Thus, we do not need to bias the optimization of Ψ for K.

For determining the parameters for Ψ, a heuristic method is used; see Appendix B.

By comparing the spectral fingerprint of species D and E, in Figure 4, to the spectral fingerprint of D of the generated dataset, Figure 3, we see that the species E obtained corresponds to the component labeled as D in Figure 3.

The memory effect for this reconstruction is $\det(\mathcal{S}) = 0.002$, which means that the memory effect for this decomposition is very strong. Additionally, the low value of the determinant is given by the fact that the reconstructed kinetic-matrix H contains some negative entries, even if they are close to zero.

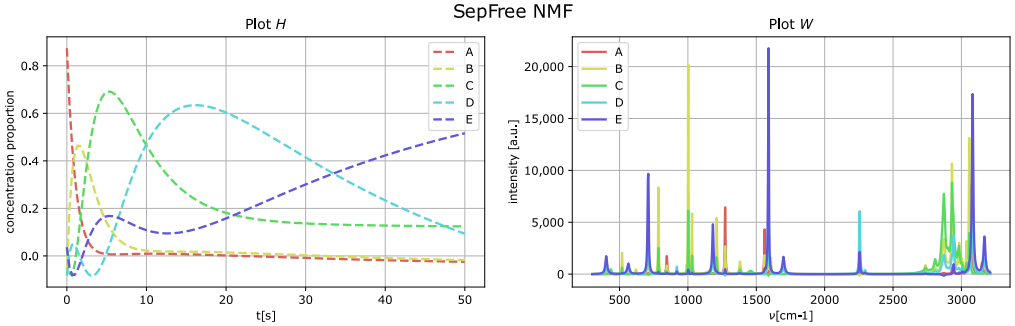

Figure 4. Reconstructed component-spectra of the compounds W on the (**right**), and reaction kinetics H on the (**left**) the for noiseless and well-separated synthetic Raman dataset. Parameter set $[-0.001, -10.0, 1.0, 0.0, 0.0]$. Minimal memory effect $\det(\mathcal{S}) = 0.002$.

4.1.1. Synthetic Dataset with Increased Interference

To increase the interference level, the peaks of the matrix W are moved next to each other. The results of the analysis with SepFree NMF are displayed in Figure 5; the decomposition is computed for the same parameter set previously used. The minimal memory effect for increasing peak-interference of 30% is $\det(\mathcal{S}) = 0.001$, so the memory effect becomes twice bigger by adding interference to the data.

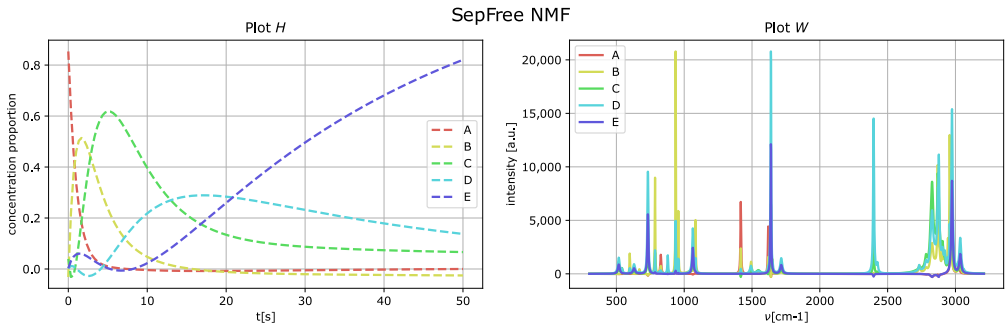

Figure 5. Reconstructed component-spectra of the compounds W on the right and reaction kinetics H on the for noiseless and 30% of increased peak interference of the synthetic Raman dataset. Paramater set $[-0.001, -10.0, 1.0, 0.0, 0.0]$. Minimal memory effect $\det(\mathcal{S}) = 0.001$.

As in the obtained results for the data without noise nor interference, the component E corresponds to the one labeled as D of the generated dataset.

