Recovering Joint PMF from Pairwise Marginals

Shahana Ibrahim
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, USA
ibrahish@oregonstate.edu

Xiao Fu
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, USA
xiao.fu@oregonstate.edu

***Abstract*—To overcome the curse of dimensionality in joint probability learning, recent work has proposed to recover the joint probability mass function (PMF) of an arbitrary number of random variables (RVs) from three-dimensional marginals, exploiting the uniqueness of tensor decomposition and the (unknown) dependence among the RVs. Nonetheless, accurately estimating three-dimensional marginals is still costly in terms of sample complexity. Tensor decomposition also poses a computationally intensive optimization problem. This work puts forth a new framework that learns the joint PMF using pairwise marginals that are relatively easy to acquire. The method is built upon *nonnegative matrix factorization* (NMF) theory, and features a Gram–Schmidt-like economical algorithm that works provably well under realistic conditions. Theoretical analysis of a recently proposed *expectation maximization* (EM) algorithm for joint PMF recovery is also presented. In particular, the EM algorithm is shown to provably improve upon the proposed pairwise marginal-based approach. Synthetic and real-data experiments are employed to showcase the effectiveness of the proposed approach.**

***Index Terms*—joint probability learning, nonnegative matrix factorization, probability tensors, two-dimensional marginals**

I. Introduction

Many learning and inference tasks in high-dimensional statistics boil down to estimating/approximating the joint probability of a set of random variables (RVs). However, in the high-dimensional regime, directly estimating the joint probability via "structure-free" methods such as sample averaging is considered not viable—due to the need of a huge amount of data. Many workarounds, e.g., linear estimators, kernels, and neural networks, have been proposed for combating this curse of dimensionality [1]. However, the fundamental challenge of estimating the joint probability from limited data remains.

Very recently, Kargas *et al.* proposed a new framework for *blindly* estimating the *joint probability mass function* (PMF) of N discrete finite-alphabet RVs [2] by modelling the N-dimensional joint PMF as an Nth-order tensor. The work in [2] shows that if the RVs are "reasonably dependent", the joint PMF can be recovered via jointly decomposing the three-dimensional marginal PMFs (which are third-order tensors). The approach does not use any *a priori* structural information of the RVs, and the recoverability of the joint PMF is provably guaranteed [2]. The work in [2] has shown promising results, but a couple of major challenges remain. First, estimating third-order marginals accurately is not a trivial task, since real-life high-dimensional data are often very sparse. Second, computing coupled third-order tensor decomposition poses a challenging and resource-consuming optimization problem, which in general does not have known polynomial-time solvers.

Instead of working with a large number of third-order marginals, a recent work in [3] offers an *expectation maximization* (EM) algorithm that directly estimates the latent factors of the Nth-order probabilistic tensor. The EM algorithm is well-motivated, since it tackles the the *maximum likelihood estimator* (MLE). In addition, it admits simple and economical updates, and thus is quite scalable. However, unlike the tensor-based approach in [2], it is unclear if the recoverability of the joint PMF can be guaranteed using EM and the MLE. In addition, since the ML estimation problem is nonconvex, convergence properties of the EM algorithm is unclear.

In this work, we propose a new framework that offers provable recoverability of the joint PMF and at the same time enjoys low sample and computational complexities. To be specific, we propose an approach that utilizes only pairwise marginals to recover the joint PMF of an arbitrary number of discrete finite-alphabet RVs. This way, the sample complexity is substantially reduced relative to the three-dimensional marginal-based approach in [2]. We propose a pragmatic and easy-to-implement joint PMF estimation procedure, which is based on performing a simple and scalable Gram–Schmidt (GS)-like algorithm (namely, the *successive projection algorithm* [4]) on a carefully constructed "virtual nonnegative matrix factorization (NMF)" model. We also show that the EM algorithm in [3] can provably recover the joint PMF tensor if initialized properly, e.g., using our GS-like algorithm. We illustrate the effectiveness of the proposed approach using a number of synthetic and real-data experiments.

II. Problem Statement

Consider a set of discrete and finite-alphabet RVs, i.e., $Z_1, \ldots, Z_N$. We will use $\mathsf{Pr}(i_1, \ldots, i_N)$ as the shorthand notation to represent $\mathsf{Pr}(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)})$ in the sequel, where $\{z_n^{(1)}, \ldots, z_n^{(I_n)}\}$ denotes the alphabet of Z_n. The work in [2] shows a connection between joint PMFs and low-rank tensors under the *canonical polyadic decompo-*

This work is supported by the National Science Foundation under NSF-ECCS 1608961 and NSF-ECCS 1808159 and the Army Research Office under ARO W911NF-19-1-0247 and W911NF-19-1-0407.

sition (CPD) model. To be specific, if an N-th order tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ has CP rank F, it can be written as:

$$\underline{\boldsymbol{X}}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \boldsymbol{\lambda}(f) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n, f), \qquad (1)$$

where $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$ is called the mode-n latent factor. In the above, $\boldsymbol{\lambda} = [\boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(F)]^T$ with $\|\boldsymbol{\lambda}\|_0 = F$ is employed to "absorb" the norms of columns. The work in [2] shows that *any* joint PMF admits a naive Bayes (NB) model *representation*. It follows that the joint PMF of $\{Z_n\}_{n=1}^N$ can always be decomposed as

$$\mathsf{Pr}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \mathsf{Pr}(f) \prod_{n=1}^{N} \mathsf{Pr}(i_n|f), \qquad (2)$$

where $\mathsf{Pr}(f) := \mathsf{Pr}(H = f)$ is the prior distribution of a latent variable H and $\mathsf{Pr}(i_n|f) := \mathsf{Pr}(Z_n = z_n^{(i_n)}|H = f)$ are the conditional distributions. Consequently, one can represent any joint PMF as an Nth-order tensor by letting $\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \mathsf{Pr}(i_1, \ldots, i_N)$ and $\boldsymbol{A}_n(i_n, f) = \mathsf{Pr}(i_n|f)$, $\boldsymbol{\lambda}(f) = \mathsf{Pr}(f)$ (see more details in [2]).

