Abstract—We consider the problem of estimating the Probability Mass Function (PMF) of a discrete random vector (RV) from partial observations, namely when some elements in each observed realization may be missing. Since the PMF takes the form of a multi-way tensor, under certain model assumptions the problem becomes closely associated with tensor factorization. Indeed, in recent studies it was shown that a low-rank PMF tensor can be fully recovered (under some mild conditions) by applying a low-rank (approximate) joint factorization to all estimated joint PMFs of subsets of fixed cardinality larger than two (e.g., triplets). The joint factorization is based on a Least Squares (LS) fit to the estimated lower-order sub-tensors. In this letter we take a different estimation approach by fitting the partial factorization directly to the observed partial data in the sense of Kullback-Leibler divergence (KLD). Consequently, we avoid the need for particular selection and direct estimation of sub-tensors of a particular order, as we inherently apply proper weighting to all the available partial data. We show that our approach essentially attains the Maximum Likelihood estimate of the full PMF tensor (under the low-rank model) and therefore enjoys its well-known properties of consistency and asymptotic efficiency. In addition, based on the Bayesian model interpretation of the low-rank model, we propose an Estimation-Maximization (EM) based approach, which is computationally cheap per iteration. Simulation results demonstrate the advantages of our proposed KLD-based hybrid approach (combining alternating-directions minimization with EM) over LS fitting of sub-tensors.

Index Terms—Probability Mass Functions (PMF), Maximum Likelihood (ML), Coupled Tensor Factorization, Kullback-Leibler Divergence, Estimation-Maximization (EM).

I. INTRODUCTION

Estimation of the probability mass function (PMF) of a discrete random vector (RV) (whose elements take values in finite alphabets) from partial observations thereof is useful in many data analysis contexts in signal processing and machine learning, such as recommender systems, data completion, classification and clustering tasks. For example, the elements of the random vector may represent different features of individual subjects, and the ultimate goal is to infer the values of unobserved features when observing a subset of the features. If the full PMF of the vector is known, the Maximum A-Posteriori (MAP) estimates of the missing features would be readily available. In practice, however, the full PMF is rarely known and needs to be estimated during a “training” or “learning” stage. If only partly observed realizations are provided during this stage as well (namely, in each observed realization some or even most of the elements are missing), reasonable estimates of the full PMF would generally require an unfeasible amount of observed data, so as to properly cover all possible co-occurrences.

However, in recent work by Kargas et al. [4], it was shown that when the PMF can be represented by a low-rank tensor, such a representation can be interpreted as a naïve Bayes model. Moreover, the full PMF tensor can be recovered from knowledge of all of its sub-tensors of a fixed degree larger than 2 (e.g., triplets or quadruples) - and these can be more realistically estimated from partial observations. It may indeed be more realistic 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 (or “joint”) factorization to the empirically obtained sub-tensors, using 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. Another difficulty posed by the LS approach is that in a scenario of partial observations some combinations of observed elements (say certain triplets) may appear more frequently than others, rendering the estimates of the respective sub-tensors of the former more accurate than those of the latter. Consequently, such a situation would normally warrant some “proper weighting” scheme, but the LS criterion does not offer a “natural” way for properly weighting the different sub-tensors in the coupled factorization.

In this letter we take a different approach, seeking the Maximum Likelihood (ML) estimate of the PMF tensor, based on the partial observations and constrained by the low-rank model assumption. We show that while the resulting ML estimate can also be interpreted as solving a coupled factorization problem, the implied similarity criterion in this context is the Kullback-Leibler divergence (KLD, [12]), rather than the LS difference (note that KLD has also been recently used in some what related contexts in [5], [13]). Additionally, the coupled factorization is essentially applied directly to all of the observed data, thereby obviating the need for prior estimation of all sub-tensors of a particular degree, and offering implicit optimal weighting.
Two computational approaches are proposed: One follows the coupled factorization approach of [4] (with some important structural differences and using KLD rather than LS), which takes an iterative alternating-directions (AD) minimization strategy; the other is based on the Estimation-Maximization (EM, e.g., [6], [7]) approach, which is well-suited to the naive Bayes model implied by the low-rank assumption, and is computationally cheaper per iteration than the AD minimization.

