entropy

MDPI

Article

Probabilistic PARAFAC2

Philip J. H. Jørgensen [1], Søren F. Nielsen [1], Jesper L. Hinrich [1], Mikkel N. Schmidt [1], Kristoffer H. Madsen [1,2] and Morten Mørup [1,*]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark; jehi@dtu.dk (J.L.H.); mnsc@dtu.dk (M.N.S.); kristofferm@drcmr.dk (K.H.M.)

[2] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Amager and Hvidovre, 2650 Hvidovre, Denmark

* Correspondence: mmor@dtu.dk

Abstract: The Parallel Factor Analysis 2 (PARAFAC2) is a multimodal factor analysis model suitable for analyzing multi-way data when one of the modes has incomparable observation units, for example, because of differences in signal sampling or batch sizes. A fully probabilistic treatment of the PARAFAC2 is desirable to improve robustness to noise and provide a principled approach for determining the number of factors, but challenging because direct model fitting requires that factor loadings be decomposed into a shared matrix specifying how the components are consistently co-expressed across samples and sample-specific orthogonality-constrained component profiles. We develop two probabilistic formulations of the PARAFAC2 model along with variational Bayesian procedures for inference: In the first approach, the mean values of the factor loadings are orthogonal leading to closed form variational updates, and in the second, the factor loadings themselves are orthogonal using a matrix Von Mises–Fisher distribution. We contrast our probabilistic formulations to the conventional direct fitting algorithm based on maximum likelihood on synthetic data and real fluorescence spectroscopy and gas chromatography–mass spectrometry data showing that the probabilistic formulations are more robust to noise and model order misspecification. The probabilistic PARAFAC2, thus, forms a promising framework for modeling multi-way data accounting for uncertainty.

Keywords: tensor decomposition; multi-way modeling; variational inference; orthogonality constraint; PARAFAC2

Citation: Jørgensen, P.J.H.; Nielsen, S.F.; Hinrich, J.L.; Schmidt, M.N.; Madsen, K.H.; Mørup, M. Probabilistic PARAFAC2. *Entropy* **2024**, *26*, 697. https://doi.org/10.3390/e26080697

Academic Editor: Nikolai Leonenko

Received: 13 June 2024
Revised: 9 August 2024
Accepted: 13 August 2024
Published: 17 August 2024

1. Introduction

Tensor decompositions are multi-way generalizations of matrix decompositions such as principal component analysis (PCA): A matrix is a second-order array with two modes, rows and columns, while a data cube is a third order array with the third mode referred to as slabs. When multi-way data have an inherent multi-linear structure, the advantage of tensor decomposition methods is that they capture this intrinsic information and often provide a unique representation without needing further constraints such as sparsity or statistical independence.

Applications of tensor factorization originated within the field of psychometrics [1,2] and have been widely useful in other fields such as chemometrics [3], for example, to model the relationship between excitation and emission spectra of samples of different mixed compounds obtained by fluorescence spectroscopy [4]. Tensor decomposition is today encountered in practically all fields of research including signal processing, neuroimaging, and information retrieval (see also [5–7]).

The two most prominent tensor decomposition methods are (i) the Tucker model [8], where the so-called core array accounts for all multi-linear interactions between the components of each mode, and (ii) the CandeComp/PARAFAC (CP) model [1,2,9], where

interactions are restricted to be between components of identical indices across modes, corresponding to a Tucker model with a diagonal core array. Both models can be considered generalizations of PCA to higher-order arrays, with the Tucker model being more flexible at the expense of reduced interpretability. The CP model has been widely used primarily due to its ease of interpretation and its uniqueness [6,10].

In the CP model, the components are assumed identical across measurements, varying only in their scaling. In many situations, this is too restrictive—for example, when the number of samples vary across a mode. Furthermore, violation of the CP structure within chemometrics can be caused by retention time shifts [11,12], whereas in neuroimaging, such violations may be induced by subject and trial variability [6] invalidating the use of the CP model. To handle variability while preserving the uniqueness of the representation, the Parallel Factor Analysis 2 (PARAFAC2) model was proposed [2]. It admits individual loading matrices for each entry in a mode while preserving uniqueness properties of the decomposition by imposing consistency of the Gram matrix (i.e., the loading matrix left multiplied by its transpose, thereby imposing consistency in how components are co-expressed across samples) [13–15]. It has since been applied within diverse application domains, including handling variations in elution profiles due to retention shifts in chromatography [11]; monitoring and fault detection facing unequal batch lengths in chemical processes [16]; in neuroimaging to analyze latency changes in frequency resolved evoked EEG potentials [17], to extract common connectivity profiles in multi-subject fMRI data accounting for individual variability [18], and to characterize dynamic functional connectivity [19]; for cross-language information retrieval [20]; as well as for music and image tagging [21,22]. Recently, efforts have been made to scale the PARAFAC2 model to large-scale data [23–25], enhance the robustness and efficiency of the conventional direct fitting algorithm [26,27], and apply a non-negativity constraint also on the varying mode [28,29] as well as broader sets of constraints based on alternating directions of the method of multipliers [30].

Traditionally, tensor decompositions have been based on maximum likelihood inference using alternating least squares estimation in which the components of a mode are estimated while keeping the components of other modes fixed. Initial probabilistic approaches defined probability distributions over the component matrices and the core array but relied on maximum likelihood estimates for determining a solution [31,32]. However, the Bayesian approach presented here makes inference with respect to the posterior distributions of the model parameters and can thus be used to assess uncertainty in the parameters and noise estimates. Most work on probabilistic tensor decomposition has focused on the TUCKER and CP models using either Markov Chain Monte Carlo (MCMC) sampling [33–35] or variational inference [36–39]. The CP and Tucker models have been extended to model sparsity [35,40,41], non-negativity [42], and non-linearity [33,43] in component loadings. Heteroscedastic noise modeling has been discussed in the context of the CP model [41,44,45] and Tucker model [46], the latter also providing a generalization of tensor decomposition to exponential family distributions. A review and toolbox for probabilistic tensor decompositions are given in [45]. For component matrices with orthogonal components, recent work has explored using the von Mises–Fisher Matrix (vMF) distribution in the CP model [47] and the block-term decomposition model defined as a sum of Tucker models [48]. The former used a MAP-based estimation—which is not a fully Bayesian approach—and the latter used a variational Bayesian inference approach. In addition to using a variational Bayesian inference approach to the vMF distribution, we also explore another orthogonal formulation that is applicable beyond the PARAFAC2 structure.

Benefits of probabilistic modeling include the ability to account for uncertainty and noise while providing tools for model order selection. Whereas probabilistic modeling can be directly applied to the CP and TUCKER models extending probabilistic PCA [49], a probabilistic treatment of the PARAFAC2 model faces the following two key challenges:

(i) The ability to impose orthogonality on variational factors (necessary for imposing the PARAFAC2 structure).
(ii) Handling the coupling of these orthogonal components.

In this paper, we address these two challenges and derive the probabilistic PARAFAC2 model. In particular, we investigate two different formulations of the orthogonality constraint and demonstrate how the orthogonality of variational factors as in the least squares estimation for conventional PARAFAC2 can be obtained in closed form using the singular value decomposition. We exploit how the probabilistic framework admits model order quantification by the evaluation of model evidence and automatic relevance determination. We contrast our probabilistic formulation to conventional maximum likelihood estimation on synthetic data as well as fluorescence spectroscopy and gas chromatography–mass spectrometry data, highlighting the utility of the probabilistic formulation facing noise and model order misspecification (A short workshop contribution in brief presenting the proposed probabilistic PARAFAC2 was presented in [50]).

2. Methods

The three-way CP model can be formulated as a series of coupled matrix decompositions,

$$\boldsymbol{X}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top + \boldsymbol{E}_k,$$

where $\boldsymbol{X}_k \in \mathbb{R}^{I\times J}$ is the k'th slab of the three-way array $\mathcal{X}$ with dimensions $I \times J \times K$. Let M be the number of components in the model; then, the matrix $\boldsymbol{A}$ with dimensions $I \times M$ contains the loadings for the first mode and $\boldsymbol{F}$ with dimensions $J \times M$ contains the loadings for the second mode. The matrices $\boldsymbol{D}_k$, $k = 1, \ldots, K$, are diagonal with dimensions $M \times M$ and contain the loadings for the third mode. These are usually written as a single matrix $\boldsymbol{C} \in \mathbb{R}^{K\times M}$, where the k'th row contains the diagonal of $\boldsymbol{D}_k$. $\boldsymbol{E}_k$ denotes the residuals for the k'th slab with dimensions $I \times J$. Notice that the structure of the first and second mode are invariant across the third mode in this model.

The PARAFAC2 model extends the CP structure by letting a mode have individual factors $\boldsymbol{F}_k$ for each slab. The extension allows for a varying number of observations in the chosen mode. This model would be as flexible as PCA on the concatenated data $[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_K]$ if not for the additional constraint that each Gram matrix of $\boldsymbol{F}_k$ be identical, $\boldsymbol{F}_k^\top \boldsymbol{F}_k = \boldsymbol{\Psi}$, which is a necessary constraint in order to obtain unique solutions [51]. The three-way PARAFAC2 model can thus be written as

$$\boldsymbol{X}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}_k^\top + \boldsymbol{E}_k \quad \text{s.t. } \boldsymbol{F}_k^\top \boldsymbol{F}_k = \boldsymbol{\Psi}.$$

Modeling $\boldsymbol{\Psi}$ explicitly can be difficult, but it is necessary and sufficient [15] to have $\boldsymbol{F}_k = \boldsymbol{P}_k\boldsymbol{F}$, with $\boldsymbol{P}_k$ being a columnwise orthogonal $J \times M$ matrix and $\boldsymbol{F}$ a $M \times M$ matrix; thus, the model can be written as

$$\boldsymbol{X}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top + \boldsymbol{E}_k \text{ s.t. } \boldsymbol{P}_k^\top\boldsymbol{P}_k = \mathbf{I}. \tag{1}$$

In the following, we describe the conventional direct fitting algorithm [15] for parameter estimation in the PARAFAC2 model before we introduce the probabilistic model formulation in Section 2.3.

2.1. Direct Fitting Algorithm

The parameters in the PARAFAC2 model in (1) can be estimated using the alternating least squares algorithm [15], minimizing the constrained least squares objective function,

$$\underset{\boldsymbol{A},\boldsymbol{F},\{\boldsymbol{P}_k,\boldsymbol{D}_k\}\,\forall_k}{\arg\min} \sum_k \|\boldsymbol{X}_k - \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top\|^2 \quad \text{s.t. } \boldsymbol{P}_k^\top\boldsymbol{P}_k = \mathbf{I}.$$

For fixed $\boldsymbol{A}$, $\boldsymbol{D}_k$, and $\boldsymbol{F}$, the $\boldsymbol{P}_k$ that minimizes the k'th term in the objective function is equal to

$$\underset{\boldsymbol{P}_k}{\arg\max}\, \mathrm{Tr}(\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{A}^\top\boldsymbol{X}_k\boldsymbol{P}_k) \tag{2}$$

and can be computed as [15,52]

$$\boldsymbol{P}_k = \boldsymbol{V}_k \boldsymbol{U}_k^\top \tag{3}$$

where $\boldsymbol{V}_k$ and $\boldsymbol{U}_k$ come from the singular value decomposition (SVD) decomposition

$$\boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^\top = \boldsymbol{F}\boldsymbol{D}_k \boldsymbol{A}^\top \boldsymbol{X}_{\boldsymbol{k}}.$$

Upon fitting $\boldsymbol{P}_k$, each slab $\boldsymbol{X}_k$ of the tensor can be projected onto $\boldsymbol{P}_k$, thereby leaving the remaining parameters to be fitted as a CP model minimizing

$$\underset{\boldsymbol{A},\boldsymbol{F},\{\boldsymbol{D}_k\}}{\arg\min} \sum_k \|\boldsymbol{X}_k \boldsymbol{P}_k - \boldsymbol{A}\boldsymbol{D}_k \boldsymbol{F}^\top\|^2. \tag{4}$$

A solution to (4) is well explained by Bro in [3]. A well-known issue with maximum likelihood methods is that it can lead to overfitting due to noise and a lack of uncertainty in the model parameters, resulting in robustness issues, which we attempt to provide a solution for by advancing the PARAFAC2 model to a fully Bayesian setting.

Model Selection

A general problem for latent variable methods is how to choose the model order, M. A popular heuristic can be formed by how well the model fits the data given as

$$\mathrm{R2} = 1 - \frac{\sum_k \|\boldsymbol{X}_k - \boldsymbol{A}\boldsymbol{D}_k \boldsymbol{F}^\top \boldsymbol{P}_k^\top\|^2}{\sum_k \|\boldsymbol{X}_k\|^2}. \tag{5}$$

However, this measure will simply increase until the model incorporates enough parameters to completely fit the data, thus eventually leading to overfitting. The model selection criterion would only be based on the expected noise level.