4.1.2. Synthetic Dataset with Noise and Increased Interference

In order to consider the effects of the (experimental) noise to the analysis and for the computation of the minimal memory effect, the normally distributed noise matrix N, $N \in \mathbb{R}^{n \times m}$ is added to the data matrix M so that

$$\tilde{M} = M + \eta N \tag{15}$$

and the noise is scaled by a factor η, for this example $\eta = 0.4$. The results of the decomposition are displayed in Figure 6. As in the obtained results for the data without noise nor interference, the component E corresponds to the species labeled as D of the generated dataset. The minimal memory effect computed with the decomposition is 0.0005. The introduction of the noise makes a remarkable difference in the memory properties of the spectrum. In comparison to the dataset with well-separated peaks and no additional noise, the minimal memory effect is four times bigger.

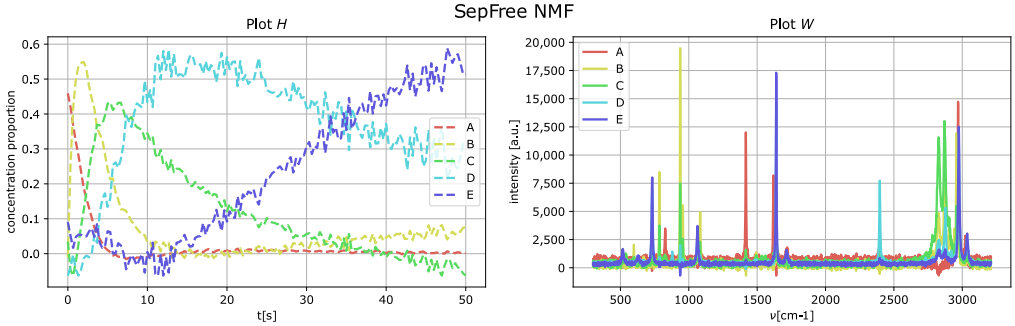

Figure 6. Reconstructed component-spectra of the compounds W on the right and reaction kinetics H on the for noise level $\eta = 0.4$ and 30% increased interference of the synthetic Raman dataset. Parameter set $[-0.001, -10.0, 1.0, 0.0, 0.0]$. Minimal memory effect $\det(\mathcal{S}) = 0.0005$.

4.2. Raman Experiment-Crystallization of Paracetamol in Methanol

In this example, we show that the determinant of $\mathcal{S}$, defined in (12), is an effective measure for the physical meaningful number of compounds.

Crystallization of paracetamol in methanol (Figure 7) shows a solvated paracetamol in methanol, a metastable intermediate phase, and a crystalline phase. For detailed explanations, including the experimental settings, see Appendix A. Raman shifts characterize different paracetamol phases (solution phase, metastable intermediate phase and form II). After applying the SepFree NMF method, the role that the obtained compounds have in the reaction can be identified by the position of the peaks in matrix W.

