Analysis of Chromatographic Data using the Probabilistic PARAFAC2

Philip J. H. Jørgensen
Department of
Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
phav@dtu.dk

Søren F. V. Nielsen
Research and Development
Sennheiser Communications
2750 Balllerup, Denmark
sfvnielsen@gmail.com

Jesper L. Hinrich
Department of
Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
jehi@dtu.dk

Mikkel N. Schmidt
Department of
Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
mnsc@dtu.dk

Kristoffer H. Madsen
Department of
Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
khma@dtu.dk

Morten Mørup
Department of
Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
mmor@dtu.dk

Abstract

PARAFAC2 is a widely applicable method often used for analyzing multi-way chromatographic data. We recently proposed a probabilistic framework for PARAFAC2[1]. The probabilistic formulations allow for a principled way of determining the number of latent components as well as modeling heteroscedastic noise. In this work we present a summary of the probabilistic PARAFAC2 models and their properties by revisiting the previous results of the analyzed data sets in a concise fashion.

1 Introduction

Multi-way analysis was originally developed within the field of psychometrics [2, 3], and since been used widely in other fields such as chemometrics [4]. Multi-way analysis appears in many fields of research including signal processing, neuroimaging, and information retrieval [5, 6]. The PARAFAC2 model, an extension of the CandeComp/PARAFAC (CP) model [2, 3, 7], was proposed by [8], has proven very useful for modeling chromatographic data handling variations occuring during experiments well[9, 10]. We recently proposed a probabilistic framework for the PARAFAC2 model for which we summarize the high-level details and more concisely present our results here.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

2 Probabilistic PARAFAC2

Using the model formulation as described by [11], the three-way PARAFAC2 model can be written as,

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

Based on this formulation of the PARAFAC2 model we developed two probabilistic PARAFAC2 formulations. The two formulations comes from the fact that in a probabilistic setting the orthogonality constraint $\boldsymbol{P}_k^\top\boldsymbol{P}_k = \boldsymbol{I}_M$ can be interpreted either as i) $\mathbb{E}[\boldsymbol{P}_k^\top\boldsymbol{P}_k] = \boldsymbol{I}_M$ or ii) $\mathbb{E}[\boldsymbol{P}_k]^\top\mathbb{E}[\boldsymbol{P}_k] = \boldsymbol{I}_M$. These formulations result in the following generative models i) and ii),

$$\boldsymbol{a}_{i\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M), \boldsymbol{f}_{m\cdot} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M), \boldsymbol{c}_{k\cdot} \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\alpha}^{-1})), \tau_k \sim \text{Gamma}(a_{\tau_k}, b_{\tau_k})$$
$$\text{i) } \boldsymbol{P}_k \sim \text{vMF}(\mathbf{0}), \text{ ii) } \boldsymbol{P}_k \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_J, \mathbf{I}_M)$$
$$\boldsymbol{X}_k \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{F}^\top\boldsymbol{P}_k^\top, \tau_k^{-1}\mathbf{I}_J)$$

Using the notation where $\boldsymbol{a}_{i\cdot}$ denotes the ith row of the matrix $\boldsymbol{A}$, and where $\boldsymbol{\alpha}$ is a vector where each element defines the length scale of a corresponding component.

Variational Inference Choosing the mean-field approximation for these model formulations lead to the factorized variational distribution given as $\text{q}(\boldsymbol{\theta}) = q(\boldsymbol{A})q(\boldsymbol{C})\prod_m q(\boldsymbol{f}_{m\cdot})\prod_k q(\boldsymbol{P}_k)q(\tau_k)$. The update rules of the parameters for this distribution follow the standard iterative scheme and are described in detail in [1], as well as the corresponding evidence lower bound (ELBO). Note that careful attention had to paid to the updates of the $\boldsymbol{F}$ matrix due to intercomponent dependencies, as well as the updates for the constrained $\boldsymbol{P}_k$ matrix. In the following we outline the details of the updates of the two different variational distributions of $\boldsymbol{P}_k$.

Matrix Von Mises-Fisher Loadings The model formulation using i) constrains the expectation of the inner product of $\boldsymbol{P}_k$ to be orthogonal. This fully conforms to the conventional PARAFAC2 model ensuring that every realization of the loadings are orthogonal. The variational distribution of $\boldsymbol{P}_k$ has the density, $\text{vMF}(\boldsymbol{P}_k|\boldsymbol{B}_{\boldsymbol{P}_k}) = \kappa(J, \boldsymbol{B}_{\boldsymbol{P}_k}^\top\boldsymbol{B}_{\boldsymbol{P}_k})^{-1}\exp\big(\text{tr}[\boldsymbol{B}_{\boldsymbol{P}_k}^\top\boldsymbol{P}_k]\big)$ which has support only on the Stiefel manifold. Details on how this was computed this can be found in [1].

Constrained Matrix Normal Loadings The model formluation using ii) constrains the expectation of the loadings themselves to the orthogonal. This results in a more flexible model than i) as the realizations of the loadings are no longer constrained to be orthogonal. However, the interpretation of the orthogonal factor becomes identical to the direct fitting method and also the update rule is in closed form in a similar manner to the direct fitting solution as shown in [1].