Another popular heuristic is the core consistency diagnostic (CCD), originally developed for the CP model [53], but that has shown useful for the PARAFAC2 model as well [54]. It is based on the observation that the CP model can be seen as a constrained Tucker model, where the core array is enforced to be a superdiagonal array of ones. The principle behind CCD is to measure how much the CP model violates this assumption of a superdiagonal core array of ones by re-estimating the core array of the CP model to fit the Tucker model, denoted $\mathcal{G}$, while keeping the loadings fixed and then calculating the CCD according to

$$\mathrm{CCD} = 100\left(1 - \frac{||\mathcal{G} - \mathcal{I}||_{\mathcal{F}}^2}{||\mathcal{I}||_{\mathcal{F}}^2}\right),$$

in which $\mathcal{I}$ is the superdiagonal core array and $||\cdot||_{\mathcal{F}}$ denotes the Frobenius norm. The PARAFAC2 model can be written as a CP model for each slab as in (4); thus, the core array can be estimated in the same way as for the standard CP model. This approach was evaluated on synthetic as well as real data sets by [54], where the conclusion was that even though the CCD can be seen as an useful parameter for determining model order, it is not recommended in practice without considering other diagnostic measures, including inspecting the residuals and the loadings.

2.2. Variational Bayesian Inference

In Bayesian modeling, the posterior distribution of the parameters $\boldsymbol{\theta}$ is computed by conditioning on the observed data $\boldsymbol{X}$ using Bayes' rule, $p(\boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{X})$. The posterior is thereby given as the product of the likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ and the prior probability of the parameters $p(\boldsymbol{\theta})$ divided by the probability of the observed data $p(\boldsymbol{X})$ under the model, also known as the marginal likelihood. Evaluating the marginal likelihood is, in general, intractable; instead, a variational approximation can be found by fitting a

distribution $q(\boldsymbol{\theta})$—called the variational distribution—to the posterior [55] minimizing the Kullback–Leibler (KL) divergence, given by

$$q^\star(\boldsymbol{\theta}) = \arg\min \mathrm{KL}\big[q(\boldsymbol{\theta}) \big\| p(\boldsymbol{\theta}|\boldsymbol{X})\big].$$

Minimizing the KL divergence is solved by maximizing a related quantity, the evidence lower bound (ELBO).

$$\mathrm{ELBO}(q(\boldsymbol{\theta})) = \mathbb{E}[\log p(\boldsymbol{\theta}, \boldsymbol{X})] - \mathbb{E}[\log q(\boldsymbol{\theta})].$$

A common choice is a variational distribution that factorizes over the parameters, known as a mean-field approximation, $q(\boldsymbol{\theta}) = \prod_j q_j(\boldsymbol{\theta}_j)$. Note that, for convenience, we choose distributions belonging to the exponential family, as this allows closed-form solutions to be found. The optimal variational distribution can then be found by iterative updates of the form

$$q_j(\theta_j) \propto \exp\left(\mathbb{E}_{-j}[\log p(\theta_j, \boldsymbol{\theta}_{-j}, \boldsymbol{X})]\right), \tag{6}$$

where $\mathbb{E}_{-j}[\cdot]$ denotes the expectation over the variational distribution except q_j. For a comprehensive overview of variational inference, see for example [56,57], and for Bayesian inference in general, see [58].

2.3. Probabilistic PARAFAC2

We propose two probabilistic PARAFAC2 variants using the formulation in (1), which differ only in how the orthogonality of $\boldsymbol{P}_k$ is handled. The constraint $\boldsymbol{P}_k^\top \boldsymbol{P}_k = \boldsymbol{I}_M$ has the probabilistic interpretation that $\mathbb{E}[\boldsymbol{P}_k^\top \boldsymbol{P}_k] = \boldsymbol{I}_M$, in which the $\boldsymbol{P}_k$ is an orthogonal matrix, which we call model (i). Another interpretation is to enforce that the expected value $\mathbb{E}[\boldsymbol{P}_k]$ is an orthogonal matrix and implies $\mathbb{E}[\boldsymbol{P}_k]^\top \mathbb{E}[\boldsymbol{P}_k] = \boldsymbol{I}_M$—which we call model (ii). The main motivation for the latter approach being the interpretation of the orthogonal factor is identical to that of the maximum likelihood estimation. However, the resulting components are no longer themselves restricted to the set of orthogonal matrices, namely, the Stiefel manifold. As such, the model (ii) becomes more flexible as only the mean parameters of the variational approximation are constrained to be orthogonal and not the expectation of their inner product, as required for every realization of the underlying distribution to conform to the PARAFAC2 model. We include the latter model formulation, as it provides simple closed-form updates similar to the conventional direct fitting PARAFAC2 algorithm, as shown below. The updates for (ii) are derived by constraining the mean of a matrix normal ($\mathcal{MN}$) distribution within the variational approximation to the Stiefel manifold, whereas the model i) formulation is based on [59] and uses a matrix von Mises–Fisher (vMF) Matrix distribution, which only has support on the Stiefel manifold. We accordingly present the following two generative models, (i) and (ii), for the probabilistic PARAFAC2:

$$\begin{aligned}
&\boldsymbol{a}_{i\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M) \\
&\boldsymbol{f}_{m\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M) \\
&\boldsymbol{c}_{k\cdot} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha}^{-1})) \\
\text{(i)} \quad &\boldsymbol{P}_k \sim \mathrm{vMF}(\mathbf{0}) \\
\text{(ii)} \quad &\boldsymbol{P}_k \sim \mathcal{MN}\big(\mathbf{0}, \mathbf{I}_J, \mathbf{I}_M\big) \quad \text{s.t.} \quad \mathbb{E}[\boldsymbol{P}_k]^\top \mathbb{E}[\boldsymbol{P}_k] = \boldsymbol{I}_M \\
&\tau_k \sim \mathrm{Gamma}(a_{\tau_k}, b_{\tau_k}) \\
&\boldsymbol{X}_k \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top \boldsymbol{P}_k^\top, \tau_k^{-1}\mathbf{I}_J),
\end{aligned}$$

where $\boldsymbol{a}_{i\cdot}$ denotes the ith row of the matrix $\boldsymbol{A}$ and similarly for $\boldsymbol{f}_{m\cdot}$ and $\boldsymbol{c}_{k\cdot}$. We denote the set of all $\{\boldsymbol{P}_k\}_{k=1,2,\ldots,K}$ as $\mathcal{P}$. For the rate-scale Gamma distribution, the hyper-parameters a_{τ_k} and b_{τ_k} are user defined. $\boldsymbol{\alpha}$ defines the length scale of each component and can thus be used for automatic relevance determination (ARD) by turning off excess components by concentrating their distributions at zero when α_m is large [56]. In this paper, we use

the MAP estimate of α_m as we are more interested in the model's pruning ability than uncertainty on α_m. Pruning excess components is a challenging task, see [45] for ARD within Bayesian inference in the CP and Tucker models, and [60] for Bayesian shrinkage priors in general. Lastly, we allow the noise τ_k to vary across slabs, thereby accounting for potential different levels of the noise (i.e., assuming heteroscedastic noise) across slabs.

2.4. Variational Update Rules

The inference is based on the following factorized distribution,

$$\mathrm{q}(\boldsymbol{\theta}) = q(\boldsymbol{A})q(\boldsymbol{C})\prod_m q(\boldsymbol{f}_{m\cdot})\prod_k q(\boldsymbol{P}_k)q(\tau_k)$$

leading to the following ELBO,

$$\begin{aligned}\mathrm{ELBO}(\mathrm{q}(\boldsymbol{\theta})) &= \mathbb{E}[\log p(\mathcal{X},\boldsymbol{\theta})] - \mathbb{E}[\log \mathrm{q}(\boldsymbol{\theta})] \\ &= \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A},\boldsymbol{C},\boldsymbol{F},\mathcal{P},\boldsymbol{\tau})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{A})] \\ &\quad + \mathbb{E}[\log \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{\alpha})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{F})] \\ &\quad + \mathbb{E}[\log \mathrm{p}(\mathcal{P})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{\tau})] \\ &\quad + \mathrm{h}(\mathrm{q}(\boldsymbol{A})) + \mathrm{h}(\mathrm{q}(\boldsymbol{C})) + \mathrm{h}(\mathrm{q}(\boldsymbol{F})) \\ &\quad + \mathrm{h}(\mathrm{q}(\mathcal{P})) + \mathrm{h}(\mathrm{q}(\boldsymbol{\tau})).\end{aligned} \tag{7}$$

Expanding the variational factors, as given by (6), the resulting variational distributions and update rules are given in Table 1. The update for the factor matrix $\boldsymbol{F}$ is non-trivial, and to obtain a closed-form solution we employ a componentwise updating scheme inspired by the non-negative matrix factorization literature [61–63]. For each latent parameter, we use (6) and moment matching to determine the optimal variational distributions.

Table 1. Overview of all the variational factors and their updates. Note that $\mathcal{P} = \{\boldsymbol{P}_k\}_{k=1,2,\ldots,K}$ is the set of projection matrices and (SVD) indicates the expression is decomposed by singular value decomposition (SVD) to obtain $\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k$.

Variational Factor	Update
$\mathrm{q}(\boldsymbol{A}) \sim \prod_i \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a}_{i\cdot}}, \boldsymbol{\Sigma}_{\boldsymbol{a}_{i\cdot}})$	$\boldsymbol{\Sigma}_{\boldsymbol{a}_{i\cdot}} = \big(\mathbf{I}_M + \sum_k \mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k]\big)^{-1}$ $\boldsymbol{\mu}_{\boldsymbol{a}_{i\cdot}} = \boldsymbol{\Sigma}_{\boldsymbol{a}_{i\cdot}} \sum_k \mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top\boldsymbol{x}_{i\cdot k}^\top]$
$\mathrm{q}(\boldsymbol{C}) \sim \prod \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{c}_{k\cdot}}, \boldsymbol{\Sigma}_{\boldsymbol{c}_{k\cdot}})$	$\boldsymbol{\Sigma}_{\boldsymbol{c}_{k\cdot}} = \big(\mathrm{diag}(\boldsymbol{\alpha}) + \mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{F}^\top\boldsymbol{P}_k^\top\boldsymbol{P}_k\boldsymbol{F}] \circ \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\big)^{-1}$ $\boldsymbol{\mu}_{\boldsymbol{c}_{k\cdot}} = \boldsymbol{\Sigma}_{\boldsymbol{c}_{k\cdot}}\mathbb{E}[\tau_k]\mathrm{diag}(\mathbb{E}[\boldsymbol{F}^\top]\mathbb{E}[\boldsymbol{P}_k^\top]\boldsymbol{X}_k^\top\mathbb{E}[\boldsymbol{A}])$
$\mathrm{q}(\boldsymbol{F}) \sim \prod_m \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{f}_{m\cdot}}, \boldsymbol{\Sigma}_{\boldsymbol{f}_{m\cdot}})$	$\boldsymbol{\Sigma}_{\boldsymbol{f}_{m\cdot}} = \big(\sum_k \mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{D}_k\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{p}_{\cdot mk}^\top\boldsymbol{p}_{\cdot mk}] + \mathbf{I}_M\big)^{-1}$ $\boldsymbol{\mu}_{\boldsymbol{f}_{m\cdot}} = \boldsymbol{\Sigma}_{\boldsymbol{f}_{m\cdot}}\big(\sum_k \mathbb{E}[\tau_k]\{\mathbb{E}[(\boldsymbol{P}_k^\top)_m]\boldsymbol{X}_k^\top\mathbb{E}[\boldsymbol{A}]\mathbb{E}[\boldsymbol{D}_k]$ $-\mathbb{E}[\boldsymbol{D}_k\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{D}_k]\sum_{m'\setminus m}\mathbb{E}[\boldsymbol{p}_{\cdot mk}^\top\boldsymbol{p}_{\cdot m'k}]\boldsymbol{f}_{m'\cdot}^\top\}\big)$
$\mathrm{q}(\mathcal{P}) \sim \prod_k \mathrm{vMF}(\boldsymbol{B}_{\boldsymbol{P}_k})$	$\boldsymbol{B}_{\boldsymbol{P}_k} = \mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{F}]\mathbb{E}[\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{A}^\top]\boldsymbol{X}_k$ $\mathbb{E}[\boldsymbol{P}_k] = \boldsymbol{V}_k\boldsymbol{\Psi}\boldsymbol{U}_k^\top$, where $\boldsymbol{B}_{\boldsymbol{P}_k} = \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\top$ (SVD) ($\boldsymbol{\Psi}$ given by [59], Appendix A.2)
$\mathrm{q}(\mathcal{P}) \sim \prod_k c\mathcal{MN}(\boldsymbol{M}_{\boldsymbol{P}_k}, \mathbf{I}_J, \boldsymbol{\Sigma}_{\boldsymbol{P}_k})$	$\boldsymbol{\Sigma}_{\boldsymbol{P}_k} = (\mathbb{E}[\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top] + \mathbf{I})^{-1}$ $\boldsymbol{M}_{\boldsymbol{P}_k} = \boldsymbol{V}_k\boldsymbol{U}_k^\top$, where $\mathbb{E}[\tau_k]\mathbb{E}[\boldsymbol{F}]\mathbb{E}[\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{A}^\top]\boldsymbol{X}_k = \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\top$ (SVD)
$\mathrm{q}(\boldsymbol{\tau}) \sim \prod_k \mathrm{Gamma}(a_{\tau_k}, b_{\tau_k})$	$a_{\tau_k} = a_{\boldsymbol{\tau}} + \frac{IJ}{2}$ $b_{\tau_k} = \big(b_{\boldsymbol{\tau}}^{-1} + \frac{1}{2}\mathrm{Tr}(\boldsymbol{X}_k\boldsymbol{X}_k^\top) + \frac{1}{2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{A}^\top)]$ $-\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top\boldsymbol{X}_k^\top)]\big)^{-1}$
$\arg\max_{\alpha_m} \mathrm{ELBO}(\alpha_m)$	$\alpha_m = K\big(\sum_k \mathbb{E}[c_{km}^2]\big)^{-1}$

These updates rules are used for implementing a computational algorithm for probabilistic PARAFAC2, where each factor $\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}$, and $\boldsymbol{\tau}$ is updated conditionally on all other factors. This leads to an alternating optimization algorithm that, given an initial solution

(randomized or starting from the MAP solution), iteratively maximizes the evidence lower bound, Equation (7), until the relative change in ELBO is below a convergence criteria or a maximum number of iterations is reached. Finding the optimal solution is a non-convex optimization problem that is sensitive to initialization and the order of the updates.