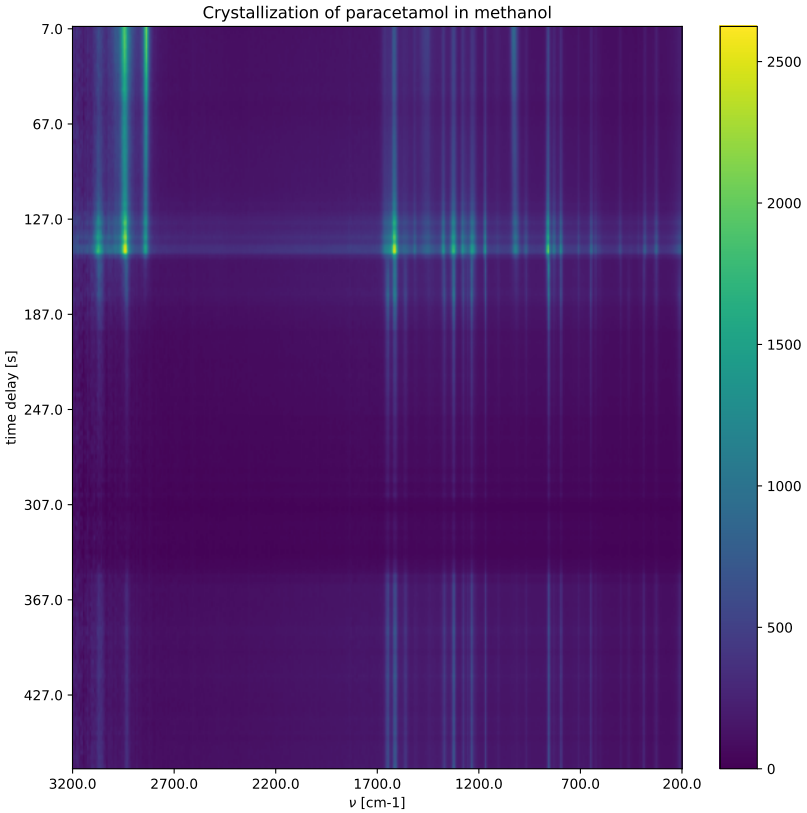

Figure 7. Raman spectrum of the crystallization of paracetamol in methanol. The intensities of the Raman shifts are presented as function of delay time t and Raman shift ν (cm^{-1}) .

Firstly, the singular value decomposition of the sequentially measured unprocessed spectral data taken as matrix suggests that the number of compounds can probably be

$r = 3, 4, 5$. Thus, SepFree NMF is tested then for these r values. The parameters used for the objective function are $\alpha = 0.1$, $\beta = 1000.0$, $\gamma = 50.0$, $\delta = 1.0$, $\mu = 1.0$.

The results of the application of the SepFree NMF are shown in Figure 8 for $r = 3$, in Figure 9 for $r = 4$, and in Figure 10 for $r = 5$. In each figure, the left plot shows the row of the matrix H as function of the delay time t; the right plot displays the peak intensities of W as function of the Raman shift (cm^{-1}). The sum of the relative concentrations $\sum H_t$ is also plotted on the left below, since the sum of the compounds can change. This change is due to variations of state occurring in the reaction, which modify the relative concentrations of the paracetamol moieties.

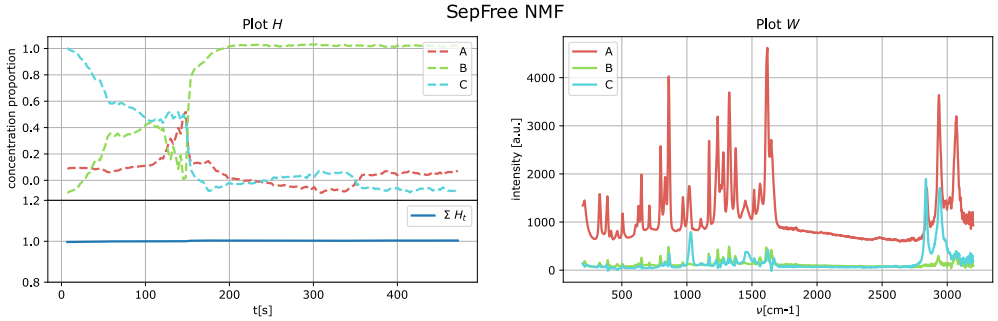

Figure 8. SepFree NMF analysis of the crystallization Raman spectrum with $r = 3$. Minimal memory effect is $\det(\mathcal{S}) = 0.106$.

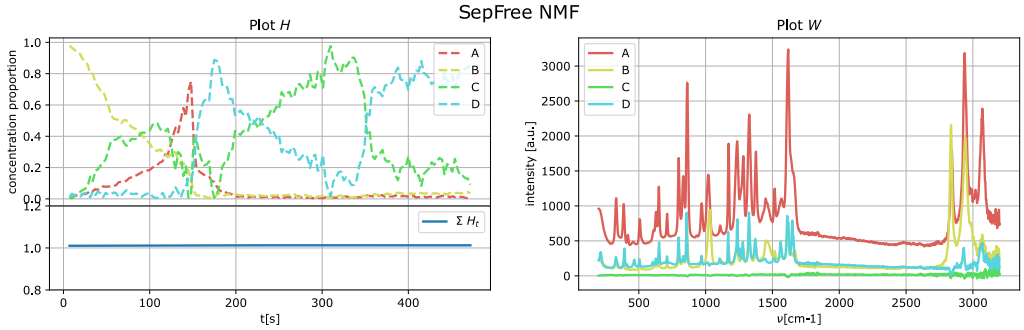

Figure 9. SepFree NMF analysis of the crystallization Raman spectrum with $r = 4$. Minimal memory effect is $\det(\mathcal{S}) = 0.031$.

