Probabilistic Tensor Train Decomposition

Jesper L. Hinrich, and Morten Mørup
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kongens Lyngby, Denmark

***Abstract*—The tensor train decomposition (TTD) has become an attractive decomposition approach due to its ease of inference by use of the singular value decomposition and flexible yet compact representations enabling efficient computations and reduced memory usage using the TTD representation for further analyses. Unfortunately, the level of complexity to use and the order in which modes should be decomposed using the TTD is unclear. We advance TTD to a fully probabilistic TTD (PTTD) using variational Bayesian inference to account for parameter uncertainty and noise. In particular, we exploit that the PTTD enables model comparisons by use of the evidence lower bound (ELBO) of the variational approximation. On synthetic data with ground truth structure and a real 3-way fluorescence spectroscopy dataset, we demonstrate how the ELBO admits quantification of model specification not only in terms of numbers of components for each factor in the TTD, but also a suitable order of the modes in which the TTD should be employed. The proposed PTTD provides a principled framework for the characterization of model uncertainty, complexity, and model- and mode-order when compressing tensor data using the TTD.**

***Index Terms*—Bayesian inference, tensor train decomposition, matrix product state, multi-modal data**

I. Introduction

Tensor decomposition approaches have become important tools for the modeling of multi-way array data in which prominent tensor decomposition approaches include the canonical polyadic decomposition (CPD), PARAFAC2, and the Tucker decomposition, see also [1]–[3] for reviews. Recently, the tensor train decomposition (TTD) has been proposed as a flexible alternative decomposition approach. The TTD has several attractive properties including compact yet flexible multi-way data representation, efficient inference through the use of the singular-value decomposition, and an attractive tensor representation format amenable to further efficient computational modeling due to the computational efficiency in which the model representations can contract modes and reduce memory storage, for details see [4].

Lately, tensor decomposition has been advanced to probabilistic modeling using variational inference. Benefits of probabilistic modeling includes robustness to model miss-specification and tools for complexity quantification through the evidence lower bound (ELBO), see also [5]–[7], while inference can be cast within traditional alternating optimization widely used for tensor decomposition optimization [1], [2]. Exploiting the advantages of probabilistic modeling the CPD [5]–[9] , PARAFAC2 [10], and Tucker decomposition [9], [11] have been advanced to fully Bayesian inference frameworks.

We advance the tensor train decomposition to variational Bayesian inference exploiting that the orthogonality structure in the decomposition can be imposed using a matrix-von-Mises-Fisher decomposition as used previously in the context of orthogonal probabilistic PCA [12] and for implementing the consistency constraints of the Gram matrices in the PARAFAC2 model [10] based on the direct fitting procedure of [13]. Notably, the proposed probabilistic tensor train decomposition (PTTD) admits model evaluation using the evidence lower bound (ELBO). We demonstrate how the ELBO can not only be used to quantify a suitable specification of rank of each factor in the tensor train, but also the order in which the modes should be decomposed. We highlight these aspects on synthetic data with ground truth TTD structure as well as on a real fluorescence spectroscopy dataset with known underlying CPD structure.

In summary, this paper investigates how the tensor train decomposition can be advanced to variational inference forming the PTTD and its inferential properties and merits are when compared to the conventional TTD.

II. Methods

Let $\boldsymbol{\mathcal{X}}$, $\mathbf{X}$, and $\mathbf{x}$, respectively denote a tensor, matrix, and a vector. Let $\boldsymbol{\mathcal{A}} \times_{[a,b]} \boldsymbol{\mathcal{B}}$ define the joint tensor contraction along mode a and b. For a M$^{\text{th}}$ order tensor $\boldsymbol{\mathcal{A}}^{I_1 \times \ldots \times I_M}$ and a set of tensors $\{\boldsymbol{\mathcal{U}}_i \in \mathbb{R}^{D_{i-1} \times I_i \times D_i}\}_{i=1,\ldots,M}$, where $D_0 = D_M = 1$, we define the sequential contraction of all $\boldsymbol{\mathcal{U}}_i$ (except the m^{th}) onto $\boldsymbol{\mathcal{A}}$ as $\texttt{contract}\left(\boldsymbol{\mathcal{A}}, \{\boldsymbol{\mathcal{U}}_i\}_{i \neq m}\right)$ resulting in a tensor $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{D_{m-1} \times I_m \times D_m}$. Using this notation the TTD approximation [4] can be written as:

$$\begin{aligned}\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{M}} = &\boldsymbol{\mathcal{U}}^{(1)} \times_{[2,1]} \boldsymbol{\mathcal{U}}^{(2)} \times_{[3,1]} \boldsymbol{\mathcal{U}}^{(3)} \times_{[4,1]} \cdots \\ &\times_{[M-1,1]} \boldsymbol{\mathcal{U}}^{(M-1)} \times_{[M,1]} (\mathbf{S}\mathbf{V}^\top)\end{aligned}$$

where $\boldsymbol{\mathcal{U}}^{(i)}, i = 1, \ldots, M-1$ are the factors or train carts, and last factor $\boldsymbol{\mathcal{U}}^{(M)} \equiv \mathbf{S}\mathbf{V}^\top$ is defined so it can handle different scaling of each component, i.e. s_{dd}, $d = 1, \ldots, D_{M-1}$.

Now, let $v\mathcal{MF}$ denote the von-Mises Fisher matrix distribution which defines a distribution of orthogonal matrices on the Stiefel-manifold, $\mathcal{TN}_{[a,b]}$ denote the truncated normal distribution on the interval $[a; b]$, and $\mathcal{G}(\alpha, \beta)$ denote a Gamma distribution with rate α and shape β, and $\mathcal{N}_{I_1 \times I_2 \times \cdots \times I_M}$ be the array normal distribution extending the conventional matrix normal distribution to higher order arrays. The generative

978-9-0827-9703-9/19/$31.00

model for the proposed probabilistic tensor train decomposition (PTTD) is then:

$$\begin{aligned}\tau \sim&\mathcal{G}(\alpha_\tau,\beta_\tau), \qquad \lambda\sim\mathcal{G}(\alpha_\lambda,\beta_\lambda),\\ s_{dd}\sim&\mathcal{TN}_{[0,\infty]}\left(0,\lambda^{-1}\right), \quad d=1,\dots,D_{M-1}\\ \mathbf{V}\sim&v\mathcal{MF}\left(\mathbf{0}^{I_M\times D_{M-1}}\right),\\ \mathbf{U}^{(m)}_{(1,2)}\sim&v\mathcal{MF}\left(\mathbf{0}^{D_{m-1}I_m\times D_m}\right), m=1,\dots,M-1\\ \mathcal{X}|\boldsymbol{\theta}\sim&\mathcal{N}_{I_1\times I_2\times\cdots\times I_M}\left(\boldsymbol{\mathcal{M}},\mathbf{I}_{I_1}\tau^{-1},\mathbf{I}_{I_2},\dots,\mathbf{I}_{I_M}\right),\end{aligned}$$

where the each factor, $\boldsymbol{\mathcal{U}}^{(m)}$, is matricized along the first and second mode $\mathbf{U}^{(m)}_{(1,2)}$ and is orthogonal in the last mode, i.e. $\mathbf{U}^{(m)\top}_{(1,2)}\mathbf{U}^{(m)}_{(1,2)}=\mathbf{I}$. The model assumes element-wise white noise $\mathcal{N}(0,\tau^{-1})$ with τ indicating the noise precision. The singular value of the last D_{M-1} components are positive and share the same scale, as indicated by the precision λ. The hyperparameters $\alpha_\tau,\beta_\tau,\alpha_\lambda,\beta_\lambda$ are fixed to 10^{-6} resulting in a broad prior on τ and λ.

The exact posterior distribution of the parameters, $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}})$, is intractable and we use variational Bayesian (VB) inference to approximate the model parameters, $\boldsymbol{\theta}=\{\tau,\lambda,\mathbf{S},\mathbf{V},\boldsymbol{\mathcal{U}}^{(1)},\boldsymbol{\mathcal{U}}^{(2)},\dots,\boldsymbol{\mathcal{U}}^{(M-1)}\}$. In VB inference the joint posterior distributions is approximated by a factorized distribution, here we use a mean-field approximation, i.e.

$$\begin{aligned}Q(\tau,\lambda,\boldsymbol{s},\mathbf{V},\mathbf{U}^{(1)},\dots,\mathbf{U}^{(M-1)})=&\\ Q(\tau)Q(\lambda)Q(\boldsymbol{s})Q(\mathbf{V})\prod_{m=1}^{M-1}Q(\mathbf{U}^{(m)}),&\end{aligned}\quad(1)$$

such that the marginal distribution is approximated by the evidence lower bound (ELBO) given by

$$\begin{aligned}&\log(P(\mathcal{X}))\geq ELBO(\mathcal{X})=\\ &E_Q[\log\left(\tfrac{P(\mathcal{X}|\mathcal{M},\tau)P(\tau)P(\lambda)P(\boldsymbol{s})P(\mathbf{V})\prod_{m=1}^{M-1}P(\mathbf{U}^{(m)})}{Q(\tau)Q(\lambda)Q(\boldsymbol{s})Q(\mathbf{V})\prod_{m=1}^{M-1}Q(\mathbf{U}^{(m)})}\right)].\end{aligned}$$