2.4.1. Von Mises–Fisher Loading

In the von Mises–Fisher model for the loading $\boldsymbol{P}_k$, the variational distribution is given by

$$\mathrm{vMF}(\boldsymbol{P}_k|\boldsymbol{B}_{\boldsymbol{P}_k}) = \kappa(J, \boldsymbol{B}_{\boldsymbol{P}_k}^\top \boldsymbol{B}_{\boldsymbol{P}_k})^{-1} \exp\left(\mathrm{tr}[\boldsymbol{B}_{\boldsymbol{P}_k}^\top \boldsymbol{P}_k]\right),$$

which is defined on the Stiefel manifold, $\boldsymbol{P}_k^\top \boldsymbol{P}_k = \boldsymbol{I}$. The normalization constant is given by $\kappa = {}_0F_1\left(\frac{1}{2}J, \frac{1}{4}\boldsymbol{B}_{\boldsymbol{P}_k}^\top \boldsymbol{B}_{\boldsymbol{P}_k}\right) \mathrm{v}_{J,M}$, where $\mathrm{v}_{J,M}$ is the volume of the J-dimensional Stiefel manifold described by M components [64]. The hypergeometric function with matrix argument ${}_0F_1(\cdot,\cdot)$ can be calculated more efficiently using the SVD of $\boldsymbol{B}_{\boldsymbol{P}_k} = \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^\top$, since ${}_0F_1\left(\frac{1}{2}J, \frac{1}{4}\boldsymbol{B}_{\boldsymbol{P}_k}^\top \boldsymbol{B}_{\boldsymbol{P}_k}\right) = {}_0F_1\left(\frac{1}{2}J, \frac{1}{4}\boldsymbol{S}_k^2\right)$ [64].

Computing expectations over the vMF matrix distribution requires evaluating the hypergeometric function and can be performed as described by [59]. († Source code for approximating the hypergeometric function is available online http://staff.utia.cz/smidl/files/mat/OVPCA.zip (accessed on 28 February 2017). This code was used with default settings and without modifications in the experiments. We also share it with the probabilistic PARAFAC2 code at https://github.com/philipjhj/VBParafac2 (accessed on 28 February 2017)). Note that it follows from the vMF matrix distribution that $\mathbb{E}[\boldsymbol{P}_k^\top \boldsymbol{P}_k] = \mathbf{I}$, but in general $\mathbb{E}[\boldsymbol{P}_k]^\top \mathbb{E}[\boldsymbol{P}_k] \neq \mathbf{I}$. However, if an orthogonal summary representation is desired, one can inspect the mode of the vMF given by $\mathbf{U}_k \mathbf{V}_k^\top$.

2.4.2. Constrained Matrix Normal Loading

In the constrained matrix normal ($c\mathcal{MN}$) model for the variational factor of the loadings $\boldsymbol{P}_k$, we consider the distribution

$$c\mathcal{MN}(\boldsymbol{P}_k|\boldsymbol{M}_{\boldsymbol{P}_k}, \boldsymbol{I}_J, \boldsymbol{\Sigma}_{\boldsymbol{P}_k}) = \frac{\exp\left\{-\frac{1}{2}\mathrm{trace}\left(\boldsymbol{\Sigma}_{\boldsymbol{P}_k}^{-1}(\boldsymbol{P}_k - \boldsymbol{M}_{\boldsymbol{P}_k})^\top \boldsymbol{I}_J^{-1}(\boldsymbol{P}_k - \boldsymbol{M}_{\boldsymbol{P}_k})\right)\right\}}{(2\pi)^{IM/2}|\boldsymbol{\Sigma}_{\boldsymbol{P}_k}|^{I/2}|\boldsymbol{I}_J|^{M/2}},$$
$$\text{s.t.} \quad \boldsymbol{M}_{\boldsymbol{P}_k}^\top \boldsymbol{M}_{\boldsymbol{P}_k} = \boldsymbol{I}.$$

Instead of using the free form variational approach, we maximize (7) as a function of the mean parameter $\boldsymbol{M}_{\boldsymbol{P}_k}$ subject to the orthogonality constraint $\boldsymbol{M}_{\boldsymbol{P}_k} \boldsymbol{M}_{\boldsymbol{P}_k}^\top = \mathbf{I}_M$.

The constraint consequently causes (7) to be constant except for the linear term of the expected log of the probability density function of the data. The reason for this is that all other terms do not depend on $\boldsymbol{M}_{\boldsymbol{P}_k}$ or only on the matrix product $\boldsymbol{M}_{\boldsymbol{P}_k} \boldsymbol{M}_{\boldsymbol{P}_k}^\top$, which is equivalent to the identity matrix, resulting in the optimization problem

$$\arg\max_{\boldsymbol{M}_{\boldsymbol{P}_k}} \mathrm{ELBO}(\boldsymbol{M}_{\boldsymbol{P}_k}) \text{ s. t. } \boldsymbol{M}_{\boldsymbol{P}_k} \boldsymbol{M}_{\boldsymbol{P}_k}{}^\top = \mathbf{I}$$

where

$$\mathrm{ELBO}(\boldsymbol{M}_{\boldsymbol{P}_k}) = \sum_k \mathbb{E}[\tau_k] \mathrm{Tr}(\mathbb{E}[\boldsymbol{F}] \mathbb{E}[\boldsymbol{D}_k] \mathbb{E}[\boldsymbol{A}^\top] \boldsymbol{X}_k \boldsymbol{M}_{\boldsymbol{P}_k}) + c.$$

This is equal to (2) except for a scalar leading to the same solution as for the maximum likelihood estimation method, as given in (3). Detailed derivations of the expression above are given in the Appendices A and B. The variance parameter $\boldsymbol{\Sigma}_{\boldsymbol{P}_k}$ in the variational distribution follows from moment matching using (6).

2.4.3. The **F** Matrix

The updates for $f_{m\cdot}$ are non-trivial due to an inter-component dependency. The quadratic term in (6) for $\boldsymbol{F}$ is

$$\begin{aligned}
&\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top] \\
&\quad=\mathbb{E}_{-\boldsymbol{F}}[\mathrm{Tr}(\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k)] \\
&\quad=\mathrm{Tr}(\boldsymbol{F}\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\boldsymbol{F}^\top\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{P}_k{}^\top\boldsymbol{P}_k]) \\
&\quad=\sum_{mm\prime}(\boldsymbol{F}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\boldsymbol{F}^\top)_{mm\prime}(\mathbb{E}[\boldsymbol{P}_k{}^\top\boldsymbol{P}_k])_{mm\prime} \\
&\quad=\sum_{mm\prime}f_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbf{f}^{\mathrm{T}}_{m\prime\cdot}\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}] \\
&\quad=\sum_{m}f_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot mk}]\mathbf{f}^{\mathrm{T}}_{m\cdot} \\
&\qquad+2\sum_{m}\sum_{m\prime\setminus m}f_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}]\mathbf{f}^{\mathrm{T}}_{m\prime\cdot},
\end{aligned}$$

where we see that the quadratic term separates into a quadratic and linear part, revealing the linear inter-component dependency.

2.4.4. Non-Trivial Expectations

An overview of all the factors and their updates are given in Table 1. Below, we detail some non-trivial expectations and the necessary steps to compute them. The first group of expectations deals with having the diagonal matrix $\boldsymbol{D}_k$ left and right multiplied with an inner term. The first case is the following expectation,

$$\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]$$

which is equivalent to the Hadamard product of the outer product of the diagonal of the surrounding matrix with itself and the inner matrix; so, we can separate the expectation into two parts

$$\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]=\mathbb{E}[\boldsymbol{c}_{k\cdot}{}^\top\boldsymbol{c}_{k\cdot}]\circ\mathbb{E}[\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}],$$

where $\boldsymbol{c}_{k\cdot}$ is the vector containing the diagonal elements of $\boldsymbol{D}_k$. The same rule applies for the following expectation:

$$\mathbb{E}[\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k]=\mathbb{E}[\boldsymbol{c}_{k\cdot}{}^\top\boldsymbol{c}_{k\cdot}]\circ\mathbb{E}[\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}],$$

where the second expectation becomes trivial when using the vMF prior (ii) as the matrix product $\boldsymbol{P}_k{}^\top\boldsymbol{P}_k$ is the identity matrix. However, when using the matrix normal distribution (i), we obtain

$$\mathbb{E}[\boldsymbol{P}_k{}^\top\boldsymbol{P}_k]=\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{P}_k})+\mathbf{I}_M,$$

which leads to the element with index ij of the expectation to be equal to

$$\begin{aligned}
\mathbb{E}[\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}]_{ij}&=\mathbb{E}[\sum_m(\boldsymbol{F}^\top)_{im}(\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F})_{mj}] \\
&=\mathbb{E}[\sum_m\boldsymbol{F}^\top_{mi}\sum_{m\prime}(\boldsymbol{P}_k{}^\top\boldsymbol{P}_k)_{mm\prime}\boldsymbol{F}_{m\prime j}] \\
&=\sum_m\sum_{m\prime}\mathbb{E}[\boldsymbol{F}^\top_{mi}\boldsymbol{F}_{m\prime j}]\mathbb{E}[(\boldsymbol{P}_k{}^\top\boldsymbol{P}_k)_{mm\prime}].
\end{aligned}$$

Since the m'th and m' components are independent, we have

$$\mathbb{E}[\boldsymbol{F}_{mi}^{\top}\boldsymbol{F}_{m'j}] = \begin{cases} \mathbb{E}[\boldsymbol{F}_{mi}^{\top}]\mathbb{E}[\boldsymbol{F}_{m'j}] + (\boldsymbol{\Sigma}_{f_m.})_{ij} & \text{for } m = m'. \\ \mathbb{E}[\boldsymbol{F}_{mi}^{\top}]\mathbb{E}[\boldsymbol{F}_{m'j}] & \text{for } m \neq m'. \end{cases}$$

These are the most involved expectations when computing the update rules, and the remaining are either simpler or depend upon the expectations derived here.

2.5. Noise Modeling

The probabilistic formulation of PARAFAC2 requires the specification and estimation of the noise precision $\boldsymbol{\tau}$. We presently consider two specifications, i.e., homoscedastic noise in which the noise of each slab $\boldsymbol{X}_k$ is identical—i.e., $\tau_1 = \ldots = \tau_K$—as assumed in the direct fitting algorithm, and heteroscedastic noise, where the model includes a separate precision for each of the K slabs.

2.6. Model Selection

A benefit of a fully probabilistic formulation of the PARAFAC2 model is that it provides model order quantification using tools from Bayesian inference, see [45,60], respectively, for details in the context of probabilistic tensor models and Bayesian inference in general. We presently exploit automatic relevance determination by learning the length scale $\boldsymbol{\alpha}$, see also [56]. In practice, we use the MAP estimates for the automatic relevance determination because we are more interested in the pruning ability than the uncertainty estimates on $\boldsymbol{\alpha}$. If desired, a variational estimate is easily found by letting α_m follow a Gamma distribution, c.f. [49]. Finally, the estimated ELBO on the data can also be used to compare different model orders.

2.7. Computational Complexity

The computational complexity of probabilistic PARAFAC2—for a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$ with M components—is the same as its maximum likelihood alternative, namely, $O(I \cdot J \cdot K \cdot M + K \cdot M^3)$ where the first term stems from the matricized tensor Khatri–Rao product (MTTKRP) and the second from the inversion (or SVD) of an $M \times M$ matrix in connection with updating $\boldsymbol{P}_k$ for $k = 1, 2, \ldots, K$. The MTTKRP cannot be avoided, but caching of the sufficient statistics can make resulting calculations more efficient, although the computational complexity remains unchanged. Usually, I, J, and K are much greater than M; so, the MTTKRP becomes the limiting factor. Importantly, a limitation of the variational Bayesian formulation of PARAFAC2 is that one cannot directly use the projection trick of PARAFAC2, where the K mode is projected such that $\boldsymbol{X}_k\boldsymbol{P}_k = \boldsymbol{Y}_k$, and $\mathcal{Y}$ then has a PARAFAC structure. The trick relies on $\boldsymbol{P}_k$ being orthogonal, but in the variational formulation, the expectation of $\mathbb{E}[\boldsymbol{P}_k]$ is used instead of $\boldsymbol{P}_k$; thus, it is no longer exactly orthogonal. A remedy to this is either using sampling or a maximum a priori estimate for which $\boldsymbol{P}_k$ is exactly orthogonal, although neither approach changes the computational complexity of probabilistic PARAFAC2.