II. PROBLEM FORMULATION

Let \( \mathbf{X} = [X_1, X_2, \ldots, X_N]^T \in \mathbb{R}^N \) be a discrete random vector with \( X_n \) taking discrete integer values in \([1, I_n]\) \((n = 1, \ldots, N)\). We denote its joint PMF tensor \( \mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N} \), where \( \mathbf{X}(i_1, i_2, \ldots, i_N) = \Pr\{X_1 = i_1, X_2 = i_2, \ldots, X_N = i_N\} \). We assume that \( \mathbf{X} \) admits a low-rank non-negative Canonical Polyadic Decomposition (CPD, e.g., [1]) with \( F \) components, namely there are \( N \) factor matrices \( A_1, A_2, \ldots, A_N \) \((A_n \in \mathbb{R}^{I_n \times F})\) and a “loading vector” \( \lambda \in \mathbb{R}^F \) such that

\[
\mathbf{X} = \sum_{f=1}^F \mathbf{A}_1(:, f) \circ \mathbf{A}_2(:, f) \circ \cdots \circ \mathbf{A}_N(:, f) = \lambda \circ [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N] \quad (1)
\]

where \( \circ \) denotes an outer product of vectors. All elements of \( \lambda \) are positive, all elements of \( \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N \) are non-negative and \( 1^T \lambda = 1, 1^T A_n = 1^T, n = 1, \ldots, N \). (1) denotes an outer product of vectors. All elements of \( \lambda \) are positive, all elements of \( \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N \) are non-negative and \( 1^T \lambda = 1, 1^T A_n = 1^T, n = 1, \ldots, N \). (1) denotes a RV obtained as a (possibly partial) observation of \( \mathbf{X} \), such that

\[
Y_n = \begin{cases} X_n & \text{w.p. } 1 - p \\ 0 & \text{w.p. } p \end{cases}, \quad n = 1, \ldots, N \quad (2)
\]

(independently for each \( n \) and independently of the values in \( Y_1, Y_2, \ldots, Y_N \)).

The goal is to estimate \( \mathbf{X} \) from a finite number \( T \) of independent, identically distributed (i.i.d.) realizations of \( Y \).

III. ESTIMATION OF THE PMF

Naturally, if all elements of \( \mathbf{X} \) are observed together (in all possible combinations) for a sufficiently large number of times, an empirical (histogram-like) estimate of the entire PMF tensor can be obtained, followed by a low-rank approximation thereof. However, since the probability of observing the full vector is \((1 - p)^N\), even small values of \( p \) can make such observations extremely rare. Combined with the exponential dependence of the total number of combinations (bins in the full histogram) on the tensor’s dimensions, direct empirical estimation of the full PMF tensor \( \mathbf{X} \) is practically hopeless.

Fortunately, however, it is shown in [4] that even in the absence of a direct empirical estimate \( \hat{\mathbf{X}} \) of the full tensor, estimates of the factors can be obtained from empirical estimates of the joint PMF in triplets, denoted \( \hat{\mathbf{X}}_{jkl} \), for a sufficient number of combinations of \((j, k, \ell)\) \((j < k < \ell)\). Note that the probability of observing a particular triplet (say \( j\)-th, \( k\)-th and \( \ell\)-th elements of \( \mathbf{X} \)) is merely \((1 - p)^3\), which is much larger than \((1 - p)^N\), especially for \( N \gg 3 \) and/or for moderate to high values of \( p \).

It is proposed in [4] to jointly factorize the triplet tensors by minimizing the LS criterion

\[
\min_{\{A_n\}_{n=1}^N} \sum_{j \geq j_1 \geq \cdots \geq j_{k+\ell}=k} \sum_{f=1}^F \sum_{i,j,k} \frac{1}{2} \left\| \hat{X}_{jkl} - [\lambda, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_\ell] \right\|_F^2 \quad (3)
\]

subject to: \( \lambda > 0, 1^T \lambda = 1 \). (3) where \( \| \cdot \|_F^2 \) denotes the squared Frobenius norm (sum of all squared elements of the enclosed tensor).

As already mentioned, while this approach is conceptually and computationally appealing, it may often lead to compromised estimates of the factors, due to the use of the LS criterion (which is statistically less appropriate in this context), and due to the restriction to only use triplets. In the following subsection we propose a ML approach for estimating the factors, which would not only be more appropriate from a statistical point of view, but would also properly exploit all of the observed data (regardless of co-occurrences in triplets), and would be asymptotically optimal in the sense of attaining the Cramér-Rao Lower Bound on the mean square error (MSE) in estimating the loading vector and the factor matrices.