The approach in [2] showed that if one has access to three-dimesional marginals, i.e, $\mathsf{Pr}(i_j, i_k, i_\ell)$ for different j, k, ℓ, then the joint PMF can be provably recovered through a latent factor-coupled tensor decomposition approach—if the tensor rank F is small (meaning that if the N RVs are reasonably dependent). However, estimating $\mathsf{Pr}(i_j, i_k, i_\ell)$ is still not easy, since one needs many co-realizations of three RVs. In addition, jointly decomposing a large number of third-order tensors poses a challenging optimization problem, whose global optimality is not guaranteed. The more recent work by Yeredor and Haardt in [3] takes an ML perspective and directly estimates the model parameters in (2) using an EM algorithm. The algorithm admits economical updates, and is effective if carefully initialized, e.g., using a coupled tensor decomposition algorithm. Nonetheless, the ML formulation's recoverability properties are unclear. Since the ML estimator is a nonconvex optimization criterion, it is also unclear if the EM algorithm converges to the desired latent factors or not.

III. Proposed Approach

Our idea is to utilize pairwise marginals instead of the three-dimensional marginals. Under the naive Bayes model, the pairwise marginals can be expressed as $\mathsf{Pr}(i_j, i_k) = \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_j|f)\mathsf{Pr}(i_k|f)$, or, equivalently,

$$\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top, \ \boldsymbol{X}_{jk}(i_j, i_k) = \mathsf{Pr}(i_j, i_k),$$

where $\boldsymbol{D}(\boldsymbol{\lambda}) = \mathrm{Diag}(\boldsymbol{\lambda})$ and $\{\boldsymbol{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ are defined as before. It is readily seen that if $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$ can be identified from the marginals, $\mathsf{Pr}(i_1, \ldots, i_N)$ can be recovered by (2).

In practice, the pairwise marginals $\boldsymbol{X}_{jk}$'s are estimated from realizations of the joint PMF. Consider a set of realizations (data samples) of $\mathsf{Pr}(Z_1, \ldots, Z_N)$, denoted as $\{\boldsymbol{d}_s \in \mathbb{R}^N\}_{s=1}^S$. Assuming that there is no missing observations, the following sample averaging estimator can employed:

$$\widehat{\boldsymbol{X}}_{jk}(i_j, i_k) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left[\boldsymbol{d}_s(j) = z_j^{(i_j)}, \boldsymbol{d}_s(k) = z_k^{(i_k)}\right],$$

where $\boldsymbol{d}_s(n)$ denotes the realization of Z_n in the s-th data sample and $\mathbb{I}[E] = 1$ if the event E happens and $\mathbb{I}[E] = 0$ otherwise. Note that using such sample averaging schemes, the pair-wise marginals can be estimated to a much higher accuracy compared to the three-dimensional ones, under the same amount of data sample; see, e.g., [5].

However, *recoverability* of the joint PMF using pairwise marginals is nontrivial to establish. The marginal distributions, i.e., $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top$ for all j, k, are matrices, and low-rank matrix decomposition is in general *nonunique*—while uniqueness of the employed factorization model was the key stepping stone in [2] to establish recoverability for the joint PMF from the three-dimensional marginals. A natural thought to handle the identifiability problem would be employing certain NMF tools [6], [7], since the latent factors are all nonnegative, per their physical interpretations. However, the identifiability of NMF models holds only if $F \leq \min\{I_j, I_k\}$ (and preferably $F \ll \min\{I_j, I_k\}$). The pairs $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top \in \mathbb{R}^{I_j \times I_k}$ inherit the inner dimension F (i.e., the column dimension of $\boldsymbol{A}_j$) from the joint PMF of all the variables, which is the tensor rank of an Nth-order tensor. Note that the tensor rank F could be much larger than the I_j's [8]. Hence, one may not directly use the available NMF uniqueness results on individual $\boldsymbol{X}_{jk}$'s to argue for joint PMF recoverability.

A. A Virtual NMF-based Approach

To see how we approach these challenges, consider a splitting of the indices of the N variables, i.e., $\mathcal{S}_1 = \{\ell_1, \ldots, \ell_M\}$ and $\mathcal{S}_2 = \{\ell_{M+1}, \ldots, \ell_N\}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, \ldots, N\}$ and $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, where $\emptyset$ denotes the empty set. Then, we construct the following matrix:

$$\widetilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_{\ell_1 \ell_{M+1}} & \ldots & \boldsymbol{X}_{\ell_1 \ell_N} \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_{\ell_M \ell_{M+1}} & \ldots & \boldsymbol{X}_{\ell_M \ell_N} \end{bmatrix} = \boldsymbol{W}\boldsymbol{H}^\top, \qquad (3)$$

where $\boldsymbol{W} = [\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_M}^\top]^\top \in \mathbb{R}^{MI \times F}$ and $\boldsymbol{H} = \boldsymbol{D}[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top]^\top \in \mathbb{R}^{(N-M)I \times F}$, respectively, if $I_1 = \ldots = I_N = I$. Note that $\boldsymbol{W}$ and $\boldsymbol{H}$ are both non-negative and thus (3) is an NMF model. By constructing $\widetilde{\boldsymbol{X}}$ such that $F \leq \min\{MI, (N-M)I\}$, the identifiability of $\boldsymbol{W}$ and $\boldsymbol{H}$ can be established using certain NMF tools. One celebrated condition is the so-called *separability* [6]:

