2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) | 979-8-3503-4452-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAMSAP58249.2023.10403469

# Bayesian Estimation of a Probability Mass Function Tensor with Automatic Rank Detection

*Joseph K. Chege*[1], *Mikus J. Grasis*[1], *Arie Yeredor*[2], *and Martin Haardt*[1]
[1]Communications Research Laboratory, Ilmenau University of Technology, Ilmenau, Germany
[2]School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel

***Abstract*—Estimating the probability mass function (PMF) of a set of discrete random variables using a low-rank model for the PMF tensor has recently gained much attention. However, detecting the rank (model order) of the PMF tensor from observed data is a challenging problem. While classical techniques such as the Akaike and the Bayesian information criteria (AIC and BIC) may be applied in this regard, they require testing a number of candidate model orders before selecting the best one, a procedure which is computationally intensive for large datasets. In this work, we propose an algorithm to estimate the PMF tensor and implicitly detect its rank. We specify appropriate prior distributions for the model parameters and develop a deterministic algorithm which enables the rank to be detected as part of the inference. Numerical results using synthetic data demonstrate that, compared to classical model selection techniques, our approach is more robust against missing observations and is computationally efficient.**

***Index Terms*—PMF estimation, tensor decomposition, variational Bayesian inference, rank detection, model selection.**

## I. Introduction

One of the most important problems in statistical data analysis is the estimation of the joint probability mass function (PMF) of a set of discrete random variables from partial observations. Applications arise in many different contexts, for instance, inferring the label corresponding to a set of features (data classification) and inferring missing ratings from a subset of ratings (recommender systems). Unfortunately, estimating the joint PMF tensor via a histogram suffers from the curse of dimensionality since the number of observations required grows exponentially with the number of random variables. Therefore, alternative estimation approaches, which impose a low-rank nonnegative canonical polyadic decomposition (CPD) model on the PMF tensor (and thus reduce the size of the problem), have been proposed in recent years [1]–[9].

In practice, the CPD rank (or the model order) of the PMF tensor is not usually known beforehand and has to be manually specified. It was shown in [2], [3] that the CPD can be interpreted as a naïve Bayes model, for which model selection techniques such as the decomposed normalized maximum likelihood (DNML) criterion as well as the well-known Akaike and Bayesian information criteria (AIC and BIC) have been presented in [10]. Using these techniques, the rank which minimizes the respective criteria is selected from a set of candidates as the model order. The rank is thus treated as a parameter that needs to be tuned. These techniques may therefore be computationally intensive, especially for large datasets.

As an alternative, probabilistic models for the CPD factorization of a data tensor have been proposed in [11]–[14]. In these approaches, the CPD rank is determined automatically through variational Bayesian (VB) inference [15]–[17]. By initializing the rank as some suitable large value (e.g., the upper bound) and imposing sparsity-promoting priors, unnecessary components can be 'pruned', leaving only those which 'explain' the data. These works [11]–[14], however, deal with the factorization of a data tensor into its constituent CPD components. In addition, the works consider continuous data and therefore the Bayesian model specification is markedly different from the discrete case.

In this paper, our goal is to estimate the CPD components and simultaneously detect the rank of a PMF tensor from observed discrete data. We specify a Bayesian model for the problem by assigning appropriate priors to the model parameters. Exact Bayesian inference of the posterior distribution (of the model parameters given the data) using the resultant model turns out to be analytically intractable. Therefore, inspired by [11]–[14], we apply the VB framework and derive a deterministic solution to approximate the posterior distributions of various model parameters. A similar problem was considered in [18], where a Markov-Chain Monte-Carlo (MCMC) algorithm was used to infer the posterior distributions. However, while MCMC algorithms theoretically yield an exact solution, they are stochastic in nature and tend to converge slowly. On the other hand, while VB inference is only an approximation, the solution is analytical and scalable [17].

Using synthetic data, we compare the rank estimation performance as well as the accuracy of our algorithm (VB-PMF) to AIC, BIC, and DNML [10]. Numerical results under a variety of scenarios show that VB-PMF is able to estimate the CPD components of the PMF tensor while implicitly and automatically detecting its rank, obviating the need to select the rank via tuning.

## II. Preliminaries

Consider $N$ random variables $X_1, \ldots, X_N$ which can take discrete values in $[1, I_n]$, $n = 1, \ldots, N$. The joint probability mass function (PMF) of these variables is described by a

The authors gratefully acknowledge the support of the German Research Foundation (DFG) under the PROMETHEUS project (reference no. HA 2239/16-1, project no. 462458843).