IV. A MAXIMUM LIKELIHOOD APPROACH

Assume a realization \( y \) of \( Y \), let \( B \in [0, N] \) denote the number of non-zero elements of \( y \), and let \( n_1, \ldots, n_B \) denote their indices. Then the probability of observing \( y \) is given by

\[
\Pr\{y\} = \begin{cases} p^{N-B}(1-p)^B & \Pr\{X_{n_1} = y_{n_1}, \ldots, X_{n_B} = y_{n_B}\} \\ 0 & \text{otherwise} \end{cases} \quad (4)
\]

where \( p = q \cdot \sum_{f=1}^F \lambda_f A_{n_1} (y_{n_1}, f) \cdots A_{n_B} (y_{n_B}, f) \),

\[
\Pr\{y\} = p (1 - p)^3 \Pr\{X_1 = 3, X_2 = 4, X_4 = 2\} \quad (5)
\]

where the last transition is due to \( 1^T A_3 = 1^T \). We can therefore write

\[
\log \Pr\{y\} = c + \sum_{f=1}^F \lambda_f \prod_{b=1}^B A_{n_b} (y_{n_b}, f), \quad (6)
\]
where \( c \triangleq \log q \) is an irrelevant constant. Given \( T \) i.i.d. observations \( y[1], \ldots, y[T] \), their joint log-likelihood is therefore given (up to an irrelevant constant) by
\[
\log Pr\{y[1], \ldots, y[T]\} = \sum_{t=1}^{T} \sum_{f=1}^{F} \sum_{b=1}^{B[t]} \lambda_f \prod_{n=1}^{A_n[t]} \left( y_{n[a][t][f]} \right),
\]
(7)
where \( B[t] \) is the number of non-zero (observed) elements in \( y[t] \), and \( n_1[t], \ldots, n_{f-1}[t] \) are their indices (into \( y[t] \)).

Our optimization problem (for ML estimation of \( A_1, \ldots, A_N \) and \( \lambda \)) therefore takes the form:
\[
\min_{\{A_n\}_{n=1}^{N}, \lambda} \left( -\sum_{t=1}^{T} \sum_{f=1}^{F} \sum_{b=1}^{B[t]} \lambda_f \prod_{n=1}^{A_n[t]} \left( y_{n[a][t][f]} \right) \right) \]
subject to: \( \lambda > 0, \ 1^T \lambda = 1 \)
\[
A_n \geq 0, \ 1^T A_n = 1^T, \ n = 1, \ldots, N. \quad (8)
\]

A possible interpretation is the following. Define a “pinning tensor” \( \mathcal{E}_{i_1, \ldots, i_N} \) as an all-zeros tensor with a 1 as its \( (i_1, \ldots, i_N) \)-th element. Note that each \( y[t] \) can be associated with a degenerate PMF, denoted \( \mathcal{P}[t] \) - a sub-tensor of the pinning-tensor, prescribed by indices of the observed values in \( y[t] \), ascribing probability 1 to the observed \( y[t] \) and 0 to all other combinations. The log-likelihood inside the outer sum in (7) can be thought of as the KLD between \( \mathcal{P}[t] \) and \( \mathcal{X} \). Thus, the minimization problem (8) essentially attempts to find the joint coupled CPD of \( T \) (sub-“pinning-tensors” \( \mathcal{P}[t] \), where the implied distance measure is the sum of KLDs between all \( \mathcal{P}[t] \) and \( \mathcal{X} \). In other words, this criterion generalizes (3) by incorporating all types of sub-tensors (not only triplets) and by substituting the LS criterion with the KLD.