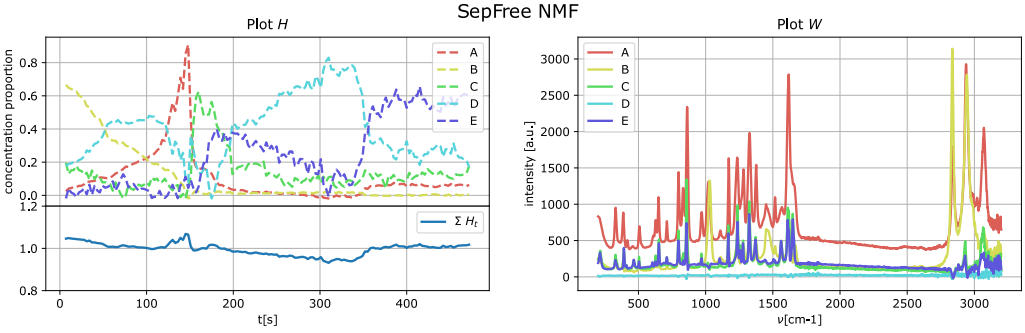

Figure 10. SepFree NMF NMF analysis of the crystallization Raman spectrum with $r = 5$. Minimal memory effect is $\det(\mathcal{S}) = 0.001$.

The decomposition with $r = 4$ and $r = 5$ compounds show kinetic curves in which the state's proportion rises, decreases and rises again, as it can be seen in Figures 9 and 10. This pattern cannot explain the state of a compound in a crystallization process. In contrast, the decomposition in Figure 8 for $r = 3$ describes a sequential kinetic process, as expected for a crystallization.

Can we estimate the number of reaction compounds only by looking at $\det(\mathcal{S})$?

This question can be answered with *yes* by Table 1. In this table, the values of $\det(\mathcal{S})$ for $r = 3, 4, 5$ are given. It can be noticed that for $r = 3$ the determinant has the highest value, in accordance with the above considerations on the kinetics.

Table 1. Number of compounds r in the decomposition and respective minimal memory effect, $\det(\mathcal{S})$. The decomposition with the smallest minimal memory effect is the one with $r = 3$. The minimal memory effect for $r = 4$ and $r = 5$ is orders of magnitude greater.

r	$\det(\mathcal{S})$
3	0.106
4	0.031
5	0.001

The SepFree NMF yields also a transition matrix $K(\tau)$ for the process [6]. Examining the transition matrix allows to understand the kinetic model underlying the reaction mechanism [18]. The crystallization process of this dataset is a sequential process.

The matrix $K(\tau)$ resulting from the decomposition

$$K(\tau) = \begin{pmatrix} 0.94 & 0.16 & -0.1 \\ 0. & 1., & 0. \\ 0.01 & -0. & 0.99 \end{pmatrix} \tag{16}$$

shows also a sequential process $C \to A \to B$ and it shows that B is an absorbing state of the Markovian process. This means that when the process reaches B, it does not leave that condition anymore, as also shown in Figure 11. The first compound for the analysis of the matrix $K(\tau)$ is C since C is the most abundant at time $t = 0$, see Figure 8.