Noise Modeling The probabilistc formulation allow for both modeling homoscedastic noise or heteroscedastic noise. Either τ_k can be updated collectively for all k or individually.

Model Selection In the probabilistic framework the scale vector $\boldsymbol{\alpha}$ is used for exploiting automatic relevance determination [12] by modeling the length scale of each component. Since the abililty to prune excess components is more of interest than uncertainty estimates on the length scales we proposed in [1] to use maximum a posteriori estimates instead of a variational estimate.

3 Results

In [1] the proposed models were evaluated on synthetic data and 3 real data sets; an amino acid fluorescence (AAF) data set and two gas chromatography mass spectrometry (GC-MS) data sets. In the following we revisit the results of the synthetic data and one of the GC-MS data sets. We refer to [1] for descriptions of the analyzed data and full details on these experiments as well as the results left out here.

Model Comparison Conventionally, different PARAFAC2 models have been compared by the ratio of explained variance and the core consistency diagnostic, respectively denoted R2 and CCD in [1] and the reminder of this work. These have been used to compare their ability to determine the correct number of components in relation to the ELBO used for the probabilistic models.

Synthetic Data The models were fitted to the synthetic data sets in order to investigate the ability to recover an underlying signal in different noise settings and noise levels, as seen in Figure 2. Also, a comparison of the different statistics for determining model order on these synthetic data sets can be seen in Figure 1.

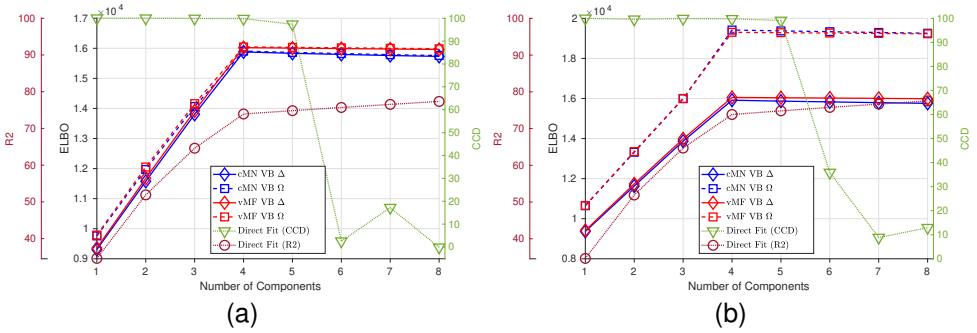

Figure 1: Model selection criteria given by the conventional PARAFAC2 and probabilistic PARAFAC2 models with 1 to 8 components on the synthetic data sets. In the legend Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

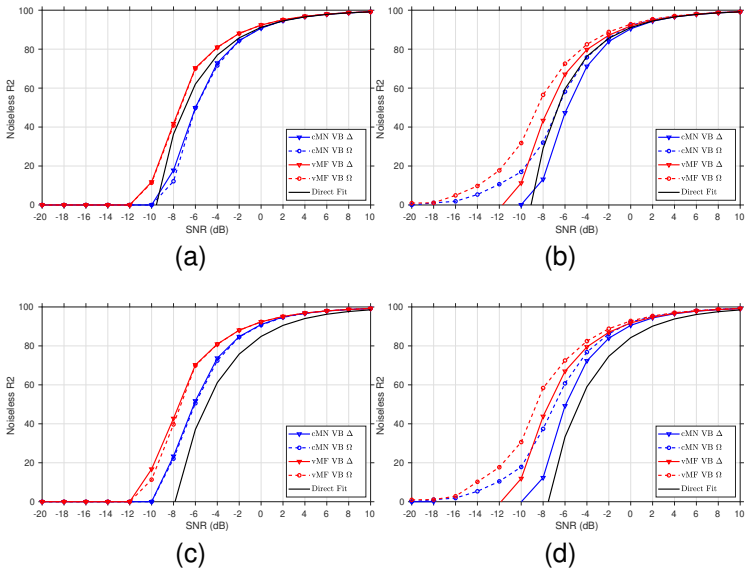

Figure 2: Noiseless R2 measured on different PARAFAC2 models fitted on synthetic data with varying levels of homoscedastic ((a),(c)) and heteroscedastic ((b),(d)) added noise. (a) and (b) show the result for models fitted with the true number of components (4 by design), and (c) and (d) for models with an overspecified number of components (6 by design) . In the legend Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

Real Data We include the results from [1] on the GC-MS data originating from tobacco (GC-MS-TOBAC). For the different models we see the model selection performance based on the R2, CCD and ELBO in Figure 3 as well as the resulting elution profiles in Figure 4.