This is obviously a non-convex optimization problem, but taking an AD approach, we can minimize w.r.t. \( \lambda \) and to each \( A_n \) separately, and each of these internal minimization problems is convex: Assuming that \( \lambda \) and all \( A_n \), except for one \( A_m \) are fixed, the minimization w.r.t. \( A_m \) takes the form:
\[
\min_{A_m} \left( -\sum_{t:y_m[t] \neq 0} \log \left( h_{m[t]}(A_m(y_m[t],:)) \right)^T \right) \]
subject to: \( A_m \geq 0, \ 1^T A_m = 1^T, \) \( (9) \)
where, adopting Matlab notations, \( A(i,:) \) denotes the \( i \)-th row of \( A \), and where \( h_m[t] \in \mathbb{R}^{F} \) is given by
\[
h_m[t] = \text{Diag}(\lambda) \cdot \prod_{b=1, n_b[t] \neq m}^{B[t]} (A_{n_b[t]}(y_{n_b[t][t][:]}(:,)))^T \quad (10)
\]
(here \( \text{Diag}(\lambda) \) creates a diagonal matrix from \( \lambda \), and \( \prod \) denotes repeated Hadamard (elementwise) products). Likewise, when minimizing w.r.t. \( \lambda \) with all \( A_n \) fixed we have
\[
\min_{\lambda} \left( -\sum_{t:y_m[t] \neq 0} \log \left( h_{m[t]}(\lambda) \right) \right) \]
subject to: \( \lambda > 0, \ 1^T \lambda = 1, \) \( (11) \)
where \( h[t] \in \mathbb{R}^{F} \) is given by
\[
h[t] = \prod_{b=1}^{B[t]} \left( A_{n_b[t]}(y_{n_b[t][t][:]}(:,)) \right)^T. \quad (12)
\]
Each of the internal (convex) minimizations can be solved, e.g., using reparameterization techniques outlined in [13].

V. USING EM

Another possible approach for the same optimization problem can be based on applying the naive Bayes model implied by the low-rank assumption (see [4] for more details on this important insight). This model conveniently lends itself to the EM algorithm as follows. Let us define the “complete data” \( z[t] = [y[t], s[t]]^T \in \mathbb{R}^{N+1} \) consisting of the observed vectors, augmented by the unknown latent Bayesian “state” \( s[t] \in \{1, 2, \ldots, F\} \) from which \( x[t] \) was drawn according to the naive Bayes model. The EM algorithm then consists of iterating between the following stages:

- **E-Step**: Denoting the value of the distribution parameters (loading vector and all factor matrices) at the beginning of the iteration as \( \theta' \), compute
\[
Q(\theta, \theta') \triangleq E[\log Pr(z[1], \ldots, z[T]; \theta)|y[1], \ldots, y[T]; \theta'] \quad (13)
\]
- **M-Step**: Update \( \theta \) by maximizing \( Q(\theta, \theta') \):
\[
\text{new } \theta' \leftarrow \text{arg max } Q(\theta, \theta') \quad (14)
\]
Considering the E-Step first, we notice that
\[
Pr\{z[t]; \theta\} = \lambda_{s[t]} \prod_{b=1}^{B[t]} A_{n_b[t]}(y_{n_b[t][t][s[t]]}), \quad (15)
\]
so that
\[
\log Pr\{z[1], \ldots, z[T]; \theta\} = \sum_{t=1}^{T} \left( \log \lambda_{s[t]} + \sum_{b=1}^{B[t]} \log A_{n_b[t]}(y_{n_b[t][t][s[t]]}) \right). \quad (16)
\]
In order to take the conditional mean in (13) we only need to average over all \( F \) possible values of each \( s[t] \) (since all \( y[t] \) are given by the condition), namely
\[
E[\log Pr(z[1], \ldots, z[T]; \theta)|y[1], \ldots, y[T]; \theta'] = \sum_{t=1}^{T} \sum_{f=1}^{F} \text{Pr}\{s[t] = f | y[t]; \theta'\}
\]
\[
\cdot \left[ \log \lambda_{s[t]} + \sum_{b=1}^{B[t]} \log A_{n_b[t]}(y_{n_b[t][t][s[t]]}) \right], \quad (17)
\]
Further denoting \( c_{t,f}(\theta') \triangleq \text{Pr}\{s[t] = f | y[t]; \theta'\} \), we have
\[
c_{t,f}(\theta') = \frac{\text{Pr}\{y[t] | s[t] = f \} \cdot \text{Pr}\{s[t] = f \}}{\sum_{g=1}^{F} \text{Pr}\{y[t] | s[t] = g \} \cdot \text{Pr}\{s[t] = g \}}, \quad (18)
\]
which are all easily calculated from the elements of \( \theta' \). Substituting these coefficients into (17) and (13) and swapping the
summation order, we get
\[
Q(\theta, \theta') = \sum_{f=1}^{F} \sum_{t=1}^{T} c_{t,f}(\theta') \cdot \left( \log \lambda_{i[t]} + \sum_{b=1}^{\beta[t]} \log A_{n_i[t]}(y_{n_i[t]}, s[t]) \right).
\]