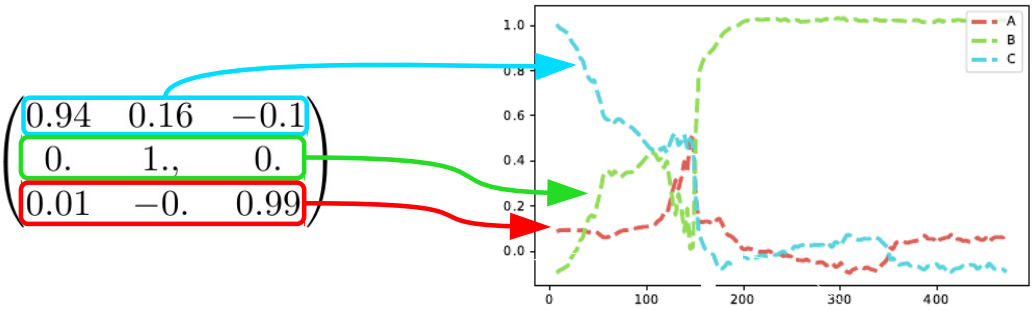

Figure 11. Each row in the transition probability matrix $K(\tau)$ represents the trend of a chemical moiety.

In the supplementary materials, we provide a similar analysis for the decomposition of ranks 4 and 5, showing that the kinetic model described is not sequential.

Recognizing the chemical moieties from the Raman shifts is advantageous for validating the number of compounds estimated with the minimal memory effect, as shown in the following analysis. To show that the results given by the decomposition of rank 3 are indeed in agreement with the experiment, we plot each peak intensity of W (Figure 8) with the peaks of form II and metastable amorphous form II. The values of the Raman peaks are reported in the supplementary materials.

Spectrum C (Figure 12) shows only the peaks of the metastable amorphous phase and two clear peaks in the 2800–3000 cm^{-1} region, which can be assigned to methanol [25].

Spectrum A (Figure 13) presents the peaks characteristic of the metastable amorphous phase and solvent, as methanol Raman shifts are still present, in agreement with [25].

Spectrum B (Figure 14) shows the peaks of the crystallized form II; neither Raman shifts of the metastable amorphous phase nor methanol shifts are clearly present.

In conclusion, the decomposition analysis reveals three stages of the reaction:

1. C: solution phase paracetamol (paracetamol + methanol);

2. A: metastable amorphous phase and methanol;
3. B: crystallized form II.

In this way, only by computing the minimal memory effect, we obtain a decomposition (a) that describes the correct kinetic process, (b) whose spectral intensities have a precise and clear interpretation.

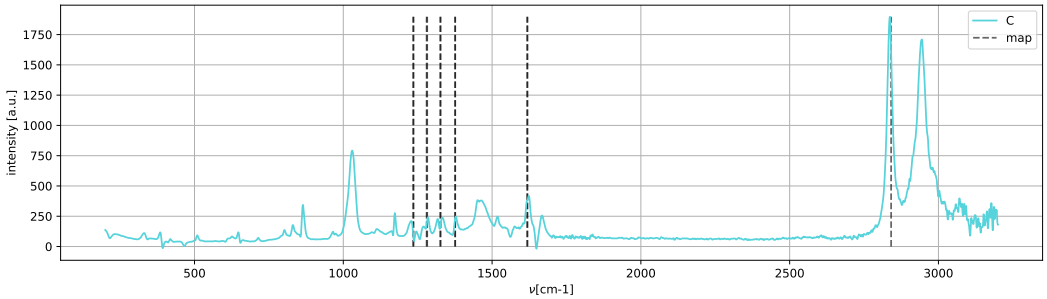

Figure 12. Spectrum C, assigned to solution phase paracetamol.

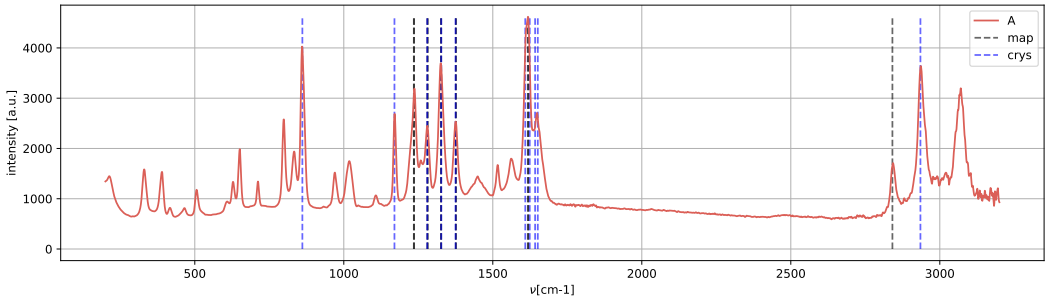

Figure 13. Spectrum A, metastable amorphous phase in presence of solvent.

