Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence

Arie Yeredor
School of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
arie@eng.tau.ac.il

Martin Haardt
Communications Research Laboratory
Ilmenau University of Technology
Ilmenau, Germany
martin.haardt@tu-ilmenau.de

***Abstract*—Recently there has been a growing interest in the estimation of the Probability Mass Function (PMF) of discrete random vectors (RVs) from partial observations thereof (namely when observed realizations of the RV are limited to random subsets of its elements). It was shown that under a low-rank assumption on the PMF tensor (and some additional mild conditions), the full tensor can be recovered, e.g., by applying an approximate coupled factorization to empirical estimates of all joint PMFs of subgroups of fixed cardinality larger than two (e.g., triplets). The coupled factorization is based on a Least Squares (LS) fit to the empirically estimated lower-order sub-tensors. In this work we take a different approach by trying to fit the coupled factorization to estimated sub-tensors in the sense of minimizing the Kullback-Leibler divergence (KLD) between the estimated and inferred tensors. We explain why the KLD-based fitting is better-suited than LS-based fitting for the problem of PMF estimation, propose an associated minimization approach and demonstrate some advantages over LS-based fitting in this context using simulation results.**

***Index Terms*—Low-Rank Tensor Factorization, Probability Mass Function (PMF), Approximate Coupled Factorization, Kullback-Leibler Divergence (KLD), Nonnegative Tensor Factorization, Canonical Polyadic Decomposition (CPD).**

I. Introduction

Tensor representations and tensor factorization are gaining increased popularity in the fields of signal processing, estimation theory and machine learning, not only as tools for multi-way data representation and analysis [1] or source separation [2], but also (more recently) in high-order performance analysis using high-order derivatives and moments [3] and in statistical representation of Probability Mass Functions (PMFs) [4]. Indeed, in recent work by Kargas *et al.* [4], it was shown that when the PMF of a discrete random vector (whose elements take values in finite alphabets) can be represented by a low-rank tensor, such a representation can be interpreted as a naïve Bayes model, and, moreover, the full PMF tensor can be recovered from knowledge of all of its sub-tensors of degrees larger than 2 (e.g., triplets or quadruples).

Often in machine learning and in related applications (e.g., recommender systems), only partial observations of the random vector are available, where some (or even most) of its elements are missing in each observed realization. Under these conditions, it may be possible to obtain empirical estimates of all sub-tensors of a certain degree from the co-occurrence histograms of small groups of the vector's elements (e.g., triplets), but not of the entire tensor. It is therefore suggested in [4] to estimate the full tensor by applying a low-rank approximate coupled factorization to the empirically obtained sub-tensors, where the criterion for the coupled factorization is the ordinary Least Squares (LS) criterion (expressed in terms of the sum of Frobenius norms of the differences between the empirical sub-tensors and the implied sub-tensors of the estimated full tensor).

While offering relatively convenient optimization procedures, the LS criterion entails a conceptual drawback in this context. In particular, it is not severely penalized when attributing an extremely small (or even zero) probability to certain elements of the estimated PMF, even when the empirical evidence may suggest that the respective vector values are feasible. The Kullback-Leibler Divergence (KLD, [10]), on the other hand, is a non-negative (asymmetric) distance measure, denoted $D(\boldsymbol{\mathcal{X}}\|\boldsymbol{\mathcal{Y}})$, between any two (equal-size) PMF tensors $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ (see the explicit notation definition in the sequel), which equals zero iff $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}}$, and approaches infinity if an element of $\boldsymbol{\mathcal{Y}}$ approaches zero while the respective element of $\boldsymbol{\mathcal{X}}$ is nonzero.

Therefore, using the KLD as a substitute to the LS criterion for measuring the fit between the estimated low-rank tensor and the empirical sub-tensors is a more "natural" choice in the context of PMF estimation. Moreover, it can be shown that under a proper setup, the resulting KLD-based estimate coincides with the Maximum Likelihood (ML) estimate of the low-rank PMF from the empirical sub-tensors, unlike the LS-based estimate[1]. In this paper we propose a modification of the approximate coupled low-rank factorization approach, which uses the KLD instead of the LS criterion, and demonstrate the resulting improvement in the PMF estimation in a small-scale

[1]In fact, the LS-based estimate can be shown to approach the ML estimate, and therefore the KLD-based estimate (under the same setup), only asymptotically.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

("toy-example") simulation scenario.