E_Q denotes expectation taken with respect to the Q distribution defined in eq. (1). The updates of each of the parameters θ_m can be found using the coordinate ascent variational inference (CAVI) procedure conditioning on all other parameters $\theta_{\setminus m}$, see [14] for details. The update of each of the parameters in the variational inference are given below. We note that as an alternative to VB inference, a Gibbs sampling procedure is easy to define and implement from these updates, by sampling the parameters $\theta_m|\theta_{\setminus m}$, instead of computing the relevant expectations of the random variables in VB.

Inferring $\boldsymbol{\mathcal{U}}^{(m)}$ for $m=1,\dots,M-1$. Contracting $\boldsymbol{\mathcal{X}}$ with all but the m$^{\text{th}}$ factor, i.e. $\{\boldsymbol{\mathcal{U}}^{(n)}\}_{n\neq m}$ and $\mathbf{SV}$, results in a tensor of size $D_{m-1}\times I_m\times D_m$ (as $\forall D_i, i\neq m-1,m$ are also contracted). For $m=1$ the contraction results in a $I_1\times D_1$ matrix and the estimation problem relates to the matrix decomposition $\mathbf{X}_{(1)}=\mathbf{U}^{(1)}\boldsymbol{\mathcal{W}}^\top_{(1)}$, where $\boldsymbol{\mathcal{W}}^{D_1\times I_2\times I_3\times\dots\times I_M}$ is the tensor train reconstruction using all but the first factor. Determining $\mathbf{U}^{(1)}$ is then done by moment matching a $v\mathcal{MF}$-distribution as derived in the context of Bayesian principal component analysis [12]. For $m=2,\dots,M-1$, $\boldsymbol{\mathcal{U}}^{(m)}$ is determined by contracting the data with $\{\boldsymbol{\mathcal{U}}^{(n)}\}_{n\neq m}$, unfolding the results, and moment matching the unfolded factor $\mathbf{U}^{(m)}_{(1,2)}$ resulting in,

$$\begin{aligned}\tilde{\boldsymbol{F}}^{(m)}=&\left\langle\tau\right\rangle\left(\texttt{contract}\left(\boldsymbol{\mathcal{X}},\{\left\langle\boldsymbol{\mathcal{U}}^{(n)}\right\rangle\}_{n\neq m}\right)\right)_{(1,2)},\\ Q(\mathbf{U}^{(m)}_{(1,2)})\sim&v\mathcal{MF}(\tilde{\boldsymbol{F}}^{(m)}), \qquad m=1,2,\dots,M-1.\end{aligned}\quad(2)$$

Where $\boldsymbol{\mathcal{U}}^{(m)}$ is obtained by reshaping $\mathbf{U}^{(m)}_{(1,2)}$ and the expected value, $\langle\mathbf{U}^{(m)}_{(1,2)}\rangle$, is determined by using singular value decomposition and the hypergeometric function, as shown in [12].

Inferring V. This is similar to the updates of $\boldsymbol{\mathcal{U}}^{(m)}$, as we can interpret $\boldsymbol{\mathcal{U}}^{(M)}\equiv\mathbf{SV}^\top$ and the update becomes,

$$\begin{aligned}\tilde{\boldsymbol{F}}^{(M)^\top}=&\left\langle\tau\right\rangle\left\langle\mathbf{S}\right\rangle\texttt{contract}\left(\boldsymbol{\mathcal{X}},\{\left\langle\boldsymbol{\mathcal{U}}^{(n)}\right\rangle\}_{n\neq M}\right),\\ Q(\mathbf{V})\sim&v\mathcal{MF}(\tilde{\boldsymbol{F}}^{(M)}).\end{aligned}\quad(3)$$