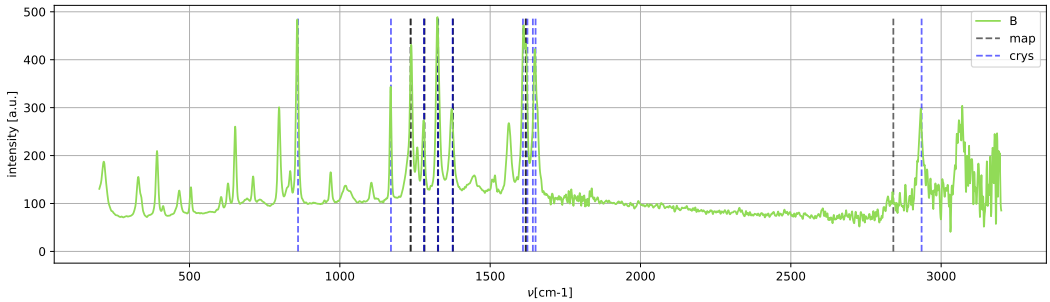

Figure 14. Spectrum B, assigned to crystallized form II.

5. Conclusions

SepFree NMF is a new method to process sequentially measured spectra and estimating the number of components. It is available as Python code using the link in the Data Availability Statement. This link includes a script for the presented figures as well as a tutorial to reproduce the results. Former methods extracting the kinetics of a molecular process observed in a sequential spectroscopic experiment need prior knowledge about the number of chemical moieties and/or their spectroscopic fingerprints. For some methods, the separability assumption has to apply, which in principle also means that the pure spectroscopic fingerprints of the compounds are to be found in the measured spectrum. SepFree NMF can be applied without this kind of prior knowledge.

SepFree NMF simultaneously generates the kinetic curves together with the spectroscopic fingerprints. How does it differ from other methods based on experimental data? On top of the measured data, it additionally uses the fact that the projected macro process is assumed to be Markovian. Thus, there should a transition matrix $K(\tau)$, and there should be as few memory effects as possible in the projection. We show that $\det(\mathcal{S})$ is a good indicator

for the memory effect. This *structural information* is the conceptual basis for formulating an optimization problem.

Using SepFree NMF, it is possible to extract the kinetics together with the number of compounds of a crystallization process observed with Raman spectroscopy. This would not be possible with methods which rely on the separability assumption.

Supplementary Materials: Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/a15090297/s1, which is from our GitHub repository: reproduction of all the figures, tutorial for the data analysis.

Author Contributions: Conceptualization, R.S., K.F. and M.W.; Investigation, R.S., K.F. and S.C.; Project administration, K.F. and M.W.; Resources, S.C.; Software, R.S.; Supervision, K.F. and M.W.; Writing—original draft, R.S. and K.F.; Writing—review and editing, R.S. and K.F. All authors have read and agreed to the published version of the manuscript.

Funding: Deutsche Forschungsgemeinschaft: EXC-2046/1, project ID 390685689.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/rena-96/SepFree-NMF/releases/tag/Algorithms (accessed on 7 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Crystallization of Paracetamol

Paracetamol is known to crystallize in five forms, forms I–V [26,27]. In practice, parameters such as the cooling rate and choice of solvent are optimized to obtain the desired polymorph. Form II is obtained when methanol is used as a solvent. We followed the crystallization process of paracetamol in methanol using sequential in situ Raman spectroscopy. The solution of paracetamol in methanol was prepared to the half of their respective saturation concentration, as reported by Granberg and Rasmuson for sample to only supersaturate in the sample holder. [28] A methanolic droplet of the aforementioned paracetamol was suspended in a custom-made acoustic levitator for conducting contact-free crystallization studies. Raman spectra were acquired using a process Raman instrument (Raman RXN1^{TM} Kaiser Optical Systems, Ecully, France) at 785 nm wavelength. A contactless probe with a probe size of 1 mm aperture was used for these experiments. The spectra were recorded for 1 s each with a processing time of 1.5 s in a total yielding resolution of 3 s was recorded. The analyzed dataset is plotted in Figure 7 as a 2D heatmap as a function of wavenumber and delay time.