Authorized licensed use limited to: TU Ilmenau. Downloaded on May 03,2024 at 13:42:36 UTC from IEEE Xplore. Restrictions apply.

PMF tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ such that $\boldsymbol{\mathcal{X}}(i_1, \ldots, i_N) = \mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N)$. Further, assume that $\boldsymbol{\mathcal{X}}$ admits a low-rank canonical polyadic decomposition (CPD) [19] with rank $R$ such that

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r \boldsymbol{A}_1(:, r) \circ \boldsymbol{A}_2(:, r) \ldots \circ \boldsymbol{A}_N(:, r), \tag{1}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^R$ is the loading vector, $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times R}$ are the factor matrices, and $\circ$ denotes the outer product. It has been shown in [3] that a rank-$R$ CPD can be interpreted as a naïve Bayes model with one latent variable taking $R$ states. Hence, $\boldsymbol{\lambda}$ and $\{\boldsymbol{A}_n\}_{n=1}^N$ are subject to a set of probability simplex constraints such that $\boldsymbol{\lambda} > \mathbf{0}$, $\boldsymbol{A}_n \geq \mathbf{0}$ (nonnegativity), and $\mathbf{1}^\mathsf{T}\boldsymbol{\lambda} = 1$, $\mathbf{1}^\mathsf{T}\boldsymbol{A}_n = \mathbf{1}^\mathsf{T}$ (sum-to-one).

In this paper, we assume the setup in [5] whereby a discrete random vector $Y = [Y_1, \ldots, Y_N]^\mathsf{T}$ is observed according to the model

$$Y_n = \begin{cases} X_n & \text{w.p. } 1 - p \\ 0 & \text{w.p. } p \end{cases}, \quad n = 1, \ldots, N, \tag{2}$$

where $p$, the outage probability, controls the fraction of missing data. Let there be $T$ i.i.d. realizations $\{\boldsymbol{y}[t]\}_{t=1}^T$ of $Y$, collected into a dataset $\boldsymbol{Y} = [\boldsymbol{y}[1], \ldots, \boldsymbol{y}[T]]$, where $\boldsymbol{y}[t] = [y_{1,t}, \ldots, y_{N,t}]^\mathsf{T}$. Estimates of the CPD components (and therefore, of the joint PMF tensor $\boldsymbol{\mathcal{X}}$) can be readily obtained from $\boldsymbol{Y}$ using maximum likelihood (ML) estimation [5]. However, this approach assumes that the CPD rank $R$ is known, a scenario which rarely arises in practice. We therefore seek to formulate the problem within the Bayesian paradigm and design an algorithm to learn the resulting probabilistic model.

## III. Bayesian Model Specification

In this section, for notational convenience, it is assumed that there are no missing observations in the data, i.e., $p = 0$. Following the model (2), the joint log-likelihood of the observations $\boldsymbol{Y}$ can be expressed (up to a constant) in terms of the CPD components (the model parameters) $\{\boldsymbol{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ as [5]

$$\log p(\boldsymbol{Y} \,|\, \boldsymbol{\lambda}, \{\boldsymbol{A}_n\}_{n=1}^N) = \sum_{t=1}^{T} \log \sum_{r=1}^{R} \lambda_r \prod_{n=1}^{N} \boldsymbol{A}_n(y_{n,t}, r). \tag{3}$$

This form of the log-likelihood is problematic to handle because the logarithm appears outside the summation over the $R$ components. To circumvent this, it is customary to introduce a set of latent variables $\boldsymbol{z}[t] \in \mathbb{R}^R$ with binary elements $z_{r,t} \in \{0, 1\}$ where $r = 1, \ldots, R$ and $\sum_{r=1}^R z_{r,t} = 1$. Each variable $\boldsymbol{z}[t]$ describes which component in $\boldsymbol{\lambda}$ gave rise to an observation $\boldsymbol{y}[t]$. In other words, if $\boldsymbol{y}[t]$ is drawn from component $s$ then $z_{r,t} = 1$ if $s = r$ and $z_{r,t} = 0$ if $s \neq r$. Thus, $\mathsf{Pr}(z_{r,t} = 1) = \lambda_r$ and since $\boldsymbol{z}[t]$ is a binary vector, we can equivalently write

$$p(\boldsymbol{z}[t]) = \prod_{r=1}^{R} \lambda_r^{z_{r,t}}. \tag{4}$$

Letting $\boldsymbol{Z} = [\boldsymbol{z}[1], \ldots, \boldsymbol{z}[T]]$, the joint log-likelihood (3) can then be rewritten (up to a constant) as

$$\log p(\boldsymbol{Y} \,|\, \boldsymbol{Z}, \{\boldsymbol{A}_n\}_{n=1}^N) = \sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{n=1}^{N} z_{r,t} \log \boldsymbol{A}_n(y_{n,t}, r). \tag{5}$$