Inferring S. First we define the matrix $\mathbf{W}^{D_{M-1}\times I_M}=\texttt{contract}\left(\boldsymbol{\mathcal{X}},\{\left\langle\boldsymbol{\mathcal{U}}^{(n)}\right\rangle\}_{n\neq M}\right)$, then using the univariate truncated normal distribution each s_d, $\forall_{d=1,\dots,D_{M-1}}$ is found by conditioning $s_{dd}|\{s_{d'd'}\}_{d'\neq d}$. The update is then,

$$\begin{aligned}\sigma^2_{s_{dd}}=&(\left\langle\tau\right\rangle+\left\langle\lambda\right\rangle)^{-1},\\ \mu_{s_{dd}}=&\sigma^2_{s_{dd}}\left\langle\mathbf{W}_d\right\rangle\left\langle\mathbf{v}_d\right\rangle\left\langle\tau\right\rangle,\\ Q(\mathbf{s}_{dd})\sim&\mathcal{TN}_{[0,\infty]}\left(\mu_{s_{dd}},\sigma^2_{s_{dd}}\right).\end{aligned}\quad(4)$$

Due to orthogonality, the dependence between d and d' disappears as $\mathbf{v}_d^\top\mathbf{V}_{\setminus d}=\mathbf{0}$ (i.e. $\mathbf{V}^\top\mathbf{V}=\mathbf{I}$). Meaning the components can be updated jointly. The first and second moment are estimated based on [15] which handles numerical issues around the tail probabilities.

Inferring λ is straight forward once $\left\langle\mathbf{SS}^T\right\rangle$ is known, the resulting update becomes,

$$\begin{aligned}\tilde{\alpha}_\lambda=\alpha_\lambda+0.5D_{M-1}\quad,\quad\tilde{\beta}_\lambda=\beta_\lambda+0.5\text{trace}(\left\langle\mathbf{SS}^\top\right\rangle),&\\ Q(\lambda)\sim\mathcal{G}(\tilde{\alpha}_\lambda,\tilde{\beta}_\lambda)&\end{aligned}\quad(5)$$

Inferring τ is simple and the computational difficulty lies in contracting all modes of the data. This simplicity arises as the second order interactions for each $\boldsymbol{\mathcal{U}}^{(m)}$ and $\mathbf{V}$ becomes the identity matrix, due to the orthogonality of the factors imposed through the von Mises-Fisher matrix distribution. Thus, the update only needs second order moments in terms of the diagonal matrix $\mathbf{S}$.

$$\begin{aligned}\tilde{\alpha}_\tau=&\alpha_\tau+0.5\prod_{m=1}^{M}I_m\\ \tilde{\beta}_\tau=&\beta_\tau+0.5\big[\text{trace}\left(\boldsymbol{\mathcal{X}}\times\boldsymbol{\mathcal{X}}\right)+\text{trace}\left(\left\langle\mathbf{SS}^\top\right\rangle\right)\\ &-2\text{trace}\left(\texttt{contract}\left(\boldsymbol{\mathcal{X}},\{\{\left\langle\boldsymbol{\mathcal{U}}^{(n)}\right\rangle\}_{n=1,\dots,M}\right)\right)\big]\\ Q(\tau)\sim&\mathcal{G}\left(\tilde{\alpha}_\tau,\tilde{\beta}_\tau\right).\end{aligned}$$

III. Results and Discussion

We analyzed the proposed PTTD in terms on synthetic data with ground truth TTD structure as well as a fluorescence spectroscopy dataset with known CPD structure.

The PTTD model and all the experiments are available, as MATLAB code, at *www.github.com/JesperLH/prob-tt*.

A. Simulated Experiments: Known or Unknown D

We simulated a TTD with five factors, where the latent dimensions are $D = (1, 6, 5, 4, 3, 1)$ and the observations in each mode are $N = (20, 19, 18, 17, 16)$. The order of the factors is an important part of how the data is generated. We refer to the generated order as the true or correct mode order, denoted by $(1, 2, 3, 4, 5)$. For a 5-way array there are $5! = 120$ possible permutations of the modes which can be ordered as $(1, 2, 3, 4, 5), (1, 2, 3, 5, 4), (1, 2, 4, 3, 5), \ldots, (5, 4, 3, 2, 1)$ using the MATLAB function `perms(5:-1:1)`.

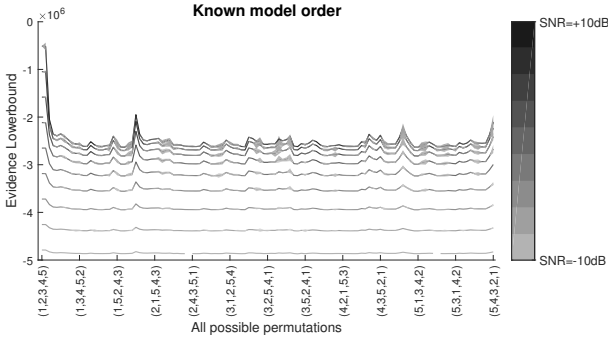
Homoscedastic Gaussian noise (white noise) is added to the generated data to obtain the desired signal-to-noise ratio (SNR). We investigate different SNRs varying from -10dB to 10dB in steps of 2.5dB, thus varying the amount of noise, but maintaining the same underlying data.