3. Results and Discussion

We evaluate the proposed models on both synthetic data and three real data sets: an amino acid fluorescence (AAF) data set and two gas chromatography–mass spectrometry (GC-MS) data sets. For comparison, we include the least squares PARAFAC2 direct-fit (Direct Fit) [15], probabilistic CP with normal distributed factors and a Gamma ARD-prior with either homoscedastic (VB PARAFAC Δ) or heteroscedastic (VB Parafac Ω) noise modeling, probabilistic Tucker (VB TUCKER) [48], and Bayesian relaxed matrix factorization (rMFT) [65]. For the proposed probabilistic PARAFAC2 methods, we initialize the model parameters as the PARAFAC2 solution computed using the direct fitting algorithm (as implemented by Bro [15] at http://www.models.life.ku.dk/go?filename=parafac2.m

(accessed on 13 October 2017)) and repeat the initialization five times for the synthetic data and 50 times for the real data to minimize the risk of getting stuck in a local extrema. The final model parameters are chosen as the parameters with the lowest R2 for the direct fitting models and the highest ELBO for the probabilistic models among the fitted models. Each model estimation is limited to 10^4 iterations for the synthetic data and 5×10^4 iterations for the real data. If the relative improvement in R2 for the direct fitting models and the ELBO for the probabilistic models after an iteration goes below 10^{-9}, we invoke an early stop. Empirically, we experienced better learning of the probabilistic models by keeping the precision parameter of the added noise fixed for some number of iterations while estimating the length scale $\boldsymbol{\alpha}$. We choose this delay to last for the first 50 iterations. The hyper-parameters of the precision were set to (shape,scale) = $(a_{\tau_k}, b_{\tau_k}) = (1, 10^{32})$ in order to be uninformative for the variational distribution, as their influence on the updated parameters is very small on the considered data sets.

3.1. Synthetic Data

To investigate the performance of the proposed model, we generate synthetic data sets in a similar manner as in [15]. We generated the data tensor $\mathcal{X}$ by sampling $\boldsymbol{A}$ from a zero-mean isotropic multivariate normal distribution with unit variance. $\boldsymbol{F}$ was taken from a Cholesky factorization of an $M \times M$ matrix with 1's in its diagonal and 0.4 in all the off-diagonal elements. This essentially keeps the M components from being too similar. Each element of $\boldsymbol{C}$ was sampled from a uniform distribution on the interval 0 to 30. $\boldsymbol{P}_k$ was constructed by the standard orthonormalization function in MATLAB of a set of vectors sampled from a zero-mean isotropic multivariate normal distribution with unit variance. The synthetic data sets were generated with either homoscedastic or heteroscedastic additive noise at different signal-to-noise ratios (SNR) in the interval $[-20, 10]$ dB, with increments of 2 dB. Each configuration was generated 10 times, resulting in 320 data sets. Each data set was given the dimensions $50 \times 50 \times 10$ with $M = 4$ components.

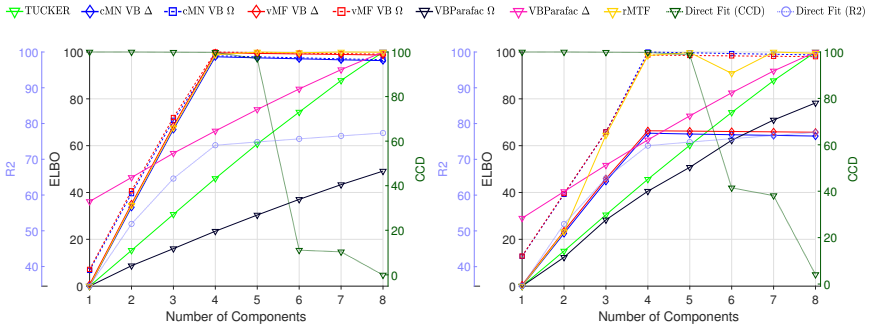
The probabilistic PARAFAC2 models were fitted to the data sets with the results on the synthetic data shown in Figures 1 and 2. To investigate the effect of the principled model selection approach based on the ELBO, we compare it to the existing model order selection heuristics by plotting the different selection criteria as a function of the number of components used in the model in Figure 1a,b. The figures show the mean result of the models fitted on the 10 synthetic data sets with four components and an SNR of 4. Overall, the ELBO suggests the same number of components as the other two criteria, R2 and CCD. When the data have heteroscedastic noise, the two probabilistic models that incorporate this have a substantially higher ELBO compared to the homoscedastic models.

The results for varying SNR using the true number of components in each model are shown in Figure 2a for data with homoscedastic noise and in Figure 2b for data with heteroscedastic noise. We report the R2 on the noiseless data, i.e., using the formula from (5), with the modification that the noise $\boldsymbol{E}_k$ has been subtracted from $\boldsymbol{X}_k$ for each slab. Thereby, we measure the different models' ability to capture the true underlying structure in the data.

On the homoscedastic data, we see a small advantage of using the two vMF models compared to the direct fitting algorithm when we decrease the SNR of the data. The $c\mathcal{MN}$ model performs slightly worse compared to the direct fitting algorithm. When we move to the heteroscedastic data, we see a stronger separation of the four different probabilistic methods. Naturally, the models with heteroscedastic noise outperform the ones with homoscedastic noise. It is also evident that the penalty of modeling the noise as heteroscedastic in a setting where the true noise is homoscedastic is small.

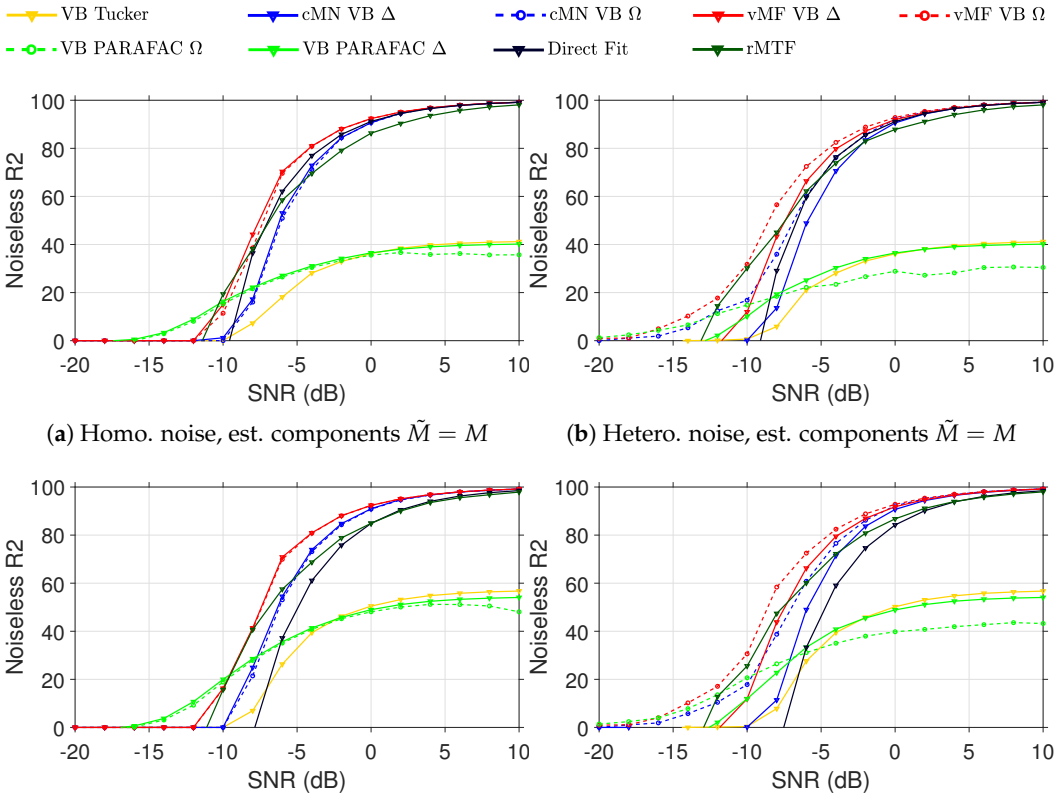
If the number of components is misspecified, see Figure 2c,d, we see a larger difference between the performance of the probabilistic models accounting for the heteroscedastic noise and the direct fitting algorithm. Here, we also observe that the vMF models perform better compared to the $c\mathcal{MN}$ parameterization and see a larger positive effect of using the probabilistic models over the direct fitting algorithm. This is mainly explained by the reduced tendency to overfit when accounting for the uncertainty and the automatic

relevance determination (ARD) pruning irrelevant components, as the Bayesian modeling promotes simpler representations by the ARD.



(**a**) (**b**)

Figure 1. Mean of model selection criteria R2 and CCD reported on the conventional PARAFAC2, and the ELBO for the TUCKER, rMTF, probabilistic PARAFAC, and probabilistic PARAFAC2 models, with 1 to 8 components on 10 synthetic data sets with added homoscedastic (**a**) and heteroscedastic (**b**) noise both with an SNR equal to 4. To make the results comparable, all ELBO values for each criterion and model (but across noise model types) have been normalized to be in the range of 0 to 100. In the legend, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.



(**a**) Homo. noise, est. components $\tilde{M} = M$ (**b**) Hetero. noise, est. components $\tilde{M} = M$

(**c**) Homo. noise, est. components $\tilde{M} = M + 2$ (**d**) Hetero. noise, est. components $\tilde{M} = M + 2$

Figure 2. Recovery of the underlying signal in synthetic data with varying levels of homoscedastic (**a**,**c**) and heteroscedastic (**b**,**d**) added noise, as measured by noiseless R2. For the conventional PARAFAC2 and probabilistic PARAFAC2 models fitted with both the true number of components ((**a**,**b**), with $M = \tilde{M} = 4$) and with an overspecified number of components ((**c**,**d**), with $\tilde{M} = 6$). In the legend, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

3.2. Real Data

As our synthetic results suggest, both formulations of the orthogonality constraint appear to be reasonable; we further investigate their performance on three real-world data sets. The first is an amino acid fluorescence (AAF) data set (available at www.ucphchemometrics.com (accessed on 28 February 2017), previously http://www.models.life.ku.dk/Amino_Acid_fluo) described in [61,66], in which the core-consistency diagnostic based on the PARAFAC2 model has previously successfully identified the three underlying constituents; tyrosine, tryptophan, and phenylalanine [54]. The data set contains five samples with 201 emission and 61 excitation intervals.

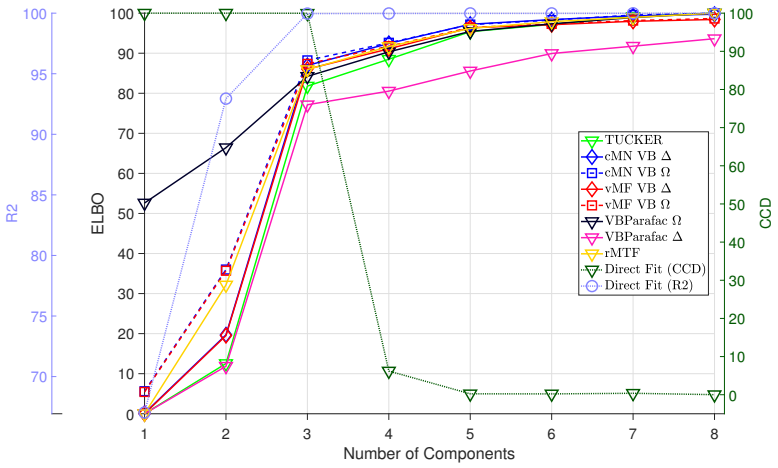
In addition, the models were evaluated on two gas chromatography–mass spectrometry (GC-MS) data sets. The first of these originated from wine (GC-MS-WINE) (available at www.ucphchemometrics.com (accessed on 28 February 2017), previously http://www.models.life.ku.dk/Wine_GCMS_FTIR) and was described in detail in [67]. PARAFAC2 has previously been used on GC-MS data obtained from measuring wine [54,68]. The second data set based on tobacco (GC-MS-TOBAC) was produced by [69] and kindly made available by the authors upon request. The GC-MS-WINE data contain 44 samples of wine; here, we specifically consider the unaligned data at the elution times 4.5903–4.7527 min over the mass range m/z 5–204. The GC-MS-TOBAC data analyzed here contain 65 samples of tobacco, and we consider the elution times between 4.95 and 5.03 min over the mass range m/z 50–350.

In Figures 3–6, we consider the estimated components using the direct fitting algorithm and the proposed probabilistic PARAFAC2 with homo- and heteroscedastic noise, respectively. In Figure 3, we report the ELBO using the probabilistic models as well as the R2 and CCD using the direct fitting algorithm, and in Figures 4–6, we present the extracted profiles for each data set.