As the experiments proceeds, the methanol evaporation leads to gradual supersaturation in the droplet. Eventually, paracetamol crystallizes in its crystalline form II. The process of crystal formation of paracetamol from its solution phase is not straightforward. While the solvent evaporates, a metastable intermediate phase of paracetamol is formed (MIP). This metastable intermediate phase is believed to be amorphous in nature and proto-crystalline in nature. Paracetamol crystallizes to form II from its metastable amorphous phase [25].

Appendix B. Heuristic Method for the Choice of the Parameters in Ψ

The coefficients of the addends in the objective function Ψ were determined with a heuristic method. This method is based on previous numerical experiments in [29].

When weighting the terms in Ψ, one does not want to overweight or underweight any of them, so they should have the same or a similar order of magnitude. For instance, if the Raman intensities are in the range 10^4, the coefficient α would be 10^{-4}, so that we can add it to the other terms of magnitude 1–10. Practically, one wants to "weaken" the weight of the spectral intensities to also consider the other terms. With the same logic, β and γ will represent H, whose entries are between $[0, 1]$, so a first guess will be of magnitude 10.

After this first step, one can tune the values of the coefficients so that the decomposition has meaningful results, e.g., all the Raman intensities in W and the state proportions in H are positive when choosing $\alpha = 1.5$ instead of $\alpha = 1.4$.