In this experiment, we only consider the probabilistic tensor train (PTTD) with either known or unknown D. If D is known then we fit PTTD with $D_{est} = D$. For unknown D, we assume we know $\texttt{max}(D) = 6$ and set $D_{est} = (1, 6, 6, 6, 6, 1)$. The models are then applied at each SNR level and for every permutation. To mitigate the effects of local optima each model fitting was repeated 10 times. The average and 10 times the standard deviation of the evidence lowerbound (ELBO) is shown in Figure 1(a) and (b) for known and unknown model order, respectively. Repeating the model fitting turned out to be unnecessary as the best model, identified by maximum ELBO, was indistinguishable from the average performance. The maximum standard deviation over all SNR levels and permutations was roughly 2 orders of magnitude lower than the mean value and barely shows in Figure 1.

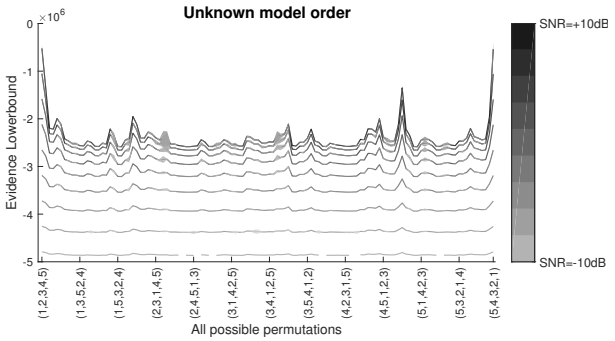
For known model order, we investigated fitting the true mode order twice, first with PTTD initialized with the true TTD, and second initialized as random orthogonal matrices. The two schemes resulted in similar mean values, but random initialization had higher standard deviation. The initializations are shown to the left in Figure 1(a), but the difference between them is negligible when compared to other mode permutations. The maximum ELBO correctly identifies the mode order $(1, 2, 3, 4, 5)$ as the true mode order, regardless of noise level.

Similarly, for unknown model order the ELBO identifies $(1, 2, 3, 4, 5)$ as the true mode order, but note the reverse mode order $(5, 4, 3, 2, 1)$ also has a high ELBO. This is happens when the latent dimensions are the same (i.e. $D_1, D_2, D_3, D_4 = 6$), so the true mode order and its reverse results in the same tensor train, but starting either from the left or right.

Contrasting the two scenarios, Figure 1(a-b) show the ELBO vary more over permutations when the model order is unknown. This is likely due to the larger subspace of the unknown tensor train. It is also observed, that as the amount of noise increases (low SNR), the difference in ELBO between permutations decreases as it becomes harder to identify the true mode order.



(a) PTTD with known model order, $D_{est} = D = (1, 6, 5, 4, 3, 1)$. The highest ELBO is achieved



(b) PTTD with unknown model order, but known $max(D) = 6$ setting $D_{est} = (1, 6, 6, 6, 6, 1)$. . The factors have the same latent dimension, i.e. $D_1, D_2, D_3, D_4 = 6$, which results in a high ELBO for both permutation (1,2,3,4,5) and (5,4,3,2,1) as the latter is the reverse of the correct mode order (i.e. the train is symmetric from end to end).

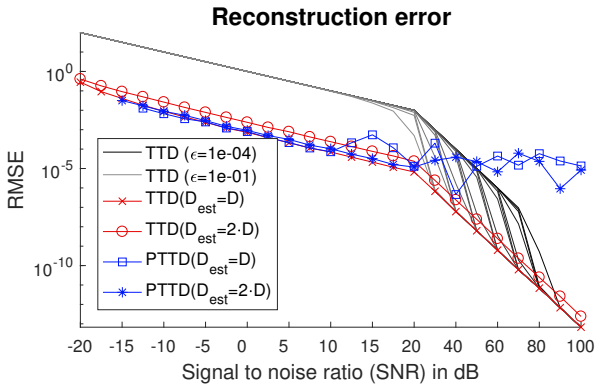
Fig. 1: **Simulated Experiments**: Performance of PTTD when the model order is known (a) and unknown (b). Higher evidence lowerbound (ELBO) indicates a better model. The permutations ($5! = 120$) are ordered from $(1, 2, 3, 4, 5), (1, 2, 3, 5, 4), (1, 2, 4, 3, 5), \ldots, (5, 4, 3, 2, 1)$.