Next, we specify the prior distributions for the latent variables $\boldsymbol{Z}$ and the CPD components $\{\boldsymbol{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$. From (4), it can be seen that $\boldsymbol{z}[t]$ is parametrized by $\boldsymbol{\lambda}$ and thus, for $T$ observations

$$p(\boldsymbol{Z} \,|\, \boldsymbol{\lambda}) = \prod_{t=1}^{T} \prod_{r=1}^{R} \lambda_r^{z_{r,t}}. \tag{6}$$

Since $\boldsymbol{\lambda}$ and $\{\boldsymbol{A}_n\}_{n=1}^N$ represent probabilities and should satisfy the probability simplex constraints, an appropriate choice for a conjugate prior is the Dirichlet distribution. Let $p(\boldsymbol{\lambda})$ and $p(\boldsymbol{a}_{n,r})$ be the prior distributions for $\boldsymbol{\lambda}$ and the $r$-th column $\boldsymbol{A}_n(:, r)$ of the $n$-th factor matrix $\boldsymbol{A}_n$, respectively. Here, we have defined $\boldsymbol{A}_n(:, r) = \boldsymbol{a}_{n,r} = [a_{n,r,1}, \ldots, a_{n,r,I_n}]^\mathsf{T} \in \mathbb{R}^{I_n}$.

Applying the Dirichlet distribution, we have

$$p(\boldsymbol{\lambda}) = \mathrm{Dir}(\boldsymbol{\lambda} \,|\, \boldsymbol{\alpha}_{\boldsymbol{\lambda}}) = C(\boldsymbol{\alpha}_{\boldsymbol{\lambda}}) \prod_{r=1}^{R} \lambda_r^{\alpha_{\boldsymbol{\lambda},r} - 1}, \tag{7}$$

$$p(\boldsymbol{a}_{n,r}) = \mathrm{Dir}(\boldsymbol{a}_{n,r} \,|\, \boldsymbol{\alpha}_{n,r}) = C(\boldsymbol{\alpha}_{n,r}) \prod_{i_n=1}^{I_n} a_{n,r,i_n}^{\alpha_{n,r,i_n} - 1}, \tag{8}$$

where $C(\cdot)$ is the respective normalization constant for the distribution (see, e.g., [17]). The Dirichlet concentration parameters $\boldsymbol{\alpha}_{\boldsymbol{\lambda}} = [\alpha_{\boldsymbol{\lambda},1}, \ldots, \alpha_{\boldsymbol{\lambda},R}]^\mathsf{T}$ and $\boldsymbol{\alpha}_{n,r} = [\alpha_{n,r,1}, \ldots, \alpha_{n,r,I_n}]^\mathsf{T}$ govern how evenly or sparsely distributed the resulting distributions are. In particular, $\alpha \to 0$ favors distributions with nearly all mass concentrated on one of their components (i.e., sparse), $\alpha \to \infty$ favors near-uniform distributions, while for $\alpha = 1$, all distributions are equally likely. In the absence of any prior information favoring one element over another, we choose symmetric Dirichlet distributions as priors, i.e., $\alpha_{\boldsymbol{\lambda},r} = \alpha_{\boldsymbol{\lambda}}\ \forall r$ and $\alpha_{n,r,i_n} = \alpha_{n,r}\ \forall i_n$.

Define $\boldsymbol{\Theta} = \{\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_N\}$ as the collection of all unknown parameters. Taking the logarithms of (6), (7), and (8) and combining with (5) gives the logarithm of the joint distribution $p(\boldsymbol{Y}, \boldsymbol{\Theta})$ as

$$\begin{aligned} \log p(\boldsymbol{Y}, \boldsymbol{\Theta}) = &\sum_{t=1}^{T} \sum_{r=1}^{R} \sum_{n=1}^{N} z_{r,t} \log a_{n,r,y_{n,t}} + \sum_{t=1}^{T} \sum_{r=1}^{R} z_{r,t} \log \lambda_r \\ &+ (\alpha_{\boldsymbol{\lambda}} - 1) \sum_{r=1}^{R} \log \lambda_r + \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{i_n=1}^{I_n} (\alpha_{n,r} - 1) \log a_{n,r,i_n} \\ &+ \text{const.} \end{aligned} \tag{9}$$