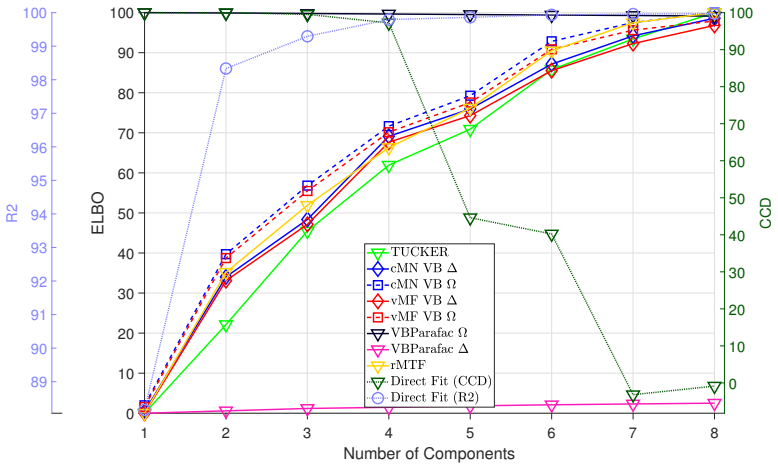
For the amino acid fluorescence data, we observe that both the R2 and CCD strongly suggest that a three-component model sufficiently describes the data, and the ELBO also finds no substantial improvements beyond three components (Figure 3a). In Figure 4, we investigate the extracted excitation loadings and observe that both the probabilistic and direct fitting PARAFAC2 models extract similar components when too few or the correct number of components are specified, i.e., $M \leq 3$. However, facing misspecification by having chosen too many components, the direct fitting algorithm extracts noisy profiles that incorrectly reflect the underlying three constituents. In contrast, the probabilistic PARAFAC2 models more robustly recover the three constituents when overspecifying the number of components—in particular, when assuming homoscedastic noise.

For the GC-MS-WINE data, the R2 and CCD point to a four- or five-component model, whereas the ELBO points to adding additional components (cf. Figure 3b). Inspecting the extracted components in Figure 5, we again observe close agreement between the extracted components using the probabilistic and direct fitting PARAFAC2 approaches when specifying a low number of components ($M \leq 5$). Furthermore, the estimated elution profiles facing model order misspecification appear less influenced by noise than the elution profiles extracted using the direct fitting algorithm, emphasizing the improved robustness by the Bayesian approach.

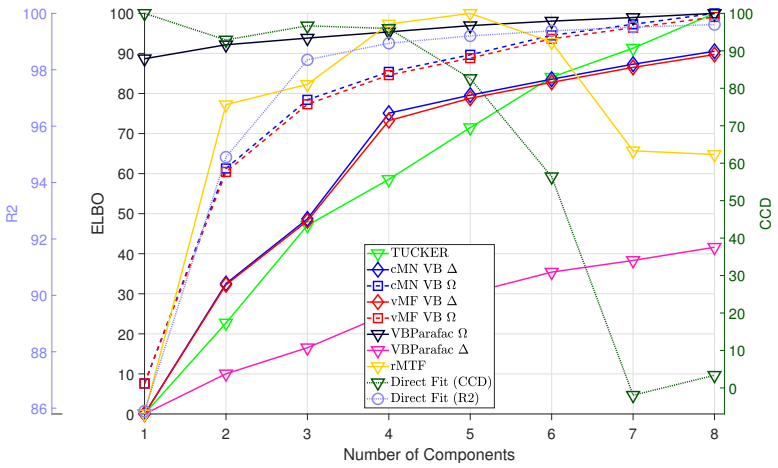
For the GC-MS-TOBAC data given in Figure 3c, we observe support for a three-component model according to R2 and CCD, whereas it is harder to decide a suitable model order based on the ELBO. The change in the ELBO from two to three components for the homoscedastic noise models suggests that local maxima have been identified. Inspecting the extracted components in Figure 6, it is also evident that local maxima have been reached for most of the probabilistic PARAFAC2 models with $M < 4$. For $M > 3$, most of the probabilistic models successfully recover the three components without using the extra components, where the direct fitting algorithm splits the three components into multiple components.



(**a**)



(**b**)



(**c**)

Figure 3. Mean of model selection criteria R2 and CCD reported on the conventional PARAFAC2, and the ELBO for the TUCKER, rMTF, probabilistic PARAFAC, and probabilistic PARAFAC2 models with 1 to 8 components on the AAF (**a**), GC-MS-WINE (**b**), and GC-MS-TOBAC (**c**) data sets. To make the results comparable, all ELBO values for each criterion and model (but across noise model types) have been normalized to be in the range of 0 to 100. In the legend, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

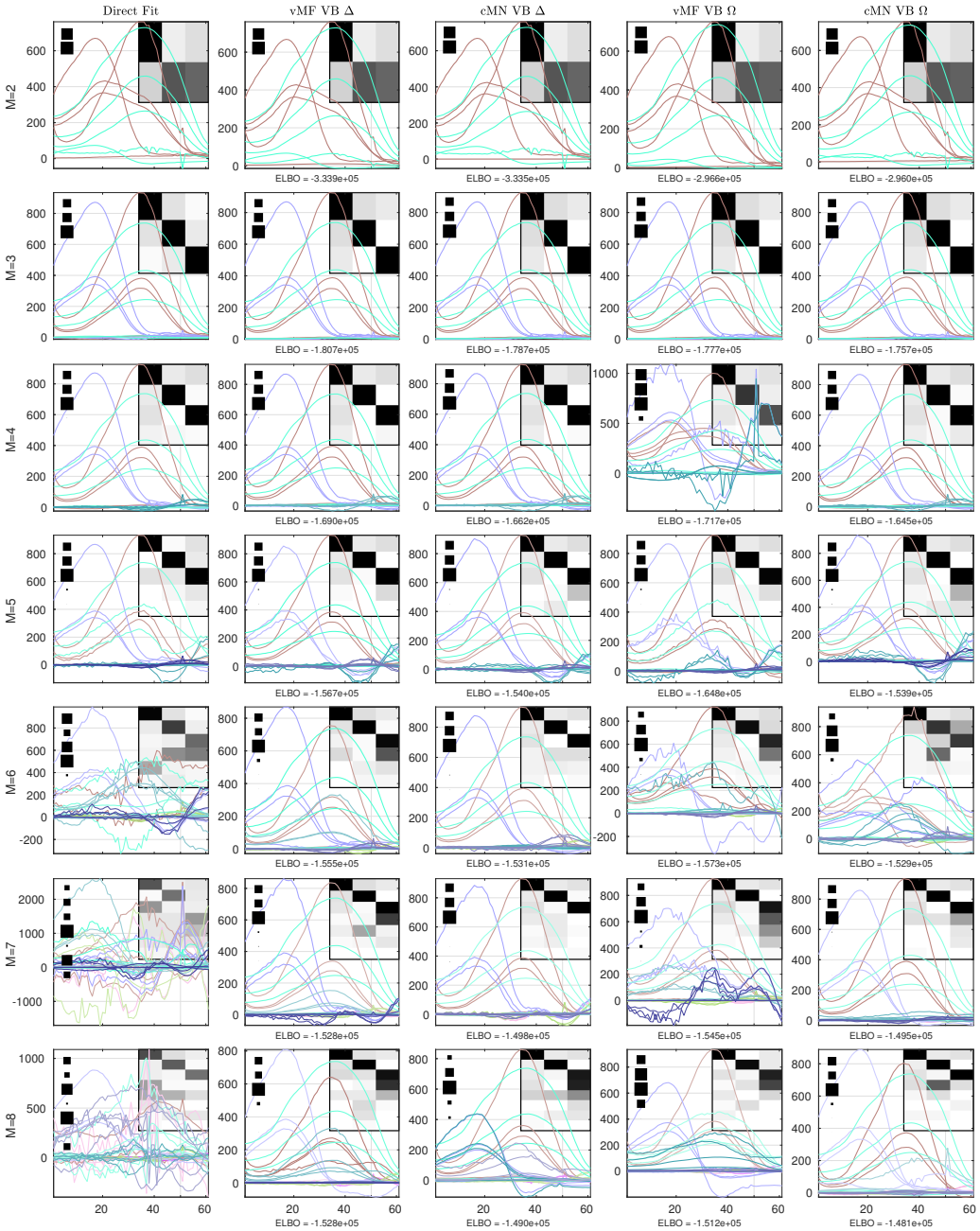


Figure 4. The excitation loadings of the AAF data given by the conventional PARAFAC2 and probabilistic PARAFAC2 models. From top to bottom, the loadings consist of 2 to 8 components. For each model, the background heatmap visualizes the correlation between the data reconstruction for each identified component and the componentwise data reconstruction of the conventional PARAFAC2 model with 3 components (ground-truth). Furthermore, to the left, a Hinton diagram indicates the relative squared Frobenius norm of the componentwise data reconstructions to the sum of them all. In the headers, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

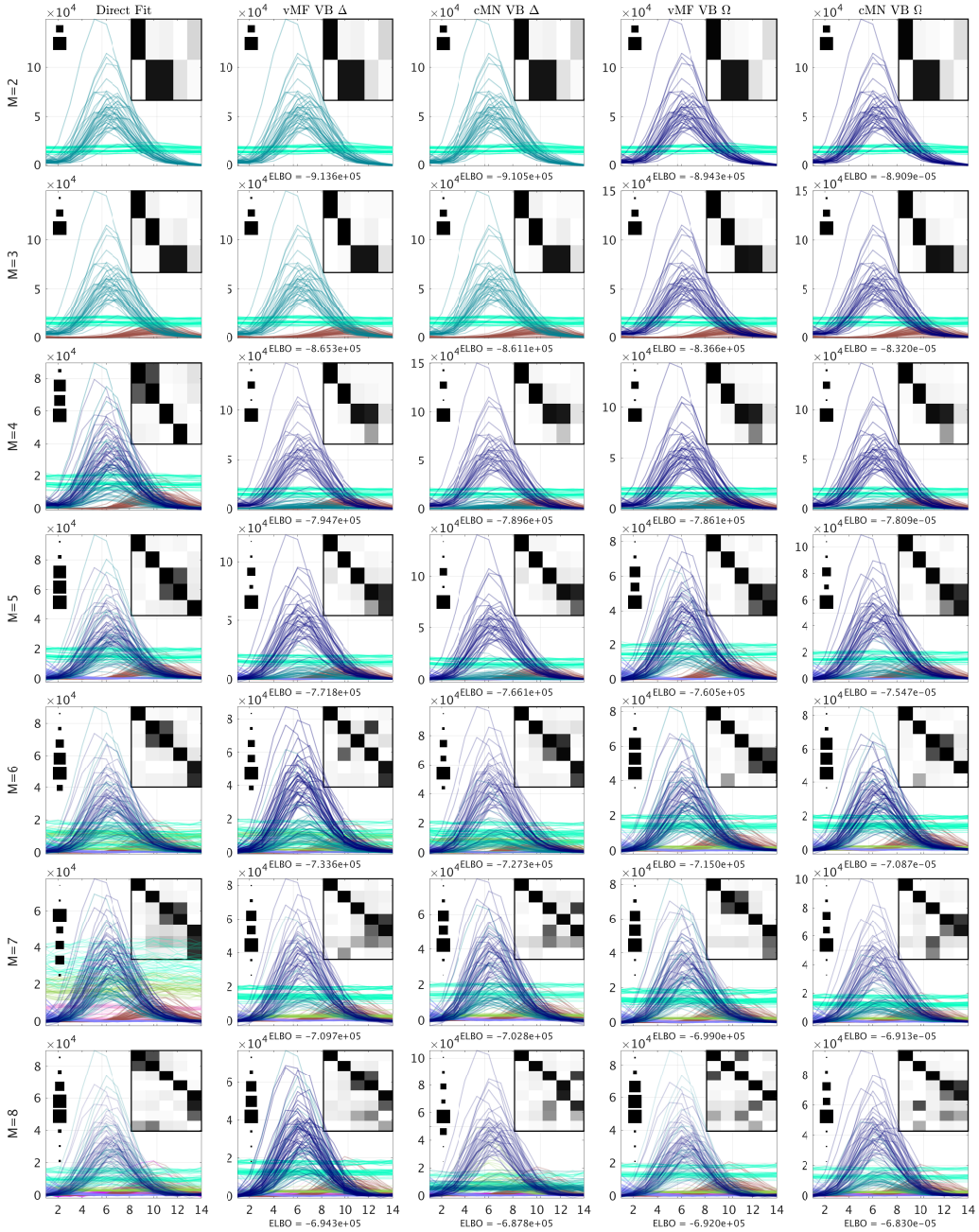


Figure 5. The elution profiles of the GC-MS-WINE data given by the conventional PARAFAC2 and probabilistic PARAFAC2 models. From top to bottom, the profiles consist of 2 to 8 components. For each model, the background heatmap visualizes the correlation between the data reconstruction for each identified component and the componentwise data reconstruction of the conventional PARAFAC2 model with 5 components (expert conclusion). Furthermore, to the left, a Hinton diagram indicates the relative squared Frobenius norm of the componentwise data reconstructions to the sum of them all. In the headers, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