Definition 1 *(Separability) If $\boldsymbol{H} \geq \boldsymbol{0}$, and $\boldsymbol{\Lambda} = \{l_1, \ldots, l_F\}$ such that $\boldsymbol{H}(\boldsymbol{\Lambda}, :) = \boldsymbol{\Sigma}$ holds, where $\boldsymbol{\Sigma} = \mathrm{Diag}(\alpha_1, \ldots, \alpha_F)$ and $\alpha_f > 0$, then, $\boldsymbol{H}$ satisfies the separability condition.*

Under the exact separability condition on $\boldsymbol{H}$, we have $\boldsymbol{H}(\boldsymbol{\Lambda}, :) = \boldsymbol{\Sigma} = \mathrm{Diag}(\alpha_1, \ldots, \alpha_F)$ and $\boldsymbol{W}\boldsymbol{\Sigma} = \boldsymbol{X}(\boldsymbol{\Lambda}, :)$. Hence, the coupled NMF task boils down to identifying the index set $\boldsymbol{\Lambda}$. The *successive projection algorithm* (SPA) from the NMF literature [4] can be employed for this purpose. Notably, this algorithm admits Gram–Schmidt-like economical and scalable updates and is provably robust to noise.

Once $\boldsymbol{W}$ is identified, we can recover $\boldsymbol{A}_{\ell_n} \in \mathbb{R}^{I_{\ell_n} \times F}$ for $\ell_n \in \mathcal{S}_1$ up to identical column permutations, by extracting the corresponding rows of $\boldsymbol{W}$. Unlike general NMF models, since we know every column of $\boldsymbol{A}_n$ is a conditional PMF, there is no scaling ambiguity. The $\boldsymbol{H}$ matrix can be estimated using (constrained) least squares, and $\boldsymbol{A}_{\ell_n}$ for $\ell_n \in \mathcal{S}_2$ can then be extracted in the similar way. Denote (3) as $\widetilde{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{D}(\boldsymbol{\lambda})\widetilde{\boldsymbol{H}}^\top$, where $\widetilde{\boldsymbol{H}} = \left[\boldsymbol{A}_{\ell_{M+1}}^\top \cdots \boldsymbol{A}_{\ell_N}^\top\right]^\top$. Then, the PMF of the latent variable can be estimated via $\boldsymbol{\lambda} = (\widetilde{\boldsymbol{H}} \odot \boldsymbol{W})^\dagger \mathrm{vec}(\widetilde{\boldsymbol{X}})$ where we have used the fact that the Khatri-Rao product $\widetilde{\boldsymbol{H}} \odot \boldsymbol{W}$ has full column rank since both $\boldsymbol{W}$ and $\widetilde{\boldsymbol{H}}$ have full column rank. Note that the permutation ambiguity across $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$ are identical. Hence, the existence of column permutations does not affect the "assembling" of $\mathsf{Pr}(i_n|f)$ and $\mathsf{Pr}(f)$ to recover $\mathsf{Pr}(i_1, \ldots, i_N)$. We refer to this procedure as *coupled NMF via SPA* (CNMF-SPA); see Algorithm 1.

Algorithm 1: CNMF-SPA

input : data samples $\{\boldsymbol{d}_s\}_{s=1}^S$ and M
estimate second order statistics $\widehat{\boldsymbol{X}}_{jk}$;
split $\{1, \ldots, N\}$ into $\mathcal{S}_1 = \{1, \ldots, M\}$ and $\mathcal{S}_2 = \{M+1, \ldots, N\}$;
Construct $\widetilde{\boldsymbol{X}}$;
Estimate $\widehat{\boldsymbol{W}}$ using the **SPA algorithm** [4] to select $\boldsymbol{\Lambda}$;
for $n = 1$ **to** M **do**
  $\widehat{\boldsymbol{A}}_n \leftarrow \widehat{\boldsymbol{W}}((n-1)I_n + 1 : nI_n, :)$;
  normalize columns of $\widehat{\boldsymbol{A}}_n$ with respect to ℓ_1 norm;
end
$\widehat{\boldsymbol{H}} \leftarrow \arg\min_{\boldsymbol{H} \geq 0} \|\widetilde{\boldsymbol{X}} - \widehat{\boldsymbol{W}}\boldsymbol{H}^\top\|_{\mathrm{F}}^2$;
for $n = M+1$ **to** N **do**
  $\widehat{\boldsymbol{A}}_n \leftarrow \widehat{\boldsymbol{H}}((n-1)I_n + 1 : nI_n, :)$;
  normalize columns of $\widehat{\boldsymbol{A}}_n$ with respect to ℓ_1 norm;
end
$\widetilde{\boldsymbol{W}}^\top \leftarrow \left[\widehat{\boldsymbol{A}}_1^\top, \ldots, \widehat{\boldsymbol{A}}_M^\top\right]$;
$\widetilde{\boldsymbol{H}}^\top \leftarrow \left[\widehat{\boldsymbol{A}}_{M+1}^\top \cdots \widehat{\boldsymbol{A}}_N^\top\right]^\top$;
$\widehat{\boldsymbol{\lambda}} \leftarrow (\widetilde{\boldsymbol{H}} \odot \widetilde{\boldsymbol{W}})^\dagger \mathrm{vec}(\widetilde{\boldsymbol{X}})$;
output: estimates $\{\widehat{\boldsymbol{A}}_n\}_{n=1}^N$, $\widehat{\boldsymbol{\lambda}}$.