Using Bayes theorem, we can calculate the posterior distribution $p(\boldsymbol{\Theta} \,|\, \boldsymbol{Y})$ as $p(\boldsymbol{\Theta} \,|\, \boldsymbol{Y}) = p(\boldsymbol{Y}, \boldsymbol{\Theta}) / \int p(\boldsymbol{Y}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta}$. However, in this case, the denominator of the expression involves integration over all latent variables as well as CPD components and is analytically intractable. We therefore resort to variational Bayesian inference to approximate $p(\boldsymbol{\Theta} \,|\, \boldsymbol{Y})$.

## IV. Variational Approximation

Under the variational Bayesian framework [15]–[17], we seek a variational distribution $q(\boldsymbol{\Theta})$ to approximate the poste-



Authorized licensed use limited to: TU Ilmenau. Downloaded on May 03,2024 at 13:42:36 UTC from IEEE Xplore. Restrictions apply.

rior distribution $p(\boldsymbol{\Theta} \mid \boldsymbol{Y})$ by minimizing the Kullback-Leibler divergence (KLD) [20] between them, i.e.,

$$\begin{aligned} D(q(\boldsymbol{\Theta}) \,\|\, p(\boldsymbol{\Theta} \mid \boldsymbol{Y})) &= \int q(\boldsymbol{\Theta}) \log \left\{ \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta} \mid \boldsymbol{Y})} \right\} \mathrm{d}\boldsymbol{\Theta} \\ &= \log p(\boldsymbol{Y}) - \underbrace{\int \boldsymbol{q}(\boldsymbol{\Theta}) \log \left\{ \frac{p(\boldsymbol{Y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} \right\} \mathrm{d}\boldsymbol{\Theta}}_{\mathcal{L}(q)}, \end{aligned} \tag{10}$$

where $\mathcal{L}(q)$ is a lower bound on the model evidence $\log p(\boldsymbol{Y})$, or the evidence lower bound (ELBO). Since the KLD is nonnegative and $\log p(\boldsymbol{Y})$ is independent of $\boldsymbol{\Theta}$, it follows from (10) that maximizing $\mathcal{L}(q)$ is equivalent to minimizing $D(q(\boldsymbol{\Theta}) \,\|\, p(\boldsymbol{\Theta} \mid \boldsymbol{Y}))$.

Based on mean-field approximation [17], it is assumed that $q(\boldsymbol{\Theta})$ factorizes between the latent variables and the CPD components such that

$$q(\boldsymbol{\Theta}) = q_Z(\boldsymbol{Z}) q_\lambda(\boldsymbol{\lambda}) \prod_{n=1}^{N} \prod_{r=1}^{R} q_{n,r}(\boldsymbol{a}_{n,r}). \tag{11}$$

Under this setup, it can be shown that the optimal distribution for the $j$-th component of $q(\boldsymbol{\Theta})$ (i.e., the distribution for which $\mathcal{L}(q)$ is largest) is

$$\log q_j^*(\boldsymbol{\theta}_j) = \mathbb{E}_{q(\boldsymbol{\Theta} \setminus \boldsymbol{\theta}_j)}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta})] + \text{const}, \tag{12}$$

where $\mathbb{E}_{q(\boldsymbol{\Theta} \setminus \boldsymbol{\theta}_j)}[\cdot]$ is the expectation with respect to $q(\boldsymbol{\Theta})$ over all components except $\boldsymbol{\theta}_j$. In the following, we derive the optimal variational distributions for $\boldsymbol{Z}$, $\boldsymbol{\lambda}$, and all factor matrix columns $\boldsymbol{a}_{n,r}$ using (12).

### A. Variational distribution of the latent variable $\boldsymbol{Z}$

From (12), taking only the terms with a functional dependence on $\boldsymbol{Z}$ and evaluating the expectations, we find that

$$\log q_Z^*(\boldsymbol{Z}) = \sum_{t=1}^{T} \sum_{r=1}^{R} z_{r,t} \Big( \sum_{n=1}^{N} \log \widetilde{a}_{n,r,y_{n,t}} + \log \widetilde{\lambda}_r \Big) + \text{const}, \tag{13}$$