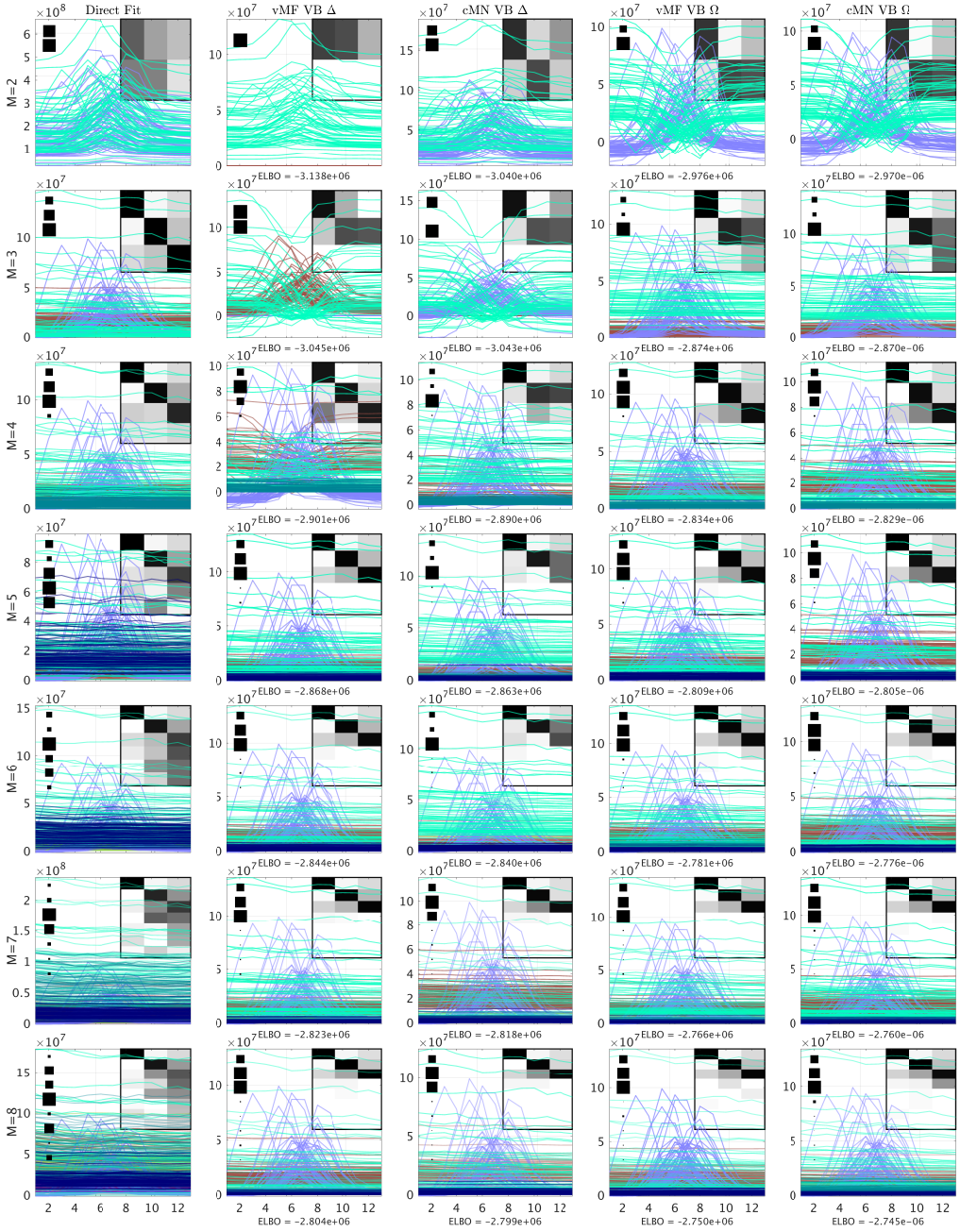


Figure 6. The elution profiles of the GC-MS-TOBAC data given by the conventional PARAFAC2 and probabilistic PARAFAC2 models. From top to bottom, the profiles consist of 2 to 8 components. For each model, the background heatmap visualizes the correlation between the data reconstruction for each identified component and the componentwise data reconstruction of the conventional PARAFAC2 model with 3 components (expert conclusion). Furthermore, to the left, a Hinton diagram indicates the relative squared Frobenius norm of the componentwise data reconstructions to the sum of them all. In the headers, Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

Of the three considered data sets, the ELBO itself does not strongly indicate an optimal number of components; however, most of the probabilistic models still manage to recover the underlying structure given by the ground-truth or expert conclusion in spite of being overspecified. This is in sharp contrast to MAP estimation, where overspecification typically leads to degenerate solutions. We attribute this to the regularization invoked by accounting for uncertainty and the automatic relevance determination promoting the pruning of

excess components. The relative importance of each component can be observed from the Hinton diagrams in Figures 4–6. Each square in the Hinton diagrams indicates the relative contribution of each component to the full data reconstruction, computed as the squared Frobenius norm of the componentwise data reconstruction divided by the sum of the squared Frobenius norms of all the componentwise data reconstructions.

4. Conclusions

We developed a fully probabilistic PARAFAC2 model and demonstrated how orthogonality can be imposed in the context of variational inference in two different ways: Firstly, using the von Mises–Fisher matrix distribution, assuming $\mathbb{E}[\boldsymbol{Y}^\top \boldsymbol{Y}] = \boldsymbol{I}$, as proposed in the context of variational PCA in [59]. Using this distribution forces all the realizations of the given matrix parameter to be orthogonal. Secondly, using the constrained matrix normal distribution, assuming $\mathbb{E}[\boldsymbol{Y}^\top]\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{I}$, in which the mean is constrained to the Stiefel manifold. This effectively results in a more flexible model as only the expectation of the realizations of the matrix are orthogonal and not the realizations themselves. For the latter approach, we presently derived a simple closed-form solution based on optimizing the lower bound.

Both probabilistic PARAFAC2 approaches were able to successfully recover the underlying signal in synthetic data when considering homoscedastic or heteroscedastic added noise. However, we found that the specification of orthogonality based on vMF was more robust to noise than the specification based on $c\mathcal{MN}$. In particular, we observed substantial noise robustness in the probabilistic PARAFAC2 models when compared to the conventional direct fitting approach, both when the correct model order was specified and when overestimating the number of components.

On the AAF data, the probabilistic PARAFAC2 framework was able to correctly identify the underlying constituents and demonstrated improved robustness to model misspecification when compared to the conventional direct fitting algorithm. The ELBOs of the probabilistic models suggest a model order of three components similar to the CCD and R2 heuristics computed from the direct fitting estimations. For the two gas chromatography–mass spectrometry data sets, GC-MS-WINE and GC-MS-TOBAC, we also observed agreement between the probabilistic and direct fitting PARAFAC2 models but with more mixed results. The model order is not so clearly evident from the ELBO on these data sets. However, we see that the automatic relevance determination suppresses unnecessary components fairly well on both data sets, ensuring robustness to overspecification of the model, which otherwise leads to degenerate solutions when the direct fitting approach is used. A few results from the probabilistic PARAFAC2 did not match the results of the direct fitting approach. This can most likely be explained by encountering local maxima, since variational methods are known to suffer from issues of underestimating uncertainty and thereby becoming overly confident on estimated parameters.

We attribute the performance improvements of probabilistic PARAFAC2 over conventional PARAFAC2 to the casting of PARAFAC2 as a Bayesian model, which approximates the posterior distribution of the parameters—rather than a point estimate as conventional PARAFAC2. Additionally, Bayesian inference, in general, enjoys more robustness to noise and overspecification of the model [48]. The proposed probabilistic PARAFAC2 models form an important step in the direction of applying probabilistic approaches to more advanced tensor decomposition approaches and a new direction for handling orthogonality constraints in probabilistic modeling—in general, using the proposed constrained matrix normal distribution framework, which has a simple variational update. In particular, we anticipate that the orthogonality constraints within a probabilistic setting may also be useful for the Tucker decomposition, in which orthogonality is typically imposed [5]; the block-term decompositions [70], in which orthogonality may be beneficial to impose within each block as previously considered using the vMF [48]; or to improve identifiability within the CP decomposition by imposing orthogonality as implemented in the n-way toolbox (http://www.models.life.ku.dk/nwaytoolbox, accessed on 28 February

2017). PARAFAC2 is actively being advanced and employed for new applications, e.g., recently, the higher-order block term decomposition has been embedded with a PARAFAC2 structure [71].

Author Contributions: Conceptualization, P.J.H.J., S.F.N., K.H.M., and M.M.; methodology, P.J.H.J., S.F.N., M.N.S., K.H.M., and M.M.; software, P.J.H.J., S.F.N., and M.M.; validation, P.J.H.J., S.F.N., J.L.H., M.N.S., K.H.M., and M.M.; formal analysis, P.J.H.J., S.F.N., and M.M.; investigation, P.J.H.J., S.F.N., J.L.H., M.N.S., K.H.M., and M.M.; data curation, P.J.H.J.; writing—original draft preparation, P.J.H.J., S.F.N., J.L.H., M.N.S., K.H., and M.M.; writing—review and editing, P.J.H.J., S.F.N., J.L.H., M.N.S., K.H., and M.M.; visualization, P.J.H.J., S.F.N., K.H., and M.M.; supervision, M.N.S., K.H., and M.M.; project administration, P.J.H.J. and J.L.H. All authors have read and agreed to the published version of the manuscript.

Funding: Philip J. H. Jørgensen was supported by the Innovation Fund Denmark through the Danish Center for Big Data Analytics and Innovation (DABAI) (Innovation Fund Denmark project nr. 10599 and. 10577). Morten Mørup (M.M.) was supported by the Novo Nordisk Foundation grant no. NNF23OC0083524. Furthermore M.M. and Jesper Løve Hinrich was supported by the Independent Research Fund Denmark (grant ID 10.46540/2035-00294B to M.M.).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

PCA	principal component analysis
SNR	signal-to-noise ratio
SVD	singular value decomposition
ARD	automatic relevance determination
ELBO	evidence lower bound
CCD	core consistency diagnostic
KL divergence	Kullback–Leibler divergence
vMF	Von Mises–Fisher
CP	CandeComp/PARAFAC

Appendix A. Software

A MATLAB implementation of the probabilistic PARAFAC2 model was used to run all experiments and generate the results in the paper. The source code is available on GitHub (https://github.com/philipjhj/VBParafac2, accessed on 28 February 2017), including a guide on setup and usage.

Appendix B. Deriving the Variational Inference

In the following, we derive the most important expressions used to identify the update rules of the model parameters. Below is a overview of the used notation.

Appendix B.1. The Evidence Lower Bound (ELBO)

An expansion of the ELBO is shown here:

$$\begin{aligned}
\mathrm{ELBO}(\mathrm{q}(\boldsymbol{\theta})) =& \mathbb{E}[\log p(\mathcal{X}, \boldsymbol{\theta})] - \mathbb{E}[\log \mathrm{q}(\boldsymbol{\theta})] \\
=& \mathbb{E}[\log \mathrm{p}(\mathcal{X}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau}, \boldsymbol{\alpha})] - \mathbb{E}[\log \mathrm{q}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau}, \boldsymbol{\alpha})] \\
=& \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau}) \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{\alpha}) \mathrm{p}(\boldsymbol{F}) \mathrm{p}(\mathcal{P}) \mathrm{p}(\boldsymbol{\tau})] \\
& - \mathbb{E}[\log \mathrm{q}(\boldsymbol{A}) \mathrm{q}(\boldsymbol{C}) \mathrm{q}(\boldsymbol{F}) \mathrm{q}(\boldsymbol{P}_k) \mathrm{q}(\boldsymbol{\tau})] \\
=& \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{A})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{\alpha})] \\
& + \mathbb{E}[\log \mathrm{p}(\boldsymbol{F})] + \mathbb{E}[\log \mathrm{p}(\mathcal{P})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{\tau})] \\
& - \mathbb{E}[\log \mathrm{q}(\boldsymbol{A})] - \mathbb{E}[\log \mathrm{q}(\boldsymbol{C})] - \mathbb{E}[\log \mathrm{q}(\boldsymbol{F})] - \mathbb{E}[\log \mathrm{q}(\mathcal{P})] \\
& - \mathbb{E}[\log \mathrm{q}(\boldsymbol{\tau})] \\
=& \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{A})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{\alpha})] \\
& + \mathbb{E}[\log \mathrm{p}(\boldsymbol{F})] + \mathbb{E}[\log \mathrm{p}(\mathcal{P})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{\tau})] \\
& + \mathrm{h}(\mathrm{q}(\boldsymbol{A})) + \mathrm{h}(\mathrm{q}(\boldsymbol{C})) + \mathrm{h}(\mathrm{q}(\boldsymbol{F})) + \mathrm{h}(\mathrm{q}(\mathcal{P})) \\
& + \mathrm{h}(\mathrm{q}(\boldsymbol{\tau}))
\end{aligned}$$

How to derive each of these terms is shown in the following.

Appendix B.2. Standard Moment Matching

As the formulation of the probabilistic PARAFAC2 model consists of the multivariate normal and gamma distribution, we expand the logarithm of their general expressions below. This will serve as a reference for identifying the parameters of the variational distribution when reading the derivations of the update rules.