B. Performance Analysis

The CNMF-SPA procedure looks simple, but several caveats exist. First, SPA only works if one can construct $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\boldsymbol{H}$ in (3) satisfies the separability condition. Testing all combinations of $\mathcal{S}_1$ and $\mathcal{S}_2$ gives a rise to a combinitorial problem, which is apparently impossible. Second, the pairwise marginals $\boldsymbol{X}_{ij}$ are estimated through a finite number of data samples, and typically there are many missing values in different data samples (i.e., not all the realizations of $Z_1, \ldots, Z_n$ are observed in any $\boldsymbol{d}_s$)—which both make the estimated pairwise marginals very noisy. It is unclear how the performance of CNMF-SPA is affected.

In practice, we observe that using the "naive" construction $\mathcal{S}_1 = \{1, \ldots, M\}$ and $\mathcal{S}_2 = \{M+1, \ldots, N\}$ seems to work reasonably well, even if the number of samples is finite and many missing values exist. To understand such effectiveness, we assume that S realizations of $\mathsf{Pr}(Z_1, \ldots, Z_N)$ are available. In each realization, every variable is observed with probability p—which determines how much is missing in the dataset. For simplicity, we also assume that $I_n = I$ for all n and utilize the following generative model for the nonnegative $\boldsymbol{A}_m$'s:

Assumption 1 *Assume that the rows of $\boldsymbol{A}_m$'s are generated from the $(F-1)$-probability simplex uniformly at random and then positively scaled, so that $\mathbf{1}^\top \boldsymbol{A}_m = 1$ is respected.*

Under the above settings, we show that the following holds:

Theorem 1 *Assume that $\|\widehat{\boldsymbol{X}}_{ij}(:, q)\|_1 \geq \eta > 0$ for any q, i, j. Also, assume that $M \geq F/I$, $p \geq \left(\frac{8}{S}\log(4/\delta)\right)^{1/2}$,*

$$S = \Omega\left(\frac{M^2 I \log(1/\delta)}{\sigma_{\max}^2(\boldsymbol{W})\eta^2\varepsilon^2 p^2}\right),$$
$$N = M + \Omega\left(\frac{\varepsilon^{-2(F-1)}}{IF}\log\left(\frac{F}{\delta}\right)\right),$$

where $\varepsilon = \frac{M\min\left(\frac{1}{2\sqrt{F-1}}, \frac{1}{4}\right)}{2\kappa(\boldsymbol{W})(1+80\kappa^2(\boldsymbol{W}))}$. Then, under the defined S, p and Assumption 1, CNMF-SPA outputs $\widehat{\boldsymbol{A}}_m$'s such that

$$\min_{\boldsymbol{\Pi}:\ \textit{permuation}} \|\widehat{\boldsymbol{A}}_m\boldsymbol{\Pi} - \boldsymbol{A}_m\|_2 = O\left(\kappa^2(\boldsymbol{W})\sqrt{F}\zeta\right) \quad (4)$$

for $m \in \mathcal{S}_1$ with a probability greater than or equal to $1-\delta$, where $\zeta = \max(\sigma_{\max}(\boldsymbol{W})\varepsilon, {}^M\sqrt{I\log(1/\delta)}/\eta p\sqrt{S})$.

The proof can be found in a long version of the paper in the appendix. Note that if $\boldsymbol{A}_m$ for all $m \in \mathcal{S}_1$ can be accurately estimated, the estimation accuracy of $\boldsymbol{A}_n$ for all $n \in \mathcal{S}_2$ and $\boldsymbol{\lambda}$ can also be guaranteed and quantified, following standard sensitivity analyses of least squares. We leave this part out of the work for conciseness.

Remark 1 *Theorem 1 is not entirely surprising. The insight behind is to model the finite sample-induced noise and the violation of separability as combined virtual noise, and then utilize the robustness of SPA [4]. The challenge lies in quantifying this virtual noise, for which we leverage concentration theorems and Assumption 1. We would like to remark that Assumption 1 is a working assumption for us to understand the effectiveness of CNMF-SPA under the naive $\mathcal{S}_1$ and $\mathcal{S}_2$. The key fact is that when $|\mathcal{S}_2|$ grows, $\boldsymbol{H}$ in (3) has a good chance to attain the separability condition approximately [9] under Assumption 1. In principle, this fact holds if the rows of $\boldsymbol{A}_m$ are drawn from any joint continuous distribution, but using the uniform distribution assumed in Assumption 1 helps simplify the analysis.*

C. Refinement via EM and Optimality Guarantees

In [3], Yeredor and Haardt proposed an EM algorithm to handle the ML estimator for the naive Bayes model in (2). Using MLE exhibits promising performance for joint PMF recovery—after all the MLE is a "gold standard" for statistical learning. However, the EM algorithm often converges to undesired solutions, if randomly initialized. This is perhaps due to the nonconvex nature of the MLE problem. As observed

in [3], using good initialization may improve the performance of EM. However, quantification for the required quality of the initial estimates has been elusive.