The use of the KLD is certainly not new to Nonnegative Matrix / Tensor Factorization (NMF / NTF) problems, and has been proposed before, e.g., by Hansen *et al.* [6], Chi and Kolda [7], and as a particular case of β-divergence also by Cichocki and Phan [8] (see also [9]). KLD was also considered recently by Kargas and Sidiropoulos in the context of using CPD for learning mixtures of smooth product distributions [5]. However, in the general context of NMF / NTF, the KLD is not associated with PMF tensors estimation, and is therefore not constrained to have all factor matrices and loading factors restricted to the probability simplex (see Section II for more details). Additionally, KLD has only been considered in the context of a single matrix / tensor factorization, but not in an approximate coupled factorization of several sub-tensors. Therefore, the main innovation in this work is in posing and solving the approximate coupled factorization associated with the PMF estimation from sample sub-tensors using the KLD under the probability simplex constraints, and in comparing the results to the LS-based counterpart estimates.

As an interesting, possibly unexpected by-product, we also observed that when using an alternating-directions scheme (decomposing the non-convex coupled factorization optimization into a series of convex optimization sub-problems with respect to each factor matrix separately), a significantly smaller number of iterations ("full sweeps") of the KLD-based optimization is required, relative to the LS-based optimization.

II. Problem Formulation

Let $\boldsymbol{X} = [X_1,\ X_2,\ \ldots,\ X_N]^T \in \mathbb{R}^N$ be a discrete random vector with with X_n taking discrete integer values in $[1, I_n]$ ($n = 1, \ldots N$). We denote its joint PMF tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, where $\mathcal{X}(i_1, i_2 \ldots, i_N) = Pr\{X_1 = i_1, X_2 = i_2 \ldots, X_N = i_N\}$. The goal is to find a non-negative Canonical Polyadic Decomposition (CPD) of $\boldsymbol{\mathcal{X}}$ with F factors, namely to find N factor matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \ldots, \boldsymbol{A}_N$ ($\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$) and a "loading vector" $\boldsymbol{\lambda} \in \mathbb{R}^F$ such that

$$\boldsymbol{\mathcal{X}} \approx \sum_{f=1}^{F} \lambda_f \cdot \boldsymbol{A}_1(:, f) \circ \boldsymbol{A}_2(:, f) \circ \cdots \circ \boldsymbol{A}_N(:, f) \triangleq [\![\boldsymbol{\lambda}, \boldsymbol{A}_1, \boldsymbol{A}_2 \ldots, \boldsymbol{A}_N]\!] \quad (1)$$

where $\circ$ denotes an outer product of vectors. All elements of $\boldsymbol{\lambda}$ are positive, all elements of $\boldsymbol{A}_1, \boldsymbol{A}_2 \ldots, \boldsymbol{A}_N$ are non-negative and $\mathbf{1}^T\boldsymbol{\lambda} = 1$, $\mathbf{1}^T\boldsymbol{A}_n = \mathbf{1}^T$, $n = 1, \ldots, N$, where $\mathbf{1}$ denotes an all-ones vector with context-implied dimensions (this set of constraints is sometimes abbreviated as "confining to the *probability simplex*").

It is shown in [4] that even in the absence of an estimate of the full $\boldsymbol{\mathcal{X}}$, the factors can be obtained from empirical estimates of the joint PMF in triplets, denoted $\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell}$, for a sufficient number of combinations of (j, k, ℓ) (with $j < k < \ell$), from which the factor matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_N$ and the loading vector $\boldsymbol{\lambda}$ can in turn be estimated by applying an approximate coupled tensor factorization to all (empirically-estimated) triplets-PMFs.

It is proposed in [4] to jointly factorize the triplet tensors by minimizing a least-squares (LS) criterion

$$\min_{\{\boldsymbol{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \sum_j \sum_{k>j} \sum_{\ell>k} \frac{1}{2} \left\| \widehat{\boldsymbol{\mathcal{X}}}_{jk\ell} - [\![\boldsymbol{\lambda}, \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!] \right\|_F^2 \quad (2)$$
$$\text{subject to:} \quad \boldsymbol{\lambda} > \mathbf{0}, \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} = 1$$
$$\boldsymbol{A}_n \geq \mathbf{0}, \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{A}_n = \mathbf{1}^{\mathrm{T}}, \quad n = 1, \ldots, N,$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm (sum of all squared elements of the enclosed tensor).