Appendix B.2.1. Multivariate Normal Distribution

Deriving the log of the probability density function of the multivariate normal distribution amounts to

$$\begin{aligned}
f(x_1, \ldots, x_k) &= \mathcal{N}([x_1, \ldots, x_k]; \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) = \mathcal{N}(\boldsymbol{X}; \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) \\
f(x_1, \ldots, x_k) &= (2\pi)^{-\frac{k}{2}} (|\boldsymbol{\Sigma}_X|)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_X)^\top \boldsymbol{\Sigma}_X^{-1} (\mathbf{X} - \boldsymbol{\mu}_X)) \\
\Rightarrow \ln f(x_1, \ldots, x_k) &= \ln \left[(2\pi)^{-\frac{k}{2}} (|\boldsymbol{\Sigma}_X|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_X)^\top \boldsymbol{\Sigma}_X^{-1} (\mathbf{X} - \boldsymbol{\mu}_X) \right) \right] \\
&= -\frac{k}{2} \ln(2\pi) - \frac{1}{2} \ln(|\boldsymbol{\Sigma}_X|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_X)^\top \boldsymbol{\Sigma}_X^{-1} (\mathbf{X} - \boldsymbol{\mu}_X) \\
&= -\frac{k}{2} \ln(2\pi) - \frac{1}{2} \ln(|\boldsymbol{\Sigma}_X|) - \frac{1}{2} \mathbf{X}^\top \boldsymbol{\Sigma}_X^{-1} \mathbf{X} - \frac{1}{2} \boldsymbol{\mu}_X{}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X + \boldsymbol{\mu}_X{}^\top \boldsymbol{\Sigma}_X^{-1} \mathbf{X} \\
&= -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_X|) - \frac{1}{2} \mathbf{X}^\top \boldsymbol{\Sigma}_X^{-1} \mathbf{X} - \frac{1}{2} \boldsymbol{\mu}_X{}^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X + \boldsymbol{\mu}_X{}^\top \boldsymbol{\Sigma}_X^{-1} \mathbf{X} + c
\end{aligned}$$

where c is the constant terms with respect to $\boldsymbol{X}_k$ and its parameters.

Appendix B.2.2. Gamma Distribution

Deriving the log density function of the gamma distribution amounts to

$$\begin{aligned} f(x;a,b) =& \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-xb^{-1}) \\ \Rightarrow \ln f(x;a,b) =& \ln\left[\frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-xb^{-1})\right] \\ =& \ln \frac{1}{\Gamma(a)b^a} + (a-1)\ln x - xb^{-1} \\ =& (a-1)\ln x - xb^{-1} + c \end{aligned}$$

where c is the constant terms with respect to x.

Appendix B.3. Non-Trivial Moment Matching

To identify the parameters for $\boldsymbol{C}$ and $\boldsymbol{F}$, non-trivial steps had to be performed.

Appendix B.3.1. The F Matrix

The variational factor for $\boldsymbol{F}$ is defined as

$$\begin{aligned} \mathrm{q}(\boldsymbol{F}) \propto& \exp \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\mathcal{X}, \boldsymbol{\theta})] \\ \propto& \exp \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\mathcal{X}, \boldsymbol{F} \mid \boldsymbol{A}, \boldsymbol{C}, \mathcal{P}, \boldsymbol{\tau})] \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\mathcal{X}, \boldsymbol{F} \mid \boldsymbol{A}, \boldsymbol{C}, \mathcal{P}, \boldsymbol{\tau})] =& \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau})] + \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\boldsymbol{F})] \\ =& \sum_k \sum_i \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\boldsymbol{x}_{i\cdot k} \mid \boldsymbol{a}_{i\cdot}, \boldsymbol{D}_k, \boldsymbol{F}, \boldsymbol{P}_k, \tau_k)] + \sum_m \mathbb{E}_{-\boldsymbol{F}}[\log \mathrm{p}(\boldsymbol{f}_{m\cdot})] \\ =& \sum_k \sum_i \mathbb{E}_{-\boldsymbol{F}}[-\frac{1}{2}(\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top)\mathbf{I}_M\tau_k(\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top)^\top \\ & + \boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top{\boldsymbol{P}_k}^\top\mathbf{I}_M\tau_k{\boldsymbol{x}_{i\cdot k}}^\top] + \sum_m \mathbb{E}_{-\boldsymbol{F}}[-\frac{1}{2}\boldsymbol{f}_{m\cdot}\mathbf{I}_M{\boldsymbol{f}_{m\cdot}}^\top] + c \\ =& -\frac{1}{2}\sum_k \sum_i \mathbb{E}_{-\boldsymbol{F}}[\tau_k(\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top{\boldsymbol{P}_k}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k{\boldsymbol{a}_{i\cdot}}^\top)] - \frac{1}{2}\sum_m \boldsymbol{f}_{m\cdot}{\boldsymbol{f}_{m\cdot}}^\top \\ & + \sum_k \sum_i \mathbb{E}_{-\boldsymbol{F}}[\tau_k\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top{\boldsymbol{P}_k}^\top{\boldsymbol{x}_{i\cdot k}}^\top] + c \\ =& -\frac{1}{2}\sum_k \mathbb{E}[\tau_k]\sum_i \mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top{\boldsymbol{P}_k}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k{\boldsymbol{a}_{i\cdot}}^\top] - \frac{1}{2}\sum_m \boldsymbol{f}_{m\cdot}{\boldsymbol{f}_{m\cdot}}^\top \\ & + \sum_k \sum_i \mathbb{E}[\tau_k]\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top{\boldsymbol{P}_k}^\top{\boldsymbol{x}_{i\cdot k}}^\top] + c. \end{aligned}$$

Again, we reorder the parameters using the trace operator to identify the quadratic term. This time, the quadratic term separates into a quadratic and linear part revealing a linear intercomponent dependency.

$$\begin{aligned}
\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top] &= \mathbb{E}_{-\boldsymbol{F}}[\mathrm{Tr}(\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top)] \\
&= \mathbb{E}_{-\boldsymbol{F}}[\mathrm{Tr}(\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{P}_k)] \\
&= \mathrm{Tr}(\boldsymbol{F}\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\boldsymbol{F}^\top\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{P}_k{}^\top\boldsymbol{P}_k]) \\
&= \sum_{mm\prime}(\boldsymbol{F}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\boldsymbol{F}^\top)_{mm\prime}(\mathbb{E}[\boldsymbol{P}_k{}^\top\boldsymbol{P}_k])_{mm\prime} \\
&= \sum_{mm\prime}\boldsymbol{f}_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbf{f}^{\mathrm{T}}_{m\prime\cdot}\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}] \\
&= \sum_{m}\boldsymbol{f}_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot mk}]\mathbf{f}^{\mathrm{T}}_{m\cdot} \\
&\quad + 2\sum_{m}\sum_{m\prime\setminus m}\boldsymbol{f}_{m\cdot}\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}]\mathbf{f}^{\mathrm{T}}_{m\prime\cdot}.
\end{aligned}$$

Again, we have to reorder and include the linear terms as before.

$$\begin{aligned}
\sum_i\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k{}^\top\boldsymbol{x}_{i\cdot k}{}^\top] &= \sum_i\sum_m\mathbb{E}_{-\boldsymbol{F}}[\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{f}_{m\cdot}{}^\top(\boldsymbol{P}_k{}^\top)_m\boldsymbol{x}_{i\cdot k}{}^\top] \\
&= \sum_i\sum_m\mathbb{E}[\boldsymbol{a}_{i\cdot}]\mathbb{E}[\boldsymbol{D}_k]\boldsymbol{f}_{m\cdot}{}^\top\mathbb{E}[(\boldsymbol{P}_k{}^\top)_m]\boldsymbol{x}_{i\cdot k}{}^\top \\
&= \sum_i\sum_m\mathbb{E}[(\boldsymbol{P}_k{}^\top)_m]\boldsymbol{x}_{i\cdot k}{}^\top\mathbb{E}[\boldsymbol{a}_{i\cdot}]\mathbb{E}[\boldsymbol{D}_k]\boldsymbol{f}_{m\cdot}{}^\top \\
&= \sum_m\mathbb{E}[(\boldsymbol{P}_k{}^\top)_m](\sum_i\boldsymbol{x}_{i\cdot k}{}^\top\mathbb{E}[\boldsymbol{a}_{i\cdot}])\mathbb{E}[\boldsymbol{D}_k]\boldsymbol{f}_{m\cdot}{}^\top
\end{aligned}$$

Accounting for all terms and matching them to the ones in Appendix B.2, we arrive at the following update rules for $\boldsymbol{F}$.

$$\mathrm{q}(\boldsymbol{F}) = \prod_m\mathcal{N}(\boldsymbol{\mu}_{f_{m\cdot}}, \boldsymbol{\Sigma}_{f_{m\cdot}}), \tag{A1}$$

$$\boldsymbol{\mu}_{f_{m\cdot}} = \boldsymbol{\Sigma}_{f_{m\cdot}}(\sum_k\mathbb{E}[\tau_k](\mathbb{E}[(\boldsymbol{P}_k{}^\top)_m]\boldsymbol{X}_k{}^\top\mathbb{E}[\boldsymbol{A}]\mathbb{E}[\boldsymbol{D}_k] - \sum_i\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\sum_{m\prime\setminus m}\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}]\mathbf{f}^{\mathrm{T}}_{m\prime\cdot})), \tag{A2}$$

$$\boldsymbol{\Sigma}_{f_{m\cdot}} = (\sum_k\mathbb{E}[\tau_k]\sum_i\mathbb{E}[\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^\top\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k]\mathbb{E}[\mathbf{p}^{\mathrm{T}}_{\cdot mk}\mathbf{p}_{\cdot m\prime k}] + \mathbf{I}_M)^{-1}. \tag{A3}$$

Appendix B.3.2. Constrained Matrix Normal Distribution

The orthogonality constraint in the model can be handled with two formulations. This section concerns the approach where the mean parameters of the variational approximation for $\boldsymbol{P}_k$ are constrained to be orthogonal, and the following section describes the solution using the von Mises–Fisher distribution. Instead of using the free form variational updates, we optimized the ELBO with respect to the mean parameters $\boldsymbol{M}_{\boldsymbol{P}_k} = \mathbb{E}[\boldsymbol{P}_k]$ constrained to be orthogonal.

$$
\begin{aligned}
\boldsymbol{M}_{\boldsymbol{P}_k} &= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \mathrm{ELBO}(\boldsymbol{M}_{\boldsymbol{P}_k}) && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau})] \\
&\quad + \mathbb{E}[\log \mathrm{p}(\boldsymbol{A})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{C} \mid \boldsymbol{\alpha})] \\
&\quad + \mathbb{E}[\log \mathrm{p}(\boldsymbol{\alpha})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{F})] + \mathbb{E}[\log \mathrm{p}(\mathcal{P})] + \mathbb{E}[\log \mathrm{p}(\boldsymbol{\tau})] \\
&\quad + \mathrm{h}(\mathrm{q}(\boldsymbol{A})) + \mathrm{h}(\mathrm{q}(\boldsymbol{C})) + \mathrm{h}(\mathrm{q}(\boldsymbol{F})) + \mathrm{h}(\mathrm{q}(\mathcal{P})) \\
&\quad + \mathrm{h}(\mathrm{q}(\boldsymbol{\tau})) + \mathrm{h}(\mathrm{q}(\boldsymbol{\alpha})) && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \mathbb{E}[\log \mathrm{p}(\mathcal{X} \mid \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{F}, \mathcal{P}, \boldsymbol{\tau})] + c_1 && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ -\frac{1}{2}\sum_k\sum_i \mathbb{E}[\tau_k(\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^{\top}\boldsymbol{P}_k{}^{\top}\boldsymbol{P}_k\boldsymbol{F}\boldsymbol{D}_k\boldsymbol{a}_{i\cdot}{}^{\top})] \\
&\quad + \sum_k\sum_i \mathbb{E}[\tau_k\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^{\top}\boldsymbol{P}_k{}^{\top}\boldsymbol{x}_{i\cdot k}{}^{\top}] + c_2 && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \sum_k\sum_i \mathbb{E}[\tau_k\boldsymbol{a}_{i\cdot}\boldsymbol{D}_k\boldsymbol{F}^{\top}\boldsymbol{P}_k{}^{\top}\boldsymbol{x}_{i\cdot k}{}^{\top}] + c_3 && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \sum_k \mathbb{E}[\tau_k]\mathrm{Tr}(\mathbb{E}[\boldsymbol{A}]\mathbb{E}[\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{F}^{\top}]\mathbb{E}[\boldsymbol{P}_k{}^{\top}]\boldsymbol{X}_k{}^{\top}) + c_3 && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I} \\
&= \underset{\boldsymbol{M}_{\boldsymbol{P}_k}}{\arg\max}\ \sum_k \mathbb{E}[\tau_k]\mathrm{Tr}(\mathbb{E}[\boldsymbol{F}]\mathbb{E}[\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{A}^{\top}]\boldsymbol{X}_k\boldsymbol{M}_{\boldsymbol{P}_k}) + c_3 && \text{s.t. } \boldsymbol{M}_{\boldsymbol{P}_k}\boldsymbol{M}_{\boldsymbol{P}_k}{}^{\top} = \mathbf{I}.
\end{aligned}
$$

Only the linear term of the probability density function of the data $\mathcal{X}$ depends on $\boldsymbol{M}_{\boldsymbol{P}_k}$ since $\boldsymbol{M}_{\boldsymbol{P}_k}$ in the quadratic terms is the identity matrix. Except for a scalar, the optimization problem reduces to the same one as finding $\boldsymbol{P}_k$ in the alternating least squares algorithm, where one maximizes $\mathrm{Tr}(\mathbb{E}[\boldsymbol{F}]\mathbb{E}[\boldsymbol{D}_k]\mathbb{E}[\boldsymbol{A}^{\top}]\boldsymbol{X}_k\boldsymbol{M}_{\boldsymbol{P}_k})$ subject to the orthogonality constraint. The solution to this is found by simply applying an SVD, as stated in the main text (The alternating least squares method is described in [15], and the solution to the optimization problem was first described in [52]).