In this work, we offer theoretical supports for the EM algorithm in [3]. To be specific, we show that, with good initial estimations for $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$ (e.g., those output by `CNMF-SPA`), the EM algorithm improves the solution towards the groud truth. To proceed, we make the following assumption:

Assumption 2 *Define $\overline{D}_1$ and $\overline{D}_2$ as follows:*

$$\overline{D}_1 = \min_{f \neq f'} \frac{1}{N} \sum_{n=1}^{N} p \mathbb{D}_{\mathrm{KL}}(\boldsymbol{A}_n(:,f), \boldsymbol{A}_n(:,f')),$$
$$\overline{D}_2 = \frac{2}{N} \min_{f \neq f'} \log(\boldsymbol{\lambda}(f)/\boldsymbol{\lambda}(f')).$$

and $\overline{D} = (\overline{D}_1 + \overline{D}_2)/2$. Assume that $\boldsymbol{A}_n, \boldsymbol{\lambda}$ and the initial estimates $\widehat{\boldsymbol{A}}_n^0, \widehat{\boldsymbol{\lambda}}^0$ satisfy $|\widehat{\boldsymbol{A}}_n^0(i,f) - \boldsymbol{A}_n(i,f)| \leq \delta_1 := \frac{4}{\rho_1(4+\overline{D})}$, $\boldsymbol{A}_n(i,f) \geq \rho_1$, $|\widehat{\boldsymbol{\lambda}^0}(f) - \boldsymbol{\lambda}(f)| \leq \delta_2 := \frac{4}{\rho_2(4+N\overline{D})}$ and $\boldsymbol{\lambda}(f) \geq \rho_2$ for all n, i, f.

Note that $\overline{D}_1$ characterizes the "conditioning" of $\boldsymbol{A}_n$ under the Kullback-Leibler (KL) divergence (denoted by $\mathbb{D}_{\mathrm{KL}}(\cdot,\cdot)$) sense, and $|\overline{D}_2|$ measures how far $\boldsymbol{\lambda}$ is away from the uniform distribution. Under the above assumption, we show that the EM algorithm improves upon the initialization:

Theorem 2 *Let $\delta_{\min} = \min(\delta_1, \delta_2)$. Assume that the following hold:*

$$N \geq \max\left(\frac{33\log(3SF/\mu)}{\rho_1 \overline{D}_1}, \frac{4\log(4SF^2/(3p\rho_2\mu))}{\overline{D}}\right),$$
$$S \geq \frac{192F^2\log(12NFI/\mu)}{p^2\rho_2^2\delta_{\min}^2},\ \overline{D} \geq \max\left\{\frac{8-4\rho_1^2}{\rho_1^2}, \frac{8-4\rho_2^2}{N\rho_2^2}\right\}.$$

Then, under Assumption 2, the EM algorithm in [3] outputs $\widehat{\boldsymbol{A}}_n(i,f), \widehat{\boldsymbol{\lambda}}(f)$ that satisfy the following with a probability greater than or equal to $1-\mu$:

$$|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)|^2 \leq \frac{48\log(12NFI/\mu)}{Sp\boldsymbol{\lambda}(f)} \leq \delta_1^2,$$
$$|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)|^2 \leq \frac{192F^2\boldsymbol{\lambda}(f)\log(12NFI/\mu)}{S} \leq \delta_2^2.$$

The proof of the theorem can be found in a longer version of the paper in the appendix. Our proof extends the analysis of a different EM algorithm proposed in [10] that is designed for learning the Dawid-Skene model in crowdsourcing. The EM algorithm there effectively learns a naive Bayes model when the latent variable is uniform. Our analysis covers the more recent EM algorithm in [3] that can handle $\boldsymbol{\lambda}$'s who have general PMFs beyond the uniform distribution.

Since `CNMF-SPA` is a natural economical initialization for Yeredor and Haardt's EM, we combine the two algorithms together and refer to this procedure as `CNMF-SPA-EM`.

IV. Experiments

In this section, we present experiments to showcase the effectiveness of the proposed framework.

TABLE I: MSE & MRE for $N = 5, F = 5, I = 10, p = 0.5$

Algorithms	Metric	$S = 10^3$	$S = 10^4$	$S = 10^5$	$S = 10^6$
CNMF-SPA [Proposed]	MSE	0.0702	0.0257	0.0211	**0.0204**
CNMF-SPA-EM [Proposed]	MSE	**0.0560**	**0.0230**	**0.0207**	**0.0204**
RAND-EM [3]	MSE	0.0855	0.0405	0.0298	0.0502
CTD [2]	MSE	0.1589	0.0260	0.0211	0.0205
CTD-EM [3]	MSE	0.1196	0.0233	0.0208	**0.0204**
CNMF-SPA [Proposed]	MRE	0.8084	0.3228	0.1137	0.0356
CNMF-SPA-EM [Proposed]	MRE	**0.6922**	**0.2077**	**0.0682**	**0.0219**
RAND-EM [3]	MRE	0.8285	0.3399	0.2226	0.3931
CTD [2]	MRE	0.9237	0.3081	0.0955	0.0309
CTD-EM [3]	MRE	0.8312	0.2180	0.0681	0.0220