where $\log \widetilde{a}_{n,r,y_{n,t}} = \psi(\widetilde{\alpha}_{n,r,y_{n,t}}) - \psi(\widehat{\alpha}_{n,r})$, $\log \widetilde{\lambda}_r = \psi(\widetilde{\alpha}_{\boldsymbol{\lambda},r}) - \psi(\widehat{\alpha}_{\boldsymbol{\lambda}})$, $\widehat{\alpha}_{\boldsymbol{\lambda}} = \sum_{r=1}^{R} \widetilde{\alpha}_{\boldsymbol{\lambda},r}$, $\widehat{\alpha}_{n,r} = \sum_{i_n=1}^{I_n} \widetilde{\alpha}_{n,r,i_n}$, and $\psi(\cdot)$ is the digamma function. Additionally, the tilde sign distinguishes the updated hyperparameters (e.g., $\widetilde{\alpha}_{\boldsymbol{\lambda},r}$) from their initial values (e.g., $\alpha_{\boldsymbol{\lambda}}$). The optimal distribution is thus given by

$$q_Z^*(\boldsymbol{Z}) = \prod_{t=1}^{T} \prod_{r=1}^{R} \rho_{r,t}^{z_{r,t}}, \tag{14}$$

where

$$\rho_{r,t} = \frac{\gamma_{r,t}}{\sum_{j=1}^{R} \gamma_{j,t}}, \tag{15}$$

and $\gamma_{r,t} = \exp\Big\{ \sum_{n=1}^{N} \log \widetilde{a}_{n,r,y_{n,t}} + \log \widetilde{\lambda}_r \Big\}$.

The value $\rho_{r,t}$ is the conditional (posterior) probability of $\boldsymbol{z}[t]$ given $\boldsymbol{y}[t]$. It can be viewed as the 'responsibility' that the $r$-th component takes for explaining the data $\boldsymbol{y}[t]$. Indeed, from (14), we can see that $\mathbb{E}[z_{r,t}] = \rho_{r,t}$, a relation that will be useful in the sequel.

### B. Variational distribution of the loading vector $\boldsymbol{\lambda}$

Beginning with (12) and only taking terms depending on $\boldsymbol{\lambda}$, we have

$$\log q_\lambda^*(\boldsymbol{\lambda}) = \sum_{r=1}^{R} \Big( M_r + \alpha_{\boldsymbol{\lambda}} - 1 \Big) \log \lambda_r + \text{const}, \tag{16}$$

where $M_r = \sum_{t=1}^{T} \rho_{r,t}$. Taking exponentials on both sides reveals that

$$q_\lambda^*(\boldsymbol{\lambda}) \propto \prod_{r=1}^{R} \lambda_r^{\widetilde{\alpha}_{\boldsymbol{\lambda},r} - 1} \Longrightarrow q_\lambda^*(\boldsymbol{\lambda}) = \mathrm{Dir}(\boldsymbol{\lambda} \mid \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\lambda}}), \tag{17}$$

where the elements of $\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\lambda}}$ are given by

$$\widetilde{\alpha}_{\boldsymbol{\lambda},r} = M_r + \alpha_{\boldsymbol{\lambda}}. \tag{18}$$

Each element of $\boldsymbol{\lambda}$ now has its own concentration parameter $\widetilde{\alpha}_{\boldsymbol{\lambda},r}$ which depends on the data via the quantity $M_r$. Components which have no role in explaining the data will tend to zero during the inference process, producing a sparse loading vector. The rank can therefore be determined by initializing with $L > R$ components and pruning out components with a value less than some specified small positive constant $\epsilon$ after convergence.

A point estimate of $\boldsymbol{\lambda}$ can be found by computing the posterior expectation over $q_\lambda^*(\boldsymbol{\lambda})$, i.e., $\widehat{\lambda}_r = \mathbb{E}_{q_\lambda^*(\boldsymbol{\lambda})}[\lambda_r] = \frac{\widetilde{\alpha}_{\boldsymbol{\lambda},r}}{\widehat{\alpha}_{\boldsymbol{\lambda}}}$, $\forall r$.