III. The Proposed Alternative

Our proposal is to substitute the LS criterion with a KLD-based criterion, measuring the distances between the estimated (empirical) PMF (in triplets) and the theoretical PMF implied by the factors.

Letting $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ denote any two 3-way PMF tensors (of the same dimensions), the KLD between them is defined (up to some vanishing constants) as

$$D(\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{Y}}) \triangleq - \sum_{i_1, i_2, i_3} \mathcal{X}(i_1, i_2, i_3) \log \frac{\mathcal{Y}(i_1, i_2, i_3)}{\mathcal{X}(i_1, i_2, i_3)} \geq 0, \quad (3)$$

where equality holds iff the two PMFs are identical (and where $0 \cdot \log 0$ and $0 \cdot \log \infty$ are both taken to be zeros). When $\boldsymbol{\mathcal{X}}$ is fixed, minimizing $D(\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{Y}})$ w.r.t. $\boldsymbol{\mathcal{Y}}$ (subject to the probability simplex constraints, which keep $\boldsymbol{\mathcal{Y}}$ a valid PMF tensor) boils down to minimizing

$$D(\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{Y}}) = C - \sum_{i_1, i_2, i_3} \mathcal{X}(i_1, i_2, i_3) \log \mathcal{Y}(i_1, i_2, i_3) = C + \|\boldsymbol{\mathcal{X}} \odot \log \boldsymbol{\mathcal{Y}}\|_1, \quad (4)$$

where C is an irrelevant constant, $\odot$ denotes the Hadamard (element-wise) product, $\log \boldsymbol{\mathcal{Y}}$ is interpreted element-wise and $\|\cdot\|_1$ denotes the ℓ_1 norm of the enclosed tensor, which is the sum of absolute values of all of its elements (which, in our case, are all non-positive). Note that since all elements of $\boldsymbol{\mathcal{X}}$ are non-negative, under the non-negativity constraint on all elements of $\boldsymbol{\mathcal{Y}}$, $D(\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{Y}})$ is a convex function of $\boldsymbol{\mathcal{Y}}$.

Returning to our joint factorization problem, our proposed criterion is therefore defined as

$$\min_{\{\boldsymbol{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \sum_j \sum_{k>j} \sum_{\ell>k} \log \left\| \widehat{\boldsymbol{\mathcal{X}}}_{jk\ell} \odot [\![\boldsymbol{\lambda}, \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!] \right\|_1 \quad (5)$$
$$\text{subject to:} \quad \boldsymbol{\lambda} > \mathbf{0}, \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} = 1$$
$$\boldsymbol{A}_n \geq \mathbf{0}, \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{A}_n = \mathbf{1}^{\mathrm{T}}, \quad n = 1, \ldots, N.$$

Such a KLD-based criterion would attain a better fit (in the context of PMF interpretation) to the empirical PMFs than the LS criterion, because it would treat smaller empirical probabilities in a different way than treating higher empirical probabilities - unlike the LS criterion, which allows (and actually promotes) similar-size deviations from all probabilities. For example, unlike the LS criterion, the KLD-based criterion would forbid a solution (a factors model) that attributes zero probability to an element of one of the tensors in which the empirical probability is non-zero, even if that probability

is very small (because then the KLD-based criterion would diverge to infinity). And indeed, a non-negative probability in an element of one of the empirical tensors means that the respective combination of values has a positive probability of occurrence, so it would not be reasonable (from a probabilistic point of view) to attribute zero probability in the model for such a combination, which means that this combination is not feasible - in contradiction to the empirical evidence.

IV. Minimization of the KLD-Based Criterion

Our "outer" minimization strategy follows the alternating directions minimization proposed in [4], which decomposes the non-convex minimization in (5) into a series of convex minimization problems w.r.t. each of the factor matrices $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$, separately. However, for the "inner" (convex) optimization problems we propose a different approach, absorbing the probability simplex' equality constraints by re-parameterizing the problem.