B. Simulated Experiment: PTTD vs. TTD

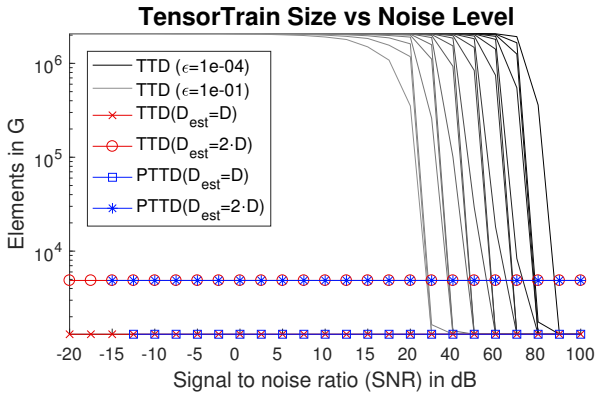
We compare the probabilistic TTD to maximum likelihood TTD and consider the scenarios where the model order is known $D_{est} = D$ or approximately known $D_{est} = 2 \cdot D$ for both PTTD and TTD, denoted as (P)TTD($D_{est} = D$) and (P)TTD($D_{est} = 2 \cdot D$). This is compared to when the model order is defined by the maximum approximation error TTD(ϵ), $\epsilon = [10^{-4}, \ldots, 10^{-1}]$, as proposed in [4].

The simulated TTD is generated precisely as in Section III-A, but we now consider more SNR levels, i.e. SNR$= [-20 : 2.5 : 20]$ dB and SNR$= [30 : 10 : 100]$ dB. We fit the model to the noisy data, $\boldsymbol{\mathcal{X}}_{noisy}$, and evaluate the performance by measuring the error between the tensor train reconstruction, $\boldsymbol{\mathcal{X}}_{recon}$ and the noiseless data $\boldsymbol{\mathcal{X}}_{truth}$.

The TTD(ϵ) uses the approximation threshold ϵ to automatically determines the size of the factors. This works well for extremely high SNR ($> +20$dB) as the noise is present on a scale that is below ϵ. However, for lower SNR the TTD(ϵ) cannot distinguish between signal and noise, and has to increase the factors sizes (e.g. more elements) to obtain an



(a) $\boldsymbol{\mathcal{X}}_{\text{truth}} - \boldsymbol{\mathcal{X}}_{\text{recon}}$



(b) $|\boldsymbol{\mathcal{G}}_{est}|$

Fig. 2: (a) The RMSE between the tensor train model and the noiseless data. (b) Number of elements of elements in the tensor train, $|\boldsymbol{\mathcal{G}}_{est}|$. The models are; PTTD (blue); TTD with fixed model order (red); TTD(ϵ) with varying $\epsilon \in [10^{-4}, 10^{-1}]$ (varying shades from dark to light grey).

ϵ-precision TTD, see Figure 2(a-b).

When the model order is (approximately) known, the performance of PTTD and TTD obtain similar results from low to high SNR ($[-20, 30]$ dB), but differ slightly at extremely high SNR, where PTTD achieve a numerical precision around 10^{-5} while TTD goes to machine precision.

C. Amino Acid Fluorescence Data

We now demonstrate how the PTTD can be used on a real dataset to determine the most likely mode and model order (i.e. mode permutation order and D) by comparing the ELBO across different model fits. The Amino Acid Fluorescence Data [16] contains five different laboratory-made samples, which are mixtures of three pure samples with known concentrations. For each mixed sample, an emission-excitation matrix was measured using fluorescence spectroscopy (excitation 250-300nm, emission 250-450mm, with resolution 1nm). The resulting dataset is a third-order tensor, $\boldsymbol{\mathcal{X}}^{5\times 201\times 61}$, with the modes being samples, emission wavelength, and excitation wavelength (e.g. original mode order $(1, 2, 3)$).

To determine the most likely mode and model order, we fit PTTD to all 6 permutations of the modes and for each permutation tested all model orders $D_1, D_2 \in [1, 30]$. Note for mode i it must be true that $D_i \leq N_i$ because of the orthogonality constraint on $\boldsymbol{\mathcal{U}}^{(i)}$. The results are shown in Table I and Figure 3.