### C. Variational distribution of the factor matrix columns $\boldsymbol{a}_{n,r}$

First, note that, by changing indices such that we sum over the discrete states $i_n$, we can rewrite the first term in (9) (cf. also (5)) as

$$\log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \{\boldsymbol{A}_n\}_{n=1}^{N}) = \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{i_n=1}^{I_n} g_{n,r,i_n} \log a_{n,r,i_n}, \tag{19}$$

where $g_{n,r,i_n} = \sum_{t:y_{n,t}=i_n} z_{r,t}$. Proceeding from (12) as before, we find

$$\log q_{n,r}^*(\boldsymbol{a}_{n,r}) = \sum_{i_n=1}^{I_n} \big( g'_{n,r,i_n} + \alpha_{n,r} - 1 \big) \log a_{n,r,i_n} + \text{const}, \tag{20}$$

where $g'_{n,r,i_n} = \sum_{t:y_{n,t}=i_n} \mathbb{E}[z_{r,t}] = \sum_{t:y_{n,t}=i_n} \rho_{r,t}$. Taking exponentials on both sides gives the optimal posterior distribution

$$q_{n,r}^*(\boldsymbol{a}_{n,r}) \propto \prod_{i_n=1}^{I_n} a_{n,r,i_n}^{\widetilde{\alpha}_{n,r,i_n} - 1} \Longrightarrow q_{n,r}^*(\boldsymbol{a}_{n,r}) = \mathrm{Dir}(\boldsymbol{a}_{n,r} \mid \widetilde{\boldsymbol{\alpha}}_{n,r}), \tag{21}$$

where the elements of $\widetilde{\boldsymbol{\alpha}}_{n,r}$ are given by

$$\widetilde{\alpha}_{n,r,i_n} = g'_{n,r,i_n} + \alpha_{n,r}. \tag{22}$$

As before, point estimates of $\boldsymbol{a}_{n,r}$ can be found by computing the posterior expectation over $q_{n,r}^*(\boldsymbol{a}_{n,r})$, i.e., $\widehat{a}_{n,r,i_n} = \mathbb{E}_{q_{n,r}^*(\boldsymbol{a}_{n,r})}[a_{n,r,i_n}] = \frac{\widetilde{\alpha}_{n,r,i_n}}{\widehat{\alpha}_{n,r}}$, $\forall n, r, i_n$.

### D. Variational lower bound $\mathcal{L}(q)$

Having derived expressions to approximate the posterior distributions, we can also find the ELBO $\mathcal{L}(q)$. Since the ELBO is a lower bound, it should increase at each iteration and can thus be used to evaluate the correctness of the



Authorized licensed use limited to: TU Ilmenau. Downloaded on May 03,2024 at 13:42:36 UTC from IEEE Xplore. Restrictions apply.

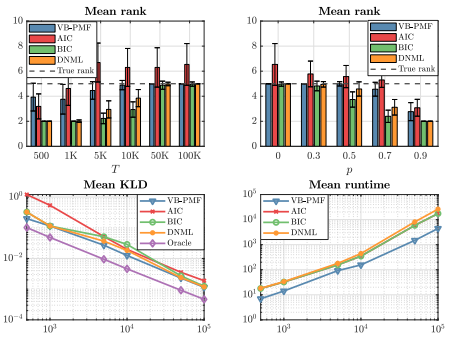


Fig. 1. Performance comparison of model selection techniques under different scenarios. From left to right: mean rank vs. $T$ ($p = 0$), mean rank vs. $p$ ($T = 10^5$), mean KLD (between the estimated and true PMF tensors) vs. $T$ ($p = 0$), and the mean CPU runtime (in seconds) vs. $T$ ($p = 0$).

mathematical expressions, as well as to test for convergence. From (10), the ELBO is given by

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\Theta})}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta})] - \mathbb{E}_{q(\boldsymbol{\Theta})}[\log q(\boldsymbol{\Theta})] \tag{23}$$

where the expectations are taken with respect to the optimal variational distributions (14), (17), and (21) in turn. Evaluating various terms in (23) gives the final expression (24) for the ELBO. The complete variational Bayes PMF estimation algorithm (VB-PMF), is summarized in Algorithm 1.

**Algorithm 1. VB-PMF**

**Input:** Dataset $\boldsymbol{Y}$
**Output:** CPD estimates $\{\widehat{\boldsymbol{A}}_n\}_{n=1}^N$ and $\widehat{\boldsymbol{\lambda}}$, rank estimate $\widehat{R}$
**Initialization:** $\alpha_{n,r}$, $\alpha_{\boldsymbol{\lambda}}$, $\rho_{r,t}$, $\forall n, r, t$
**repeat**
  Update the posterior $q_\lambda(\boldsymbol{\lambda})$ using (18)
  **for** $n = 1$ **to** $N$, $r = 1$ **to** $R$
    Update the posterior $q_{n,r}(\boldsymbol{a}_{n,r})$ using (22)
  **end for**
  Update the posterior $q_Z(\boldsymbol{Z})$ using (15)
  Evaluate the ELBO using (24)
**until convergence**
Compute posterior expectations to find $\{\widehat{\boldsymbol{A}}_n\}_{n=1}^N$ and $\widehat{\boldsymbol{\lambda}}$
Find $\widehat{R}$ by pruning out components corresponding to $\widehat{\lambda}_r < \epsilon$