We begin with a brief description of the "outer", alternating-directions minimization. When minimizing w.r.t. one of the factor matrices, say $\boldsymbol{A}_m$, alone, we assume that all other factor matrices, as well as $\boldsymbol{\lambda}$, are fixed (at the values from their latest estimates). We then need to solve

$$\min_{\boldsymbol{A}_m} \sum_{k\neq m} \sum_{\substack{\ell>k \\ \ell\neq m}} \left\| \widehat{\boldsymbol{\mathcal{X}}}^{(1)}_{mk\ell} \odot \log\left[(\boldsymbol{A}_\ell \diamond \boldsymbol{A}_k)\mathrm{Diag}(\boldsymbol{\lambda})\boldsymbol{A}_m^{\mathrm{T}}\right] \right\|_1 \tag{6}$$

$$\text{subject to:} \quad \boldsymbol{A}_m \geq \boldsymbol{0}, \quad \boldsymbol{1}^{\mathrm{T}}\boldsymbol{A}_m = \boldsymbol{1}^{\mathrm{T}},$$

where $\widehat{\boldsymbol{\mathcal{X}}}^{(1)}_{mk\ell}$ denotes the 1-mode unfolding of $\widehat{\boldsymbol{\mathcal{X}}}_{mk\ell}$ (see [1], [4]), $\diamond$ denotes the Khatri-Rao (column-wise Kronecker) product, $\mathrm{Diag}(\boldsymbol{\lambda}) \in \mathbb{R}^{F\times F}$ denotes a diagonal matrix with the elements of $\boldsymbol{\lambda}$ along its diagonal and $\geq$ for matrices denotes elementwise inequality. Then, when minimizing w.r.t. $\boldsymbol{\lambda}$ (assuming all $\boldsymbol{A}_n$ are fixed) we need to solve:

$$\min_{\boldsymbol{\lambda}} -\sum_j \sum_{k>j} \sum_{\ell>k} \mathrm{vec}^{\mathrm{T}}\left(\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell}\right) \cdot \log\left[(\boldsymbol{A}_\ell \diamond \boldsymbol{A}_k \diamond \boldsymbol{A}_j)\boldsymbol{\lambda}\right] \tag{7}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} > \boldsymbol{0}, \quad \boldsymbol{1}^{\mathrm{T}}\boldsymbol{\lambda} = 1,$$

where $\mathrm{vec}\left(\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell}\right) \in \mathbb{R}^{I_j I_k I_\ell}$ denotes the vectorized form composed of the concatenation of all columns of $\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell}$ (in "natural" order, see [4]).

We refer to an "outer" iteration, namely to a complete cycle through all $N+1$ "inner" minimization problems (w.r.t. $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_N$ and $\boldsymbol{\lambda}$) as a "sweep". Since after each inner minimization the overall value of the global KLD-based criterion in (5) is guaranteed not to increase (and usually to decrease) and is bounded below (by zero), convergence of the criterion to a stationary point is guaranteed (although the point of convergence is not necessarily the global minimum of the criterion).

To solve the inner (convex) minimization problems we propose a different approach than [4], as we prefer to re-parameterize the problem in terms of (slightly) fewer parameters, automatically accounting for the probability simplex' equality constraints. To this end, observe that any vector $\boldsymbol{a} \in \mathbb{R}^K$ of the form $\boldsymbol{a} = \boldsymbol{e}_1 - \boldsymbol{Q}\boldsymbol{b}$, where $\boldsymbol{e}_k$ denotes the k-th column of the $K \times K$ identity matrix, the matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times(K-1)}$ is defined as

$$\boldsymbol{Q} \triangleq [\boldsymbol{e}_1 - \boldsymbol{e}_2, \ \boldsymbol{e}_2 - \boldsymbol{e}_3, \ \ldots, \ \boldsymbol{e}_{K-1} - \boldsymbol{e}_K] \tag{8}$$

and $\boldsymbol{b} \in \mathbb{R}^{K-1}$ is an arbitrary vector, is easily seen to satisfy the equality constraint $\boldsymbol{1}^T\boldsymbol{a} = 1$ (since $\boldsymbol{Q}$ is a basis for the null-space of $\boldsymbol{1}^{\mathrm{T}}$). In order for $\boldsymbol{a}$ to also satisfy the non-negativity constraint, note that, due to the structure of $\boldsymbol{Q}$ we have, from $\boldsymbol{a} = \boldsymbol{e}_1 - \boldsymbol{Q}\boldsymbol{b}$,

$$\begin{aligned} a_1 = 1 - b_1, \ \ a_2 = b_1 - b_2, \ \ a_3 = b_2 - b_3, \ \ \cdots \\ \cdots \ \ a_{K-1} = b_{K-2} - b_{K-1}, \ \ a_K = b_{K-1}. \end{aligned} \tag{9}$$

We therefore conclude that $\boldsymbol{a} \geq \boldsymbol{0}$ iff the elements of $\boldsymbol{b}$ satisfy the ordering constraint $1 \geq b_1 \geq b_2 \geq \cdots \geq b_{K-1} \geq 0$.