4 Discussion

The probabilistc PARAFAC2 model recently developed and analyzed in [1] shows promising results for delivering important properties such as the principled approach of performing model selection through automatic relevance determination and handling varying noise settings and increased noise

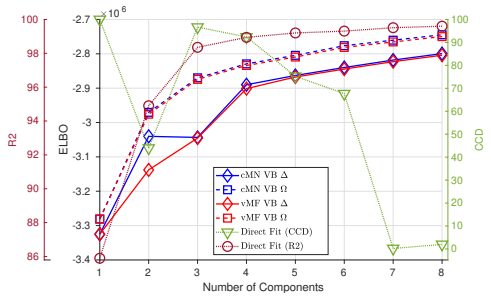

Figure 3: Model selection criteria given by the conventional PARAFAC2 and probabilistic PARAFAC2 models with 1 to 8 components on GC-MS-TOBAC data set. In the legend Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

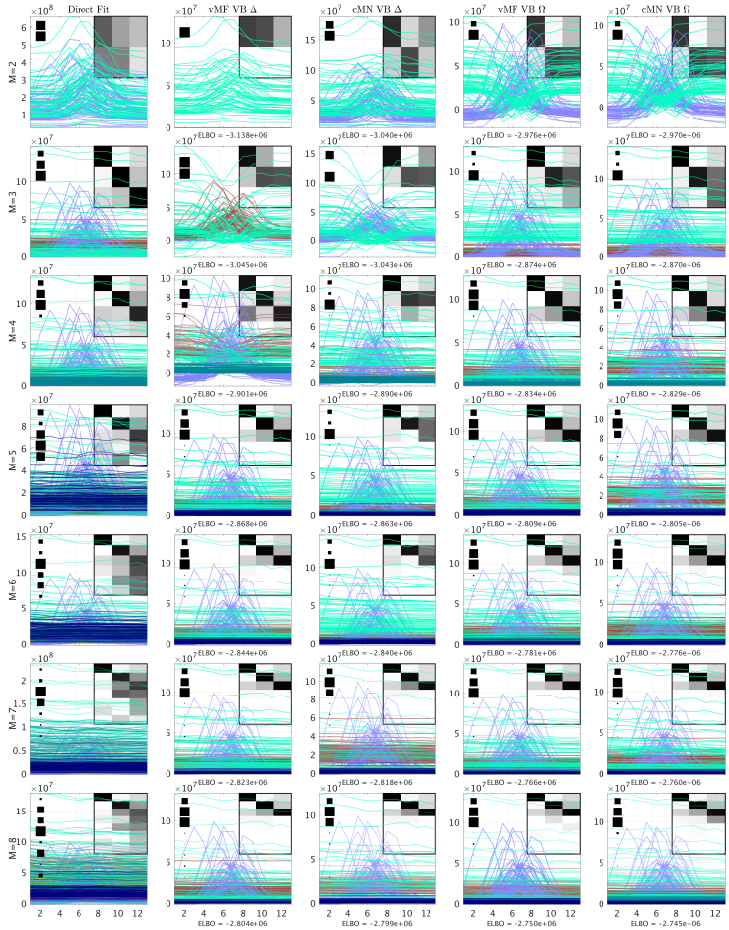

Figure 4: The resulting elution profiles of the GC-MS-TOBAC data given by the different PARAFAC2 models. From top to bottom the models is specified using model 2 to 8 components. The background heatmap visualizes the correlation between the data reconstruction for each identified component and the componentwise data reconstruction of the conventional PARAFAC2 model with 3 components (expert conclusion). Hinton diagrams indicate the relative squared Frobenius norm of the componentwise data reconstructions to the sum of them all to the left of each plot. In the headers Δ indicates a homoscedastic noise model and Ω indicates a heteroscedastic noise model.

levels, although known limitations of variational inference such as encountering local maxima are still present.

References

[1] Philip J. H. Jørgensen, Søren F. V. Nielsen, Jesper L. Hinrich, Mikkel N. Schmidt, Kristoffer H. Madsen, and Morten Mørup. Probabilistic parafac2, 2018, 1806.08195.

[2] J. Douglas Carroll and Jih Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3): 283–319, 1970. ISSN 00333123.

[3] Richard a Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16(10):1– 84, 1970.

[4] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[5] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51 (3):455–500, 2009.

[6] Morten Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1): 24–40, 2011.

[7] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Studies in Applied Mathematics*, 6(1-4):164–189, 1927.

[8] Richard A. Harshman. PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics*, 22(10):30–44, 1972. URL `http://www.bibsonomy.org/bibtex/2a964ff885ba59d4c7be518a3914f737a/threemode`.

[9] José Manuel Amigo, Thomas Skov, Rasmus Bro, Jordi Coello, and Santiago Maspoch. Solving GC-MS problems with PARAFAC2. *TrAC - Trends in Analytical Chemistry*, 27(8):714–725, 2008.

[10] Lea G Johnsen, José Manuel Amigo, Thomas Skov, and Rasmus Bro. Automated resolution of overlapping peaks in chromatographic data. *J. Chemom.*, 28(2):71–82, 2014.

[11] Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. Parafac2-part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics*, 13(3-4):275–294, 1999.

[12] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. 2006. ISBN 9780387310732. URL `http://www.library.wisc.edu/selectedtocs/bg0137.pdf`.