Mode Order	ELBO	D_{ELBO}	$\lvert\boldsymbol{\mathcal{G}}_{ELBO}\rvert$
(1,2,3)	0.774	(1,4,3,1)	806
(1,3,2)	0.882	(1,4,5,1)	1330
(2,1,3)	0.969	(1,7,4,1)	1671
(2,3,1)	1.000	(1,6,5,1)	1536
(3,1,2)	0.967	(1,10,4,1)	1434
(3,2,1)	0.804	(1,4,5,1)	1274

TABLE I: **Amino Acid**: For each mode order, the best performance identified by maximum ELBO (scaled to $\max = 1$), as well as the corresponding model order $D_{ELBO} = (D_0, D_1, D_2, D_3)$ and number of elements in the tensor train $|\boldsymbol{\mathcal{G}}_{ELBO}|$ is given.

The maximum ELBO of PTTD for every permutation of the data (mode order) and for varying model order, $D = (1, D_1, D_2, 1)$, is shown in Table I. The best mode order is $(2, 3, 1)$, e.g. emission wavelength $\times$ excitation wavelength $\times$ samples, with a model order of $D = (1, 6, 5, 1)$.

The true physical model for the data is a three component non-negative CPD (actually also uni-modal, but this is not enforced). Since the five samples are mixtures of three pure samples, with known concentration, we can calculate the correlation between the estimated concentrations (i.e. factor loadings in the sample mode) and true concentrations. When the true physical model is fit to the raw data, this correlation is $\rho = 0.9993$. The exact same correlation is obtained if the model is fit to the data reconstructed by probabilistic TTD, for all the mode and model orders shown in Table I. The raw data contains 61305 element and the best tensor train has $|\boldsymbol{\mathcal{G}}| = 1536$ elements. Thus, the TTD gives a 40 times reduction in the number of elements required to represent the data while still obtaining the correct solution.

The raw data, the best PTTD, and the true physical model (fit to the raw data) are shown in Figure 3. The difference between the raw and reconstructed data (either by PTTD or non-negative CPD) is also shown. The main difference is that non-negative CPD removes both unstructured (e.g. random Gaussian) and structured (caused by Rayleigh scattering) noise while PTTD only remove unstructured noise as the Rayleigh scattering is considered part of the signal.

This experiment illustrates that PTTD can determine the best TTD mode and model order, as well as remove unstructured noise obtaining a high compressing of the original data.

IV. Conclusion

In this article, we introduced the probabilistic (Bayesian) tensor train decomposition (PTTD) and showed how the posterior distribution could be approximated using variational Bayesian (VB) inference. On both simulated and a real dataset, we showed that for VB based inference the evidence lower-bound (ELBO) can be use to identify both the mode and model order, i.e. permutation and size of TTD carts, respectively.