We can therefore re-parameterize each of the factor matrices as $\boldsymbol{A}_n = \boldsymbol{E}_n - \boldsymbol{Q}\boldsymbol{B}_n$, where $\boldsymbol{E}_n \in \mathbb{R}^{I_n\times F}$ is an all-zeros matrix except for all-ones on its first row, and where $\boldsymbol{B}_n \in \mathbb{R}^{(K-1)\times F}$ is an arbitrary matrix. The constrained minimization problem (6) can then be recast in terms of $\boldsymbol{B}_m$ as

$$\min_{\boldsymbol{B}_m} \sum_{k\neq m} \sum_{\substack{\ell>k \\ \ell\neq m}} \left\| \widehat{\boldsymbol{\mathcal{X}}}^{(1)}_{mk\ell} \odot \log\left[\boldsymbol{C}_{\ell k}(\boldsymbol{E}_m - \boldsymbol{Q}\boldsymbol{B}_m)^{\mathrm{T}}\right] \right\|_1 \tag{10}$$

$$\text{s.t.:} 1 \geq B_m(1,f) \geq B_m(2,f) \geq \cdots \geq B_m(K-1,f) \geq 0, \quad f = 1, \ldots, F,$$

where $\boldsymbol{C}_{\ell k} \triangleq (\boldsymbol{A}_\ell \diamond \boldsymbol{A}_k)\mathrm{Diag}(\boldsymbol{\lambda})$.

The derivative of the inner term (summand) of (10) w.r.t. $B_m(i,f)$ is given by

$$-\left\| \widehat{\boldsymbol{\mathcal{X}}}^{(1)}_{mk\ell} \oslash \left[\boldsymbol{C}_{\ell k}(\boldsymbol{E}_m - \boldsymbol{Q}\boldsymbol{B}_m)^{\mathrm{T}}\right] \odot \left(\boldsymbol{C}_{\ell k}(:,f)(\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{e}_{i+1}^{\mathrm{T}})\right) \right\|_1$$

where $\oslash$ denotes element-wise division. Adding these terms up (according to the summation in (10)) we get the derivative w.r.t. each element of $\boldsymbol{B}_m$. Starting with any feasible solution, e.g., $\boldsymbol{B}_m(i,f) = (K-i)/K$ for all f, we take the following update strategy to maintain feasibility: Starting with the strongest (positive / negative) derivative, we modify the respective element of $\boldsymbol{B}_m$ with a nominal step-size of some sufficiently small negative constant α times the derivative (increasing or decreasing, depending on the sign), bounded by its neighboring element (above or below, resp.) on the same column of $\boldsymbol{B}_m$ - so as to maintain the ordering constraint in (10). We then proceed through the other elements of $\boldsymbol{B}_m$ in descending order of the absolute values of their derivatives. Once all elements of $\boldsymbol{B}_m$ have been updated, the derivatives are recalculated and the process is repeated until convergence is attained.

For the minimization w.r.t. $\boldsymbol{\lambda}$ we take a similar strategy after re-parameterizing $\boldsymbol{\lambda} = \boldsymbol{e}_1 - \boldsymbol{Q}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} \in \mathbb{R}^{F-1}$, satisfying $1 > \gamma_1 > \gamma_2 > \cdots > \gamma_{F-1} > 0$ (note that $\boldsymbol{\lambda}$ should be strictly positive, hence the strict inequalities).

V. Simulation Experiments

To demonstrate the advantages of KLD-based minimization over LS-based minimization, we present simulation results of a small-scale "toy-example" experiment.

We generated a rank-2 ($F = 2$) 5-way ($N = 5$) PMF tensor $\boldsymbol{\mathcal{X}}$ of dimensions $[2, 3, 4, 3, 2]$ using the following (arbitrarily chosen) factor matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 0.1 & 0.4 \\ 0.9 & 0.6 \end{bmatrix} \boldsymbol{A}_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.1 & 0.4 \\ 0.4 & 0.1 \end{bmatrix} \boldsymbol{A}_3 = \begin{bmatrix} 0.2 & 0.4 \\ 0.5 & 0.1 \\ 0.1 & 0.1 \\ 0.2 & 0.4 \end{bmatrix}$$

$$\boldsymbol{A}_4 = \begin{bmatrix} 0.1 & 0.2 \\ 0.1 & 0.4 \\ 0.8 & 0.4 \end{bmatrix} \quad \text{and} \quad \boldsymbol{A}_5 = \begin{bmatrix} 0.5 & 0.2 \\ 0.5 & 0.8 \end{bmatrix} \tag{11}$$

and the loading vector $\boldsymbol{\lambda} = [0.7, \ 0.3]^{\mathrm{T}}$.