References

1. Risken, H. *The Fokker-Planck Equation*; Springer: Berlin/Heidelberg, Germany, 1996.
2. Zwanzig, R. Memory Effects in Irreversible Thermodynamics. *Phys. Rev.* **1961**, *124*, 983–992. [CrossRef]
3. Vauquelin, G.; Bricca, G.; Van Liefde, I. Avidity and positive allosteric modulation/cooperativity act hand in hand to increase the residence time of bivalent receptor ligands. *Fundam. Clin. Pharmacol.* **2014**, *28*, 530–543. [CrossRef] [PubMed]
4. Satija, R.; Das, A.; Makarov, D.E. Transition path times reveal memory effects and anomalous diffusion in the dynamics of protein folding. *J. Chem. Phys.* **2017**, *147*, 152707. [CrossRef] [PubMed]
5. Weber, M.; Fackeldey, K. Computing the Minimal Rebinding Effect Included in a Given Kinetics. *Multiscale Model. Simul.* **2014**, *12*, 318–334. [CrossRef]
6. Fackeldey, K.; Röhm, J.; Niknejad, A.; Chewle, S.; Weber, M. Analyzing Raman spectral data without separabiliy assumption. *J. Math. Chem.* **2021**, *59*, 575–596. [CrossRef]
7. Schrader, B. *Infrared and Raman Spectroscopy*; Wiley-VCH: Weinheim, Germany, 1995.
8. Ferraro, J.R.; Nakamotot, K.; Brown, C.W. *Introductory Raman Spectroscopy*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2003.
9. Smith, E.; Dent, G. *Modern Raman Spectroscopy*; Wiley-VCH: Weinheim, Germany, 2005.
10. Ruckebusch, C.; Blanchet, L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal. Chim. Acta* **2013**, *765*, 28–36. [CrossRef]
11. Mack, E.T.; Snyder, P.W.; Perez-Castillejos, R.; Whitesides, G.M. Using Covalent Dimers of Human Carbonic Anhydrase II To Model Bivalency in Immunoglobulins. *J. Am. Chem. Soc.* **2011**, *133*, 11701–11715. [CrossRef] [PubMed]
12. Rao, J.; Lahiri, J.; Weis, R.M.; Whitesides, G.M. Design, Synthesis, and Characterization of a High-Affinity Trivalent System Derived from Vancomycin and l-Lys-d-Ala-d-Ala. *J. Am. Chem. Soc.* **2000**, *122*, 2698–2710. [CrossRef]
13. Errington, W.J.; Bruncsics, B.; Sarkar, C.A. Mechanisms of noncanonical binding dynamics in multivalent protein–protein interactions. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 25659–25667. [CrossRef] [PubMed]
14. Luce, R.; Hildebrandt, P.; Kuhlmann, U.; Liesen, J. Using separable nonnegative matrix factorization techniques for the analysis of time-resolved raman spectra. *Appl. Spectrosc.* **2016**, *70*, 1464–1475. [CrossRef] [PubMed]
15. Kumar, A.; Sindhwani, V.; Kambadur, P. Fast Conical Hull Algorithms for Near-separable Non-negative Matrix Factorization. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; Dasgupta, S., McAllester, D., Eds.; PMLR: Atlanta, GA, USA, 2013; Volume 28, pp. 231–239.
16. Kijima, M. *Markov Processes for Stochastic Modeling*; Springer: Berlin/Heidelberg, Germany, 2013.
17. Vavasis, S.A. On the Complexity of Nonnegative Matrix Factorization. *SIAM J. Optim.* **2010**, *20*, 1364–1377. [CrossRef]
18. Sechi, R. Unravelling the Kinetics of Time-Resolved Spectra by Matrix-Factorization without Separability Assumption and by Markov State Modeling with PCCA+ Projection. Master's Thesis, FU Berlin, Berlin, Germany, 2021. [CrossRef]
19. Röblitz, S. Statistical Error Estimation and Grid-Free Hierarchical Refinement in Conformation Dynamics. Ph.D. Thesis, FU Berlin, Berlin, Germany, 2009.
20. Kube, S.; Weber, M. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.* **2007**, *126*, 024103. [CrossRef] [PubMed]
21. Williams, M.O.;Kevrekidis, I.G.; Rowley, C.W. A Data–Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *J. Nonlinear Sci.* **2015**, *25*, 1307–1346. [CrossRef]
22. Degennaro, A.; Urban, N.M. Scalable Extended Dynamic Mode Decomposition Using Random Kernel Approximation. *SIAM J. Sci. Comput.* **2019**, *41*, A1482–A1499. [CrossRef]
23. Klus, S. Data-Driven Analysis of Complex Dynamical Systems. Ph.D. Thesis, FU Berlin, Berlin, Germany, 2020.
24. Heida, M.; Kantner, M.; Stephan, A. Consistency and convergence for a family of finite volume discretizations of the Fokker-Planck operator. *arXiv* **2021**, arXiv:2002.09385.
25. Nguyen Thi, Y.; Rademann, K.; Emmerling, F. Direct evidence of polyamorphism in paracetamol. *CrystEngComm* **2015**, *17*, 9029–9036. [CrossRef]
26. Szelagiewicz, M.; Marcolli, C.; Cianferani, S.; Hard, A.; Vit, A.; Burkhard, A.; Von Raumer, M.; Hofmeier, U.; Zilian, A.; Francotte, E.; et al. In situ characterization of polymorphic forms: The potential of Raman techniques. *J. Therm. Anal. Calorim.* **1999**, *57*, 23–43. [CrossRef]
27. Perrin, M.A.; Neumann, M.A.; Elmaleh, H.; Zaske, L. Crystal structure determination of the elusive paracetamol Form III. *Chem. Commun.* **2009**, *22*, 3181–3183. [CrossRef] [PubMed]
28. Granberg, R.A.; Rasmuson, Å.C. Solubility of paracetamol in pure solvents. *J. Chem. Eng. Data* **1999**, *44*, 1391–1395. [CrossRef]
29. Röhm, J. Non-Negative Matrix Factorization for Raman Data Spectral Analysis. Master's Thesis, Technische Universitat Berlin, Berlin, Germany, 2017.