Further defining \( C_f' \triangleq \sum_{t=1}^{T} c_{t,f}(\theta') \) and \( K_f'(n, i) \triangleq \sum_{c_v y_{n_i[t]} = c} c_{t,f}(\theta') \) for \( n = 1, ..., N, \ i = 1, ..., I_n \), we can rewrite (2) as
\[
Q(\theta, \theta') = \sum_{f=1}^{F} \left( C_f' \log \lambda_f + \sum_{n=1}^{N} \sum_{i=1}^{I_n} K_f'(n, i) \log A_n(i, f) \right).
\]

In the M-Step we maximize \( Q(\theta, \theta') \) w.r.t. \( \theta \), namely w.r.t. the loading vector and all factor matrices, subject to the constraints \( \lambda > 0, 1^T \lambda = 1 \) and \( A_n \geq 0, 1^T A_n = 1^T \) \( n = 1, ..., N \). Fortunately, all the elements of \( \theta \) in (20) can be decoupled inside the outer sum, so we only need to maximize
\[
\max_{\lambda} \sum_{f=1}^{F} C_f' \log \lambda_f \quad \text{s.t.} \quad \sum_{f=1}^{F} \lambda_f = 1
\]
\[
\max_{A_n(:, f)} \sum_{i=1}^{I_n} K_f'(n, i) \log A_n(i, f) \quad \text{s.t.} \quad \sum_{i=1}^{I_n} A_n(i, f) = 1
\]
for all \( n, f \) (where the non-negativity constraint is implicitly satisfied). The maximizing solutions can be easily shown (using Lagrange multipliers) to be given by:
\[
\lambda_f = \frac{C_f'}{\sum_{g=1}^{F} C_g'}, \quad A_n(i, f) = \frac{K_f'(n, i)}{\sum_{j=1}^{I_n} K_f'(n, j)}.
\]

The main advantage of EM over the AD KLD minimization is the significantly lower computation load per iteration ensuing from the simple closed-form updates (21) of the entire vector \( \theta \) in each iteration. However, a main disadvantage of EM is its relatively slow convergence and its high sensitivity to initialization. We therefore chose (for our simulation experiment below) to use a hybrid scheme: following random initialization, we first run a few “sweeps” of the (computationally extensive) AD scheme, and then, having hopefully reached the vicinity of the global maximum, switch to the (computationally cheaper) EM for the final refinement. Choosing the number of initial AD iterations to run poses a trade-off between the overall computation load and the probability of success of EM. For our simulation experiment we chose (after testing a few options) to run a fixed number of 20 AD iterations (in each trial) before switching to EM.

VI. SIMULATION RESULTS

In order to demonstrate the advantages of our proposed ML-based estimates, we created a medium-scale simulation similar to the experimental setup in [4], with \( N = 5 \) random variables, each taking one of \( I_n = 10 \) discrete values (for \( n = 1, ..., N \)), hence our full PMF tensor has \( 10^{10} \) elements. We generated the true PMF tensor so as to have rank \( F = 5 \), drawing all elements of the factor matrices \( A_n \in \mathbb{R}^{10 \times 5} \) and of the loading vector \( \lambda \in \mathbb{R}^5 \) independently from a \( U(0, 1) \) distribution, and normalizing as per the probability simplex constraints. The PMF tensor was drawn once and used for generating the random data for all simulation trials as follows.

In compliance with the naïve Bayes model implied by the low-rank structure [4], in each trial \( t = 1, ..., T \), we first draw the latent state \( s[t] \) according to \( \lambda \), and then draw the 5 elements of \( x[t] \), such that each \( x[n][t] \) is drawn independently according to \( A_n(:, s[t]) \). The observations \( y[t] \) are then obtained by randomly and independently “hiding” (zeroing-out) elements of \( x[t] \) with outage probability \( p = 0.25 \) (see (2)).

Fig. 1 shows the resulting error in estimating the full PMF tensor from the observed data vs. the observations length \( T \). The error is presented both in terms of the Mean Relative Error (MRE), which is the normalized Frobenius norm of the difference between the true and estimated tensors [4]; and in terms of the KLD between the true and estimated tensors. The error of the proposed ML approach (as obtained by using random initialization, applying 20 iterations of the AD minimization and then switching to EM for final convergence) is compared to the triplets-based LS factorization [4], as well as to the (unattainable) lower bound of the “Oracle” ML estimate, which is based on observing the true latent state \( s[t] \) (in addition to \( y[t] \)) in each observation. Our ML-based estimate is seen to outperform the triplets-LS-based estimate in terms of both error-measures (except for KLD for the very short observation lengths).
REFERENCES