Then T random vectors (of dimension 5) were drawn according to the resulting PMF, with T taking the (logarithmically equally-spaced) values 10,000, 25,119, 63,096, 158,489, 398,107 and 1,000,000. The empirical 3-way tensors representing all the empirical 3-way marginal distributions were calculated based on the drawn vectors, and the joint factorization criteria (2) (LS) and (5) (KLD) were applied in order to obtain estimates of the true PMF tensor $\boldsymbol{\mathcal{X}}$ based on all (ten) 3-way empirical tensors $\{\mathcal{X}_{123}, \mathcal{X}_{124} \ldots, \mathcal{X}_{345}\}$.

We then applied alternating directions minimization (as described in the previous section) with respect to each factor matrix and to $\boldsymbol{\lambda}$ according to the two different criteria. Both minimization procedures started with the same initial guess, which was generated as follows. First, we find the F largest elements (probabilities) in the full empirical tensor $\widehat{\boldsymbol{\mathcal{X}}}$ and denote their values as p_f and their indices vectors as $I_f = [i_f^{(1)}, \ i_f^{(2)} \ \ldots, \ i_f^{(N)}]^{\mathrm{T}}$ (for $f = 1, \ldots, F$). Then we initialize $\boldsymbol{\lambda} = [p_1, \ldots, p_F]^{\mathrm{T}} / \sum p_f$ and set the f-th column of the initial $\boldsymbol{A}_n$ to all-zeros except for a 1 as its $i_f^{(n)}$-th element (for all $n = 1, \ldots, N$ and $f = 1, \ldots, F$). We note that this initialization strategy is based on knowledge of the full empirical tensor $\widehat{\boldsymbol{\mathcal{X}}}$, which is usually not available in practice (in fact, its presumed availability obviates the need for coupled factorization...) but since the initialization strategy is not an issue here, we chose to use this slightly impractical scheme in order to obtain a fair comparison between the minimization criteria.

Figure 1 shows typical convergence patterns (of the respective criteria (2) and (5)), as a function of the sweep number until a convergence criterion is met. Both algorithms feature a monotonic descent, as could be expected for alternating-directions type minimization.

In order to compare the PMF estimation performance of the two approaches in this experiment, we first define the following performance measures:

- LS-fit between the empirical and the true PMF:

$$\left\| \widehat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}} \right\|_F^2$$

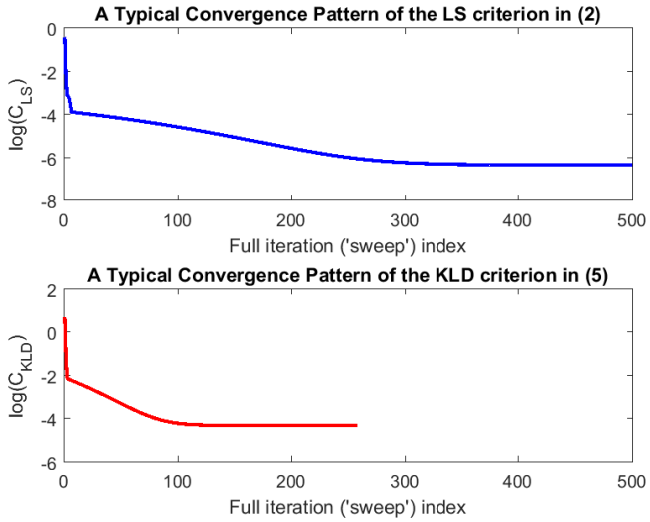

Fig. 1. Typical convergence patterns (log of the criterion vs. full iteration ["sweep"] index) for the LS and KLD criteria. The minimization based on the LS criterion is typically slower than the minimization based on the KLD criterion in terms of the number of sweeps.