TABLE II: MovieLens Action Movies set

Algorithm	RMSE	MAE	Time (s)
CNMF-SPA [Proposed]	0.8536±0.0071	0.6679±0.0061	0.029
CNMF-SPA-EM [Proposed]	**0.7761±0.0039**	**0.5936±0.0039**	2.554
CTD [2]	0.8792±0.0137	0.6640±0.0104	17.808
CTD-EM [3]	0.7872±0.0053	0.6038±0.0076	20.578
BMF [11]	0.8020±0.0015	0.6268±0.0017	45.967
Global Average	0.9471±0.0010	0.6954±0.0010	–
User Average	0.8960±0.0012	0.6834±0.0007	–
Movie Average	0.8844±0.0009	0.6979±0.0008	–

A. Synthetic Data

We consider $N = 5$ RV's where each variable takes $I = 10$ discrete values. The columns of the conditional PMF matrices (factor matrices) $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$ and the prior probability vector $\boldsymbol{\lambda} \in \mathbb{R}^F$ are generated with $F = 5$. The so-called ε-separability condition on $\boldsymbol{H}$ in (3) holds with $\varepsilon = 0.1$; see the definition of ε-separability in [9]. We generate S realizations of the joint PMF and randomly hide each variable's realization with probability $p = 0.5$. We fix $M = 3$. The *mean squared error* (MSE) of the factors and the *mean relative error* (MRE) of the recovered joint PMFs (see [2]) are evaluated. MRE is more preferred for evaluation, but it is hard to compute (due to memory issues) for large N. The results are averaged from 20 random trials. We benchmark our method using the EM algorithm by Yeredor and Haardt [3] initialized by random guesses (denoted as `RAND-EM`) and an alternating optimization (AO) algorithm-based coupled tensor decomposition (CTD) (denoted as `CTD-EM`), respectively.

Table I shows that the proposed approaches `CNMF-SPA` and `CNMF-SPA-EM` exhibit promising performance. In particular, `CNMF-SPA` is effective for initializing the EM algorithm whereas `RAND-EM` sometimes struggles to attain good performance. `CTD-EM` also works well when S is large, perhaps because the CTD stage needs a large S to estimate the three-dimensional marginals accurately.

B. Real Data : Recommender Systems

We test the approaches using the MovieLens 20M dataset [12], which has many missing values. We first round the ratings to the closest integers so that every movie's rating resides in $\{1, 2, \ldots, 5\}$. We choose select a subset of movies from the action movie genre. This way, Z_i for $i = 1, \ldots, 30$ represent the ratings of movie i, and all Z_i's alphabets are $\{1, 2, \ldots, 5\}$. We predict the rating for a movie (e.g., movie N) by user k via computing $\mathbb{E}[i_N | r_k(1), \ldots, r_k(N-1)]$ (i.e., using the MMSE estimator), where $r_k(i)$ denotes the rating of movie i by user k. This can be done via estimating $\mathsf{Pr}(i_1, \ldots, i_N)$.

TABLE III: UCI Dataset Car

Algorithm	Accuracy (%)	Time (s)
CNMF-SPA [Proposed]	69.26±2.28	0.007
CNMF-SPA-EM [Proposed]	**86.61±1.76**	0.018
CTD [2]	83.47±2.34	0.845
CTD-EM [3]	85.72±1.88	0.955
SVM	83.65±1.58	0.147
Linear Regression	80.68±1.61	0.029
Neural Net	85.00±3.22	0.193
SVM-RBF	76.22±3.93	0.793
Naive Bayes	83.42±2.15	0.026

TABLE IV: UCI Dataset Mushroom

Algorithm	Accuracy (%)	Time (sec.)
CNMF-SPA [Proposed]	92.23+/-6.15	0.025
CNMF-SPA-EM [Proposed]	**99.47+/-0.80**	0.242
CTD [2]	96.40+/-0.59	13.695
CTD-EM [3]	97.18+/-1.21	13.931
SVM	97.47+/-0.46	37.213
Linear Regression	93.38+/-0.59	0.040
Neural Net	98.98+/-1.97	1.036
SVM-RBF	98.89+/-0.34	2.291
Naive Bayes	94.84+/-0.55	0.048

We create the validation and testing sets by randomly hiding 20% and 30% of the dataset for each trial. The remianing 50% is used for training (learning joint PMF in our approach). In this task, we also use one of the popular recommender system algorithms, *biased matrix factorization* (BMF) method [11] as a baseline. The rank F for all the methods (ranging from 5 to 25) and the number of iterations needed for EM are chosen using the validation set. The results are taken from 20 random trials. We report the *root mean squared error* (RMSE) and *mean absolute error* (MAE) of the predicted ratings.

From Table II, one can see that the proposed methods are promising in terms of prediction accuracy and runtime. Note BMF is specialized for recommender systems, while the proposed approaches are for generic joint PMF recovery. The fact that our methods perform better suggests that the underlying joint PMF is well captured by the proposed CNMF approach. Another important observation is that the CPD method in [2] does not perform as well compared to the proposed pairwise marginals based methods. This may be because of the noisy estimation for three-dimensional marginals, due to the sparse nature of the user-movie datasets. In particular, `CNMF-SPA` is fast with acceptable prediction accuracy. In addition, `CNMF-SPA-EM` presents a good accuracy and speed tradeoff—it exhibits the lowest RMSEs and MAEs, and is 8 times faster than the state-of-the-art algorithm, i.e., `CTD-EM`.