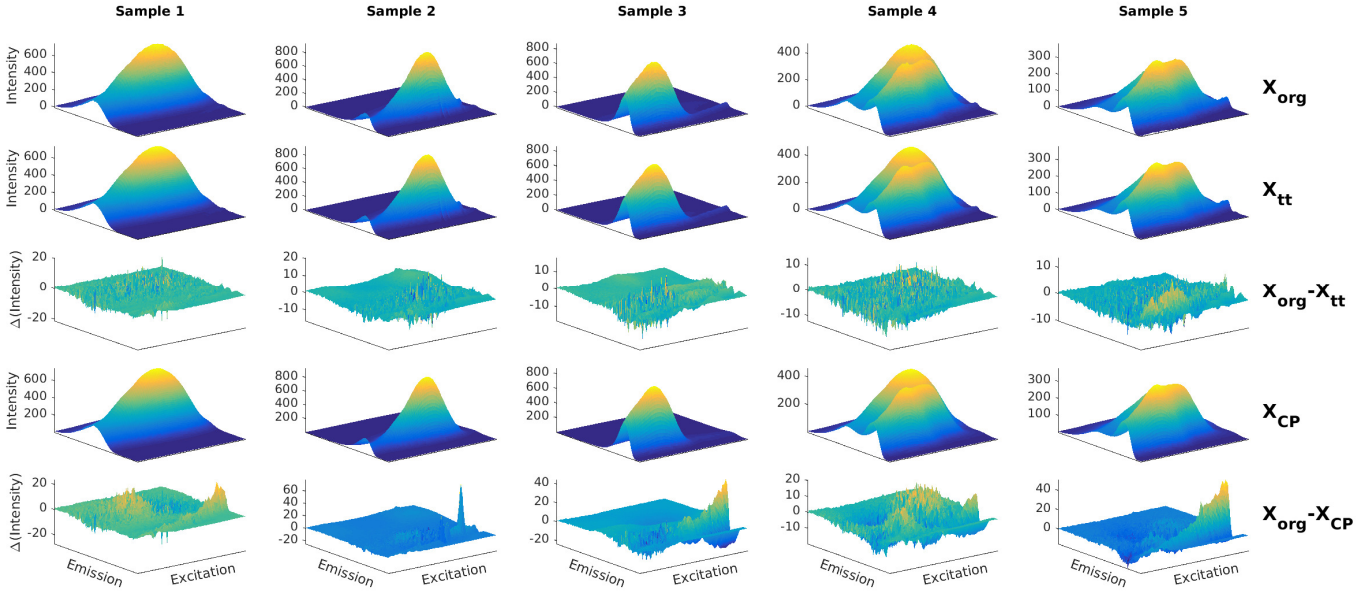


Fig. 3: **Amino Acid**: Emission Excitation Matrix for the original data (1st row), reconstruction via the best (ELBO) tensor train approximation (2nd row), and their difference $\boldsymbol{\mathcal{X}}_{\text{org}} - \boldsymbol{\mathcal{X}}_{\text{tt}}$ (3rd row). Reconstruction from the true physical model, and its difference from the observed data (4th and 5th row, respectively).

On the simulated data, we further showed that maximum likelihood TTD based on an approximation threshold ϵ (see [4]) is unable to separate signal from noise (Figure 2) unless the signal to noise ratio (SNR) is extremely high ($> +20dB$). In contrast PTTD and TTD with fixed rank perform well even at very low SNR ($\sim -15dB$).

The experiments illustrate how PTTD can be applied for data compression and to remove unstructured noise (but not structured, e.g. Rayleigh scattering). Future work should investigate how the proposed VB inference for PTTD can be scaled to large arrays and how additional constraints (c.f. [17]) for interpretability or to model physical aspects of the data imposed in the context of PTTD.

We only considered starting with a full tensor and decomposing it into a TTD. An important future extension is to consider TTD input (with large model order) and how the TTD structure of the data can be used to efficiently perform probabilistic TTD, especially efficient computation of contract() as it is the main bottleneck of our implementation. Another extension, inspired by [4], is to explore how operations such as addition, multiplication, mode contraction, etc. are carried out for PTTD and how it affects the underlying distributions.

References

[1] Tamara G Kolda and Brett W Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[2] Morten Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 24–40, 2011.

[3] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[4] Ivan V Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[5] Piyush Rai, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson, and Lawrence Carin, "Scalable bayesian low-rank decomposition of incomplete multiway tensors," in *International Conference on Machine Learning*, 2014, pp. 1800–1808.

[6] Qibin Zhao, Liqing Zhang, and Andrzej Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.

[7] Jesper L Hinrich, Søren FV Nielsen, Kristoffer H Madsen, and Morten Mørup, "Variational bayesian partially observed non-negative tensor factorization," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[8] Beyza Ermis and A Taylan Cemgil, "A bayesian tensor factorization model via variational inference for link prediction," *arXiv preprint arXiv:1409.8276*, 2014.

[9] Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-Ichi Amari, "Bayesian robust tensor factorization for incomplete multiway data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 736–748, 2016.

[10] Philip JH Jørgensen, Søren FV Nielsen, Jesper L Hinrich, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup, "Probabilistic parafac2," *arXiv preprint arXiv:1806.08195*, 2018.

[11] Peter D Hoff et al., "Equivariant and scale-free tucker decomposition models," *Bayesian Analysis*, vol. 11, no. 3, pp. 627–648, 2016.

[12] Václav ŧSmídl and Anthony Quinn, "On bayesian principal component analysis," *Computational statistics & data analysis*, vol. 51, no. 9, pp. 4101–4123, 2007.

[13] Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro, "Parafac2part i. a direct fitting algorithm for the parafac2 model," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 13, no. 3-4, pp. 275–294, 1999.

[14] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[15] John P Cunningham, Philipp Hennig, and Simon Lacoste-Julien, "Gaussian probabilities and expectation propagation," *arXiv preprint arXiv:1111.6832*, 2011.

[16] Rasmus Bro, "Parafac. tutorial and applications," *Chemometrics and intelligent laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.

[17] Michael Jauch, Peter D Hoff, and David B Dunson, "Random orthogonal matrices and the cayley transform," *arXiv preprint arXiv:1810.02881*, 2018.