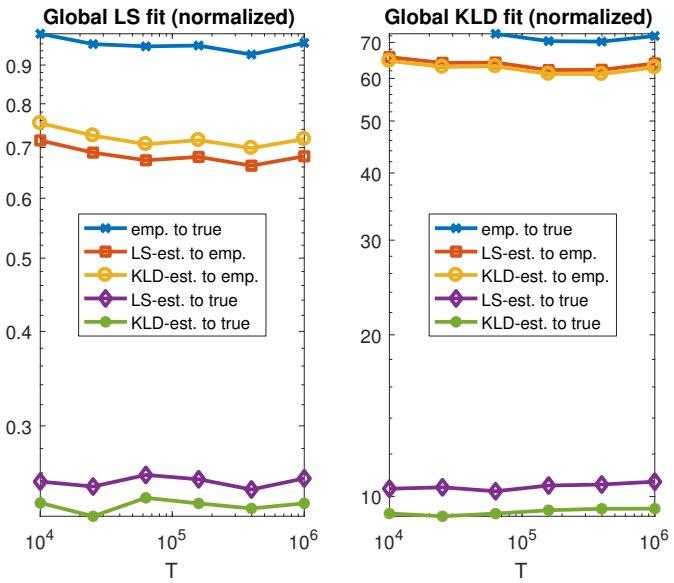

Fig. 2. The (T-normalized) LS-fit (left) and KLD-fit (right) between: the sample- and true PMF; the LS-based estimate and the sample-PMF; the KLD-based estimate and the sample-PMF; the LS-based estimate and the true PMF; the KLD-based estimate and the true PMF. Results are based on averaging 100 independent trials.

- LS-fit between an estimated and the empirical PMF:

$$\left\| \widehat{\boldsymbol{\mathcal{X}}} - \left[\left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \widehat{\boldsymbol{A}}_3, \widehat{\boldsymbol{A}}_4, \widehat{\boldsymbol{A}}_5 \right] \right] \right\|_F^2$$

- LS-fit between an estimated and the true PMF:

$$\left\| \boldsymbol{\mathcal{X}} - \left[\left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \widehat{\boldsymbol{A}}_3, \widehat{\boldsymbol{A}}_4, \widehat{\boldsymbol{A}}_5 \right] \right] \right\|_F^2$$

and, similarly,

- KLD-fit between the empirical and the true PMF:

$$\left\| \boldsymbol{\mathcal{X}} \odot (\log \widehat{\boldsymbol{\mathcal{X}}} - \log \boldsymbol{\mathcal{X}}) \right\|_1 .$$

- KLD-fit between an estimated and the empirical PMF:

$$\left\| \widehat{\boldsymbol{\mathcal{X}}} \odot (\log \left[\!\left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \widehat{\boldsymbol{A}}_3, \widehat{\boldsymbol{A}}_4, \widehat{\boldsymbol{A}}_5 \right]\!\right] - \log \widehat{\boldsymbol{\mathcal{X}}}) \right\|_1 .$$

- KLD-fit between an estimated and the true PMF:

$$\left\| \boldsymbol{\mathcal{X}} \odot (\log \left[\!\left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \widehat{\boldsymbol{A}}_3, \widehat{\boldsymbol{A}}_4, \widehat{\boldsymbol{A}}_5 \right]\!\right] - \log \boldsymbol{\mathcal{X}}) \right\|_1 .$$

Figure 2 shows these normalized performance measures for both types (LS-based and KLD-based) of estimates, vs. the observation time (number of available independent realizations T). In order to elucidate the differences, the performance measures in these plots were normalized by multiplication with T, so that they all appear at (roughly) constant logarithmic scales, which reflect an actual $1/T$-type decrease in all of these criteria. All the values shown in these plots are based on averaging the results from 100 independent experiments, in which both algorithms were applied using the same empirical data.

The worst fit is observed between the empirical and the true PMF, both in terms of the LS fit and in terms of the KLD fit. This is rather expected, as the low-rank estimates are expected to be closer than the raw empirical tensor to the true (low-rank) tensor, due to the "denoising" effect of the low-rank approximation. Observing the fit of the estimates to the empirical tensor, we note that, as could be expected, each estimate outperforms the other under its "own" criterion, namely, under the LS-fit performance measure, the LS-based estimate offers a better fit (to the empirical PMF) than the KLD-based estimate - and vice-versa. However, when it comes to fitting the *true* tensor, the KLD-based estimate is seen to provide a better fit not only under the KLD-fit performance measure, but also under the LS-fit performance measure. Note that there is no contradiction to common-sense here, since the estimation criterion is not based on fitting the true tensor, but only on fitting the empirical tensor (or, rather, all triplets sub-tensors of this empirical tensor) - the obvious conclusion here is that (at least in this experiment) KLD-based estimation outperforms LS-based estimation with respect to estimating the true PMF, under both performance measures.