C. Real Data : Data Classification

We consider UCI datasets for classification tasks. We split each dataset into training, validation and testing sets in the ratio of $50:20:30$. For our approach, we estimate the joint PMF of the features and the label using the training set, and then predict the labels on the testing data by constructing a *Maximum A Posterior* (MAP) predictor (i.e., predicting the labels conditioned on the features). For each dataset, we perform 20 trials with randomly partitioned training/testing/validation sets and take average of the results.

Tables III and IV show results on the UCI datasets Car and Mushroom. Car has 1,728 data samples from 4 classes, and Mushroom has 8,124 data samples belonging to two classes. We have $N = 7$ and $N = 22$ for the two datasets, respectively, and the average I's are 4 and 6, respectively. We set $M = 5$ in all experiments, and select rank F as before. In all the cases, one can see that the proposed combination `CNMF-SPA-EM` gives the most promising results. In particular, the solely using `CNMF-SPA` is not as promising, perhaps because the N and I are not large enough to ensure a high-accuracy performance of the `CNMF-SPA` stage. However, when initialized using `CNMF-SPA`, the `CNMF-SPA-EM` outputs the best classification accuracy and is at least 50 times faster than `CTD-EM` that is suggested in [3]. This suggests that even under critical scenarios, `CNMF-SPA` still offers useful initialization for EM.

V. Conclusion

We proposed a new framework for recovering joint PMF of any number of discrete RVs from marginal distributions. Unlike a recent approach that relies on three-dimensional marginals, our approach only uses two-dimensional marginals, which naturally has reduced-sample complexity and lighter computational burden. We proposed a virtual NMF framework and employed a Gram-Schmidt-like scalable algorithm for handling our formulation. We showed that the proposed framework is effectiveness under realistic conditions, e.g., finite samples. We also showed that an existing EM algorithm can provably improve the output of our NMF approach, using theoretical analysis and experimental validation. The combined NMF and EM approach admits economical updates and exhibits appealing joint PMF recovery accuracy.

References

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, learning, and 'Kolmogorov extension'for finite-alphabet random vectors," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, 2018.

[3] 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.

[4] N. Gillis and S. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, April 2014.

[5] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6343–6354, 2015.

[6] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, March 2019.

[7] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, p. 257, 2014.

[8] 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. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[9] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," *arXiv preprint arXiv:1909.12325*, 2019.

[10] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2014, pp. 1260–1268.

[11] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.
[12] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 19:1–19:19, Dec. 2015.
[13] A. Anandkumar, D. Hsu, and S. Kakade, "A method of moments for mixture models and hidden markov models," *Journal of Machine Learning Research*, vol. 23, 03 2012.

APPENDIX A
PROOF OF THEOREM 1

Consider the noisy matrix factorization model as below:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{H}^{\top} + \boldsymbol{N}, \tag{5}$$

where $\boldsymbol{W} \in \mathbb{R}^{L\times F}$, $\boldsymbol{H} \in \mathbb{R}^{K\times F}$, $\boldsymbol{W} \geq \boldsymbol{0}$ and $\boldsymbol{H} \geq \boldsymbol{0}$, and $\boldsymbol{N} \in \mathbb{R}^{L\times K}$ represents the noise. Also assume that $\mathrm{rank}(\boldsymbol{W}) = F$ and $\boldsymbol{H}\boldsymbol{1} = \boldsymbol{1}$, $\boldsymbol{H} = \boldsymbol{\Pi}\begin{bmatrix}\boldsymbol{I}_F \\ \boldsymbol{H}^*\end{bmatrix}$ where $\boldsymbol{I}_F$ is the identity matrix of size F and $\boldsymbol{\Pi}$ is the permutation matrix. This implies that $\boldsymbol{H}$ satisfies *seperability* condition and that there exists $\boldsymbol{\Lambda} = \{l_1, \ldots, l_F\}$ such that $\boldsymbol{H}(\boldsymbol{\Lambda}, :) = \boldsymbol{I}_F$. Gillis and Vavasis [4] have shown that under the model in (5), SPA is provably robust to noise in estimating the factor matrix $\boldsymbol{W}$(see Theorem 3).

First, we characterize the noise in our data model given by (3) by re-expressing it as a 'virtual' separable NMF as in (5). Then, we utilize Theorem 3 to characterize the error in estimating $\boldsymbol{W}$ via SPA algorithm.

Consider the pairwise marginals $\boldsymbol{X}_{jk}$'s used to construct the matrix $\widetilde{\boldsymbol{X}}$ in (3). $\boldsymbol{X}_{jk}$'s are estimated by sample averaging of a finite number of realizations and thus the estimated $\boldsymbol{X}_{jk}$ (denoted as $\widehat{\boldsymbol{X}}_{jk}$) is always noisy; i.e., we have

$$\widehat{\boldsymbol{X}}_{jk} = \boldsymbol{X}_{jk} + \boldsymbol{N}_{jk}, \tag{6}$$

where the noise matrix $\boldsymbol{N}_{jk} \in \mathbb{R}^{I\times I}$, assuming $I_n = I$ for all $n \in \{1, \ldots, N\}$.

In order to characterize the estimation accuracy of $\widehat{\boldsymbol{X}}_{jk}$ using finite number of realizations, we have the following proposition:

Proposition 1 *Let $p \in (0, 1]$ be the probability that an RV is observed. Let S be the number of available realizations of N RVs. Assume that $p \geq (\frac{8}{S}\log(2/\delta))^{1/2}$. Then, with probability at least $1-\delta$,*

$$\|\boldsymbol{X}_{jk} - \widehat{\boldsymbol{X}}_{jk}\|_{\mathrm{F}} = \|\boldsymbol{N}_{jk}\|_{\mathrm{F}} \leq \phi,$$

holds for any distinct j, k where $\phi = \frac{\sqrt{2}(1+\sqrt{\log(2/\delta)})}{(p\sqrt{S})}$.