References

1. Carroll, J.D.; Chang, J.J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **1970**, *35*, 283–319. [CrossRef]
2. Harshman, R.A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Work. Pap. Phon.* **1970**, *16*, 84.
3. Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171. [CrossRef]
4. Appellof, C.J.; Davidson, E.R. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* **1981**, *53*, 2053–2056. [CrossRef]
5. Kolda, T.G.; Bader, B.W. Tensor decompositions and applications. *SIAM Rev.* **2009**, *51*, 455–500. [CrossRef]
6. Mørup, M. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 24–40. [CrossRef]
7. Sidiropoulos, N.D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E.E.; Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* **2017**, *65*, 3551–3582. [CrossRef]
8. Tucker, L.R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **1966**, *31*, 279–311. [CrossRef]
9. Hitchcock, F.L. The expression of a tensor or a polyadic as a sum of products. *Stud. Appl. Math.* **1927**, *6*, 164–189. [CrossRef]
10. Kruskal, J.B. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Its Appl.* **1977**, *18*, 95–138. [CrossRef]
11. Bro, R.; Andersson, C.A.; Kiers, H.A. PARAFAC2-Part II. Modeling chromatographic data with retention time shifts. *J. Chemom.* **1999**, *13*, 295–309. [CrossRef]
12. Johnsen, L.G.; Amigo, J.M.; Skov, T.; Bro, R. Automated resolution of overlapping peaks in chromatographic data. *J. Chemom.s* **2014**, *28*, 71–82. [CrossRef]
13. Harshman, R.A.; Lundy, M.E. Uniqueness proof for a family of models sharing features of Tucker's three-mode factor analysis and PARAFAC/CANDECOMP. *Psychometrika* **1996**, *61*, 133–154. [CrossRef]

14. ten Berge, J.M.; Kiers, H.A. Some uniqueness results for PARAFAC2. *Psychometrika* **1996**, *61*, 123–132. [CrossRef]
15. Kiers, H.A.; Ten Berge, J.M.; Bro, R. PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemom.* **1999**, *13*, 275–294. [CrossRef]
16. Wise, B.M.; Gallagher, N.B.; Martin, E.B. Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *J. Chemom.* **2001**, *15*, 285–298. [CrossRef]
17. Weis, M.; Jannek, D.; Roemer, F.; Guenther, T.; Haardt, M.; Husar, P. Multi-dimensional PARAFAC2 component analysis of multi-channel EEG data including temporal tracking. In Proceedings of the Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 5375–5378.
18. Madsen, K.H.; Churchill, N.W.; Mørup, M. Quantifying functional connectivity in multi-subject fMRI data using component models. *Hum. Brain Mapp.* **2016**, *38*, 882–899. [CrossRef]
19. Acar, E.; Roald, M.; Hossain, K.M.; Calhoun, V.D.; Adali, T. Tracing evolving networks using tensor factorizations vs. ica-based approaches. *Front. Neurosci.* **2022**, *16*, 861402. [CrossRef]
20. Chew, P.A.; Bader, B.W.; Kolda, T.G.; Abdelali, A. Cross-language information retrieval using PARAFAC2. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Jose, CA, USA, 12–15 August 2007; pp. 143–152.
21. Panagakis, Y.; Kotropoulos, C. Automatic music tagging via PARAFAC2. In Proceedings of the Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 481–484.
22. Pantraki, E.; Kotropoulos, C. Automatic image tagging and recommendation via PARAFAC2. In Proceedings of the Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
23. Perros, I.; Papalexakis, E.E.; Wang, F.; Vuduc, R.; Searles, E.; Thompson, M.; Sun, J. SPARTan: Scalable PARAFAC2 for Large & Sparse Data. *arXiv* **2017**, arXiv:1703.04219.
24. Gujral, E.; Theocharous, G.; Papalexakis, E.E. Spade: Streaming parafac2 de composition for large datasets. In Proceedings of the 2020 SIAM International Conference on Data Mining, Cincinnati, OH, USA, 7–9 May 2020; pp. 577–585.
25. Jang, J.G.; Kang, U. Dpar2: Fast and scalable parafac2 decomposition for irregular dense tensors. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 2454–2467.
26. Yu, H.; Bro, R. PARAFAC2 and local minima. *Chemom. Intell. Lab. Syst.* **2021**, *219*, 104446. [CrossRef]
27. Cheng, Y.; Haardt, M. Enhanced Direct Fitting Algorithms for PARAFAC2 with Algebraic Ingredients. *IEEE Signal Process. Lett.* **2019**, *26*, 533–537. [CrossRef]
28. Cohen, J.E.; Bro, R. Nonnegative PARAFAC2: A Flexible Coupling Approach. In Proceedings of the Latent Variable Analysis and Signal Separation, Guildford, UK, 2–5 July 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 89–98.
29. Van Benthem, M.H.; Keller, T.J.; Gillispie, G.D.; DeJong, S.A. Getting to the core of PARAFAC2, a nonnegative approach. *Chemom. Intell. Lab. Syst.* **2020**, *206*, 104127. [CrossRef]
30. Roald, M.; Schenker, C.; Calhoun, V.D.; Adali, T.; Bro, R.; Cohen, J.E.; Acar, E. An AO-ADMM approach to constraining PARAFAC2 on all modes. *SIAM J. Math. Data Sci.* **2022**, *4*, 1191–1222. [CrossRef]
31. Chu, W.; Ghahramani, Z. Probabilistic Models for Incomplete Multi-dimensional Arrays. In Proceedings of the AISTATS, Clearwater Beach, FL, USA, 16–18 April 2009; pp. 89–96.
32. Mørup, M.; Hansen, L.K. Automatic relevance determination for multi-way models. *J. Chemom.* **2009**, *23*, 352–363. [CrossRef]
33. Porteous, I.; Bart, E.; Welling, M. Multi-HDP: A Non Parametric Bayesian Model for Tensor Factorization. In Proceedings of the AAAI, Chicago, IL, USA, 13–17 July 2008; Volume 8, pp. 1487–1490.
34. Sheng, G.; Denoyer, L.; Gallinari, P.; Jun, G. Probabilistic latent tensor factorization model for link pattern prediction in multi-relational networks. *J. China Univ. Posts Telecommun.* **2012**, *19*, 172–181.
35. Bhattacharya, A.; Dunson, D.B. Sparse Bayesian infinite factor models. *Biometrika* **2011**, *98*, 291–306. [CrossRef]
36. Shan, H.; Banerjee, A.; Natarajan, R. *Probabilistic Tensor Factorization for Tensor Completion*; University Digital Conservancy: Saint Paul, MN, USA, 2011.
37. Xu, Z.; Yan, F.; Qi, A. Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), Edinburgh, UK, 26 June–1 July 2012; pp. 1023–1030.
38. Zhao, Q.; Zhang, L.; Cichocki, A. Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1751–1763. [CrossRef] [PubMed]
39. Ermis, B.; Yılmaz, Y.K.; Cemgil, A.T.; Acar, E. Variational Inference for Probabilistic Latent Tensor Factorization with KL Divergence. *arXiv* **2014**, arXiv:1409.8083.
40. Hore, V.; Viñuela, A.; Buil, A.; Knight, J.; McCarthy, M.I.; Small, K.; Marchini, J. Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **2016**, *48*, 1094–1100. [CrossRef] [PubMed]
41. Beliveau, V.; Papoutsakis, G.; Hinrich, J.L.; Mørup, M. Sparse Probabilistic Parallel Factor Analysis for the modeling of PET and task-fMRI data. In *Proceedings of the Bayesian and Graphiphal Models for Biomedical Imaging, MICCAI*; Springer: Cham, Switzerland, 2016.
42. Schmidt, M.N.; Mohamed, S. Probabilistic non-negative tensor factorization using Markov chain Monte Carlo. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, UK, 24–28 August 2009; pp. 1918–1922.
43. Xu, Z.; Yan, F.; Qi, Y. Bayesian nonparametric models for multiway data analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 475–487. [CrossRef]

44. Zhao, Q.; Zhou, G.; Zhang, L.; Cichocki, A.; Amari, S.I. Bayesian robust tensor factorization for incomplete multiway data. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 736–748. [CrossRef] [PubMed]
45. Hinrich, J.L.; Madsen, K.H.; Mørup, M. The probabilistic tensor decomposition toolbox. *Mach. Learn. Sci. Technol.* **2020**, *1*, 025011. [CrossRef]
46. Hayashi, K.; Takenouchi, T.; Shibata, T.; Kamiya, Y.; Kato, D.; Kunieda, K.; Yamada, K.; Ikeda, K. Exponential family tensor factorization: An online extension and applications. *Knowl. Inf. Syst.* **2012**, *33*, 57–88. [CrossRef]
47. Cheng, L.; Wu, Y.C.; Poor, H.V. Probabilistic Tensor Canonical Polyadic Decomposition with Orthogonal Factors. *IEEE Trans. Signal Process.* **2017**, *65*, 663–676. [CrossRef]
48. Hinrich, J.L.; Mørup, M. Probabilistic Block Term Decomposition for the Modelling of Higher-order Arrays. *Comput. Sci. Eng.* **2024**. [CrossRef]
49. Bishop, C.M. Variational principal components. In Proceedings of the 9th International Conference on Artificial Neural Networks ICANN 99, Edinburgh, UK, 7–10 September 1999; pp. 509–514. [CrossRef]
50. Jørgensen, P.; Nielsen, S.; Hinrich, J.; Schmidt, M.; Madsen, K.; Mørup, M. Analysis of Chromatographic Data using the Probabilistic PARAFAC2. In Proceedings of the Second Workshop on Machine Learning and the Physical Sciences, 33rd Conference on Neural Information Processing Systems, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019.
51. Harshman, R.A. PARAFAC2: Mathematical and technical notes. *UCLA Work. Pap. Phon.* **1972**, *22*, 30–44.
52. Green, B.F. The Orthogonal Approximation of An Oblique Staructre in Factor Analysis. *Psychometrika* **1952**, *17*, 429–440. [CrossRef]
53. Bro, R.; Kiers, H.A. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* **2003**, *17*, 274–286. [CrossRef]
54. Kamstrup-Nielsen, M.H.; Johnsen, L.G.; Bro, R. Core consistency diagnostic in PARAFAC2. *J. Chemom.* **2013**, *27*, 99–105. [CrossRef]
55. Attias, H. A Variational Baysian Framework for Graphical Models. In Proceedings of the NIPS 1999, Denver, CO, USA, 29 November–4 December 1999; Volume 12.
56. Bishop, C.M. Pattern Recognition and Machine Learning. *J. Electron. Imaging* **2006**, *16*, 049901. [CrossRef]
57. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
58. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1995.
59. Šmídl, V.; Quinn, A. On Bayesian principal component analysis. *Comput. Stat. Data Anal.* **2007**, *51*, 4101–4123. [CrossRef]
60. Bhattacharya, A.; Pati, D.; Pillai, N.S.; Dunson, D.B. Bayesian shrinkage. *arXiv* **2012**, arXiv:1212.6088.
61. Bro, R. Multi-Way Analysis in the Food Industry: Models, Algorithms, and Applications. Ph.D. Thesis, Københavns Universitet, Det Biovidenskabelige Fakultet for Fødevarer, Veterinærmedicin, Copenhagen, Denmark, 1998.
62. Gillis, N.; Glineur, F. Nonnegative factorization and the maximum edge biclique problem. *arXiv* **2008**, arXiv:0810.4225.
63. Nielsen, S.F.V.; Mørup, M. Non-negative tensor factorization with missing data for the modeling of gene expressions in the human brain. In Proceedings of the Machine Learning for Signal Processing (MLSP), Reims, France, 21–24 September 2014; pp. 1–6.
64. Khatri, C.; Mardia, K. The von Mises-Fisher matrix distribution in orientation statistics. *J. R. Stat. Soc. Ser. (Methodol.)* **1977**, *39*, 95–106. [CrossRef]
65. Khan, S.A.; Leppäaho, E.; Kaski, S. Bayesian multi-tensor factorization. *Mach. Learn.* **2016**, *105*, 233–253. [CrossRef]
66. Kiers, H.A. A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. *J. Chemom.* **1998**, *12*, 155–171. [CrossRef]
67. Skov, T.; Ballabio, D.; Bro, R. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Anal. Chim. Acta* **2008**, *615*, 18–29. [CrossRef]
68. Amigo, J.M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. Solving GC-MS problems with parafac2. *TrAC Trends Anal. Chem.* **2008**, *27*, 714–725. [CrossRef]
69. Tian, K.; Wu, L.; Min, S.; Bro, R. Geometric search: A new approach for fitting PARAFAC2 models on GC-MS data. *Talanta* **2018**, *185*, 378–386. [CrossRef]
70. De Lathauwer, L. Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness. *SIAM J. Matrix Anal. Appl.* **2008**, *30*, 1033–1066. [CrossRef]
71. Chatzichristos, C.; Kofidis, E.; Morante, M.; Theodoridis, S. Blind fMRI source unmixing via higher-order tensor decompositions. *J. Neurosci. Methods* **2019**, *315*, 17–47. [CrossRef] [PubMed]