As a final note we show in Figure 3 the average number of full iterations (sweeps) to convergence vs. the observation length T for the two approaches. The KLD-based minimization is seen to converge faster (at least in this example) than the LS-based minimization. Also, as T gets larger, the required number of sweeps (for both methods) increases, possibly due to the improving accuracy of the sample-PMF, which enables a better fit of the low-rank model, resulting in a longer refinement (the convergence (stopping) criterion is a streak of $L = 10$ sweeps in which the change of the log of the minimized criterion remains below a threshold). We note, however, that the computation complexity per sweep is somewhat higher for the KLD-based minimization than for the LS-based minimization.

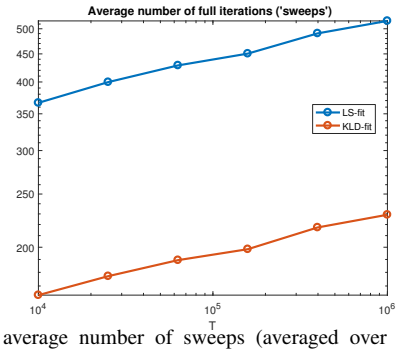

Fig. 3. The average number of sweeps (averaged over 100 independent trials) required for convergence of the two criteria in alternating-directions minimization. The KLD-based estimation is seen to feature faster convergence (in terms of the number of sweeps) than the LS-based estimation (whereas the running time per iteration for KLD is about 50% higher than for LS).

VI. Conclusion

We have addressed the problem of estimating a low-rank PMF tensor (of a discrete random vector taking values over a finite alphabet) from empirical sub-tensors, using approximate coupled factorization. The main message of our work is that a KLD-based criterion for the approximate joint factorization is better suited to this particular estimation problem than the LS-based criterion (used in [4]) and attains a more accurate estimate of the true tensor, not only in terms of the resulting KLD fit, but also in terms of the resulting LS-fit (which is also the mean square estimation error). In future work we intend to show that the KLD-based estimate coincides (under a slight change of the paradigm) with the ML estimate w.r.t. the (partial) observations of the random vector's realizations.

References

[1] N. D. Sidiropoulos, L. De Lathauwer, X, Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor Decomposition for Signal Processing and Machine Learning," IEEE Trans. on Signal Processing, vol. 65, no. 13, pp. 3551–3582, 2017.

[2] P. Tichavský and Z. Koldovský, "Weight Adjusted Tensor Method for Blind Separation of Underdetermined Mixtures of Nonstationary Sources," IEEE Trans. on Signal Processing, vol. 59, no. 3, pp. 1037–1047, 2011.

[3] A. Yeredor, A. Weiss, and A. J. Weiss, "High-Order Analysis of the Efficiency Gap for Maximum Likelihood Estimation in Nonlinear Gaussian Models," IEEE Trans. on Signal Processing, vol. 66, no. 18, pp. 4782–4795, 2018.

[4] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, Learning, and Kolmogorov Extension for Finite-Alphabet Random Vectors," IEEE Trans. on Signal Processing, vol. 60, no. 18, pp. 4854–4868, 2018.

[5] N. Kargas, N. D. Sidiropoulos, "Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm", Proceedings of Machine Learning Research, vol. 89, pp. 388-396, 2019.

[6] S. Hansen, T. Plantega, and T. G. Kolda, "Newton-based optimization for Kullback Leibler nonnegative tensor factorizations," Optimization Methods and Software, vol. 30, no. 5, pp. 1002–1029, 2015.

[7] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," SIAM J. Matrix Anal. Appl. vol. 33, no. 4, pp. 1272-1299, 2012.

[8] A. Cichocki and A. H. Phan, "Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations", IEICE Trans. on Fundamentals of Electronics, Communications and Computer Science, vol. 92-A, no. 3, pp 708–721, 2009.

[9] A. Cichocki, R. Zdunek, A. H. Phan and S.-I. Amari, "Nonnegative Matrix and Tensor Factorizations", Wiley, 2009.

[10] S. Kullback and R. A. Leibler, "On Information and Sufficiency", The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.