## V. Results

To evaluate the performance of VB-PMF, we generate synthetic data from a 5-way ($N = 5$) PMF tensor $\boldsymbol{\mathcal{X}}$ of rank $R = 5$ and dimensions $[10, 10, 10, 10, 10]$. The tensor $\boldsymbol{\mathcal{X}}$ is constructed from CPD components $\boldsymbol{\lambda}$ and $\{\boldsymbol{A}_n\}_{n=1}^N$ drawn randomly from $U(0, 1)$ and normalized to fulfill the probability simplex constraints. Each observation $\boldsymbol{y}[t]$ is obtained by sampling a vector $\boldsymbol{x}[t]$ from the PMF $\boldsymbol{\mathcal{X}}$ randomly and independently zeroing out elements of $\boldsymbol{x}[t]$ according to the outage probability $p$ (cf. (2)).

We compare VB-PMF to the AIC, BIC, and DNML model selection techniques [10]. The hyperparameters for VB-PMF are initialized as follows: $\alpha_{\boldsymbol{\lambda}} = 10^{-6}$ in order to obtain a sparse solution for $\boldsymbol{\lambda}$ such that unnecessary CPD components can be pruned out; $\alpha_{n,r} = 1\ \forall n, r$ such that all distributions for $\{\boldsymbol{A}_n\}_{n=1}^N$ are equally likely; $\rho_{r,t}$ are randomly drawn from $U(0, 1)$ and normalized such that $\sum_{r=1}^R \rho_{r,t} = 1\ \forall t$ (cf. (15)). Moreover, VB-PMF is initialized with $L = 10$ components and, after convergence, all components with a weight $\widehat{\lambda}_r < 10^{-5}$ are discarded. The number of remaining elements of $\widehat{\boldsymbol{\lambda}}$ provide an estimate $\widehat{R}$ of the rank. For AIC, BIC, and DNML, we use an accelerated form of the EM algorithm [21] to estimate the CPD components and test candidate ranks $[2, 3, \ldots, 10]$, selecting the candidate which minimizes the respective criterion as the rank estimate. We average the results over 100 independent trials.

From Fig. 1, it can be seen that VB-PMF, BIC, and DNML converge to the true rank as the number of observations $T$ is increased. Compared to the other techniques, VB-PMF detects the correct rank with fewer observations, e.g., at $T = 10^4$, the correct rank is detected in almost all trials. As expected, AIC is biased towards models with higher ranks because it is not a consistent estimator [22].

VB-PMF is also quite robust against missing observations, mostly detecting the correct rank even for 70 % outage ($p = 0.7$). On the other hand, the performance of the other techniques deteriorates significantly as the outage is increased. The accuracy of the PMF tensor estimates is evaluated in terms of the KLD. Here, the 'Oracle' KLD [5] provides an empirical lower bound on the KLD. We see that VB-PMF provides more accurate estimates for smaller values of $T$ and a comparable performance as $T$ is increased. Finally, we observe that, compared to the other techniques, VB-PMF has the smallest CPU runtime, even though the runtime increases with $T$.

## VI. Conclusions

We have investigated the problem of detecting the model order of a low-rank PMF tensor from observed data. By formulating the problem within the Bayesian paradigm, we have proposed VB-PMF, an algorithm which is able to estimate the PMF tensor while implicitly detecting its rank as a part of the inference process. Compared to classical model selection techniques, VB-PMF is computationally efficient and is robust in the presence of missing observations while providing accurate estimates of the PMF tensor.

$$\begin{aligned}\mathcal{L}(q) = &\sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{n=1}^{N} \rho_{r,t} \log \widetilde{a}_{n,r,y_{n,t}} + \sum_{t=1}^{T}\sum_{r=1}^{R} \rho_{r,t} \log \widetilde{\lambda}_r + \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{i_n=1}^{I_n} \Big( \log C(\boldsymbol{\alpha}_{n,r}) + (\alpha_{n,r} - 1) \log \widetilde{a}_{n,r,i_n} \Big) + \log C(\boldsymbol{\alpha}_{\boldsymbol{\lambda}}) \\ &+ (\alpha_{\boldsymbol{\lambda}} - 1) \sum_{r=1}^{R} \log \widetilde{\lambda}_r - \sum_{t=1}^{T}\sum_{r=1}^{R} \rho_{r,t} \log \rho_{r,t} - \sum_{n=1}^{N}\sum_{r=1}^{R}\sum_{i_n=1}^{I_n} \Big( \log C(\widetilde{\boldsymbol{\alpha}}_{n,r}) + (\widetilde{\alpha}_{n,r,i_n} - 1) \log \widetilde{a}_{n,r,i_n} \Big) - \log C(\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\lambda}}) \\ &- \sum_{r=1}^{R} (\widetilde{\alpha}_{\boldsymbol{\lambda},r} - 1) \log \widetilde{\lambda}_r. \end{aligned} \tag{24}$$



Authorized licensed use limited to: TU Ilmenau. Downloaded on May 03,2024 at 13:42:36 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] M. Ishteva, "Tensors and Latent Variable Models," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Cham: Springer International Publishing, 2015, pp. 49–55.

[2] N. Kargas and N. D. Sidiropoulos, "Completing a joint PMF from projections: A low-rank coupled tensor factorization approach," in *Proc. 2017 Information Theory and Applications Workshop (ITA)*, Feb. 2017.

[3] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, Learning, and "Kolmogorov Extension" for Finite-Alphabet Random Vectors," *IEEE Transactions on Signal Processing*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

[4] A. Yeredor and M. Haardt, "Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence," in *Proc. IEEE 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

[5] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

[6] M. Amiridi, N. Kargas, and N. D. Sidiropoulos, "Statistical Learning Using Hierarchical Modeling of Probability Tensors," in *Proc. 2019 IEEE Data Science Workshop (DSW)*, Jun. 2019, pp. 290–294.

[7] S. Ibrahim and X. Fu, "Recovering Joint Probability of Discrete Random Variables from Pairwise Marginals," *IEEE Transactions on Signal Processing*, pp. 4116–4131, 2021.

[8] J. Vora, K. S. Gurumoorthy, and A. Rajwade, "Recovery of Joint Probability Distribution from One-Way Marginals: Low Rank Tensors and Random Projections," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Jul. 2021.

[9] P. Flores, G. Harlé, A.-B. Notarantonio, K. Usevich, M. d'Aveni, S. Grandemange, M.-T. Rubio, and D. Brie, "Coupled Tensor Factorization for Flow Cytometry Data Analysis," in *Proc. IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, Xi'an, China, Aug. 2022.

[10] K. Yamanishi, T. Wu, S. Sugawara, and M. Okada, "The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models," *Data Min Knowl Disc*, vol. 33, no. 4, pp. 1017–1058, Jul. 2019.

[11] Q. Zhao, L. Zhang, and A. 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, Sep. 2015.

[12] L. Cheng, Y.-C. Wu, and H. V. Poor, "Probabilistic Tensor Canonical Polyadic Decomposition With Orthogonal Factors," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 663–676, Feb. 2017.

[13] L. Cheng, X. Tong, S. Wang, Y.-C. Wu, and H. V. Poor, "Learning Nonnegative Factors From Tensor Data: Probabilistic Modeling and Inference Algorithm," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1792–1806, 2020.

[14] L. Cheng, Z. Chen, Q. Shi, Y.-C. Wu, and S. Theodoridis, "Towards Flexible Sparsity-Aware Modeling: Automatic Tensor Rank Learning Using the Generalized Hyperbolic Prior," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1834–1849, 2022.

[15] A. Corduneanu and C. Bishop, "Variational bayesian model selection for mixture distributions," in *Proc. Eighth International Conference on Artificial Intelligence and Statistics*. Morgan Kaufmann, 2001, pp. 27–34.

[16] J. Winn and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, p. 661–694, Dec. 2005.

[17] C. M. Bishop, *Pattern recognition and machine learning, 5th Edition*, ser. Information science and statistics. Springer, 2007.

[18] D. B. Dunson and C. Xing, "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1042–1051, Sep. 2009.

[19] T. G. Kolda and B. W. Bader, "Tensor Decompositions and Applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[21] J. K. Chege, M. J. Grasis, A. Manina, A. Yeredor, and M. Haardt, "Efficient Probability Mass Function Estimation from Partially Observed Data," in *Proc. IEEE 56th Asilomar Conference on Signals, Systems, and Computers*, Oct. 2022.

[22] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Comp Stat*, vol. 11, no. 3, May 2019.




Authorized licensed use limited to: TU Ilmenau. Downloaded on May 03,2024 at 13:42:36 UTC from IEEE Xplore. Restrictions apply.