The proof of Proposition 1 is given in Sec. C.

By the definition of the Frobenius norm, we have

$$\sum_{c=1}^{I} \|\boldsymbol{N}_{ij}(:, c)\|_2^2 = \|\boldsymbol{N}_{ij}\|_{\mathrm{F}}^2 \leq \phi^2.$$

Applying norm equivalence $\frac{\|\boldsymbol{N}_{ij}(:,c)\|_1}{\sqrt{I}} \leq \|\boldsymbol{N}_{ij}(:, c)\|_2$, we get

$$\sum_{c=1}^{I} \|\boldsymbol{N}_{ij}(:, c)\|_1^2 \leq I\phi^2. \tag{7}$$

Eq. (7) implies that, for all $c \in \{1, \ldots, I\}$ and any i, j where $i \neq j$,

$$\begin{aligned} &\|\boldsymbol{N}_{ij}(:, c)\|_1^2 \leq I\phi^2, \\ \implies &\|\boldsymbol{N}_{ij}(:, c)\|_1 \leq \sqrt{I}\phi. \end{aligned} \tag{8}$$

By using the estimates $\widehat{\boldsymbol{X}}_{jk}$, the model given by (3) can be represented as

$$\begin{aligned} \widehat{\boldsymbol{X}} &= \begin{bmatrix} \widehat{\boldsymbol{X}}_{\ell_1\ell_{M+1}} & \cdots & \widehat{\boldsymbol{X}}_{\ell_1\ell_N} \\ \vdots & \vdots & \vdots \\ \widehat{\boldsymbol{X}}_{\ell_M\ell_{M+1}} & \cdots & \widehat{\boldsymbol{X}}_{\ell_M\ell_N} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix}}_{\boldsymbol{W}} \underbrace{\boldsymbol{D}(\boldsymbol{\lambda})[\boldsymbol{A}_{\ell_{M+1}}^{\top}, \ldots, \boldsymbol{A}_{\ell_N}^{\top}]}_{\boldsymbol{H}^{\top}} + \widetilde{\boldsymbol{N}} \\ &= \widetilde{\boldsymbol{X}} + \widetilde{\boldsymbol{N}}. \end{aligned} \tag{9}$$

Note that $\widehat{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{N}}$ all have the same size of $L \times K$. Assuming $I_n = I$ for all $n \in \{1, \ldots, N\}$, we have $L = MI$ and $K = (N - M)I$. Also note that $\boldsymbol{W}$ has a size of $L \times F$ and $\boldsymbol{H}$ has a size of $K \times F$.

Since any column of $\widetilde{\boldsymbol{N}}$ formed from the columns of M number of $\boldsymbol{N}_{ij}$'s, we have

$$\|\widetilde{\boldsymbol{N}}(:,q)\|_1 \leq M\sqrt{I}\phi, \tag{10}$$

where the last inequality is obtained by using triangle inequality and (8).

Next, we consider estimating $\boldsymbol{W}$ given in (9) using SPA algorithm. Before performing SPA to the data matrix $\widehat{\boldsymbol{X}}$, the columns of $\widehat{\boldsymbol{X}}$ are normalized with respect to the ℓ_1-norm. Let us denote the normalized data model as follows:

$$\overline{\boldsymbol{X}} = \overline{\boldsymbol{W}}\boldsymbol{H}^\top + \overline{\boldsymbol{N}}, \tag{11}$$

where $\overline{\boldsymbol{X}}$ and $\overline{\boldsymbol{W}}$ are column normalized versions (with respect to the ℓ_1 norm) of $\widehat{\boldsymbol{X}}$ and $\boldsymbol{W}$, respectively, and $\overline{\boldsymbol{H}}$ is row normalized version of $\boldsymbol{H}$.

Since the matrix $\widehat{\boldsymbol{X}}$ is noisy, the effect of normalization on $\widetilde{\boldsymbol{N}}$ can be characterized by Lemma 6. From the assumption $\|\widehat{\boldsymbol{X}}_{ij}(:,c)\|_1 \geq \eta$ for any $i \neq j$ and $c \in \{1,\ldots,I\}$, we get $\|\widehat{\boldsymbol{X}}(:,q)\|_1 \geq M\eta$ for any q. Combining Lemma 6 and Eq. (10), we get

$$\|\overline{\boldsymbol{N}}(:,q)\|_1 \leq \frac{2\sqrt{I}\phi}{\eta}. \tag{12}$$

Applying norm equivalence, we further have $\|\overline{\boldsymbol{N}}(:,q)\|_2 \leq \|\overline{\boldsymbol{N}}(:,q)\|_1$ and hence we get

$$\|\overline{\boldsymbol{N}}(:,q)\|_2 \leq \frac{2\sqrt{I}\phi}{\eta}. \tag{13}$$

Lemma 1 *Assume that $\|\overline{\boldsymbol{N}}(:,q)\|_2 \leq \varphi$ for any q and that $\overline{\boldsymbol{H}}$ satisfies ε-separability assumption in the model* (11)*. Suppose*

$$\left(\sigma_{\max}(\overline{\boldsymbol{W}})\varepsilon + \varphi\right) \leq \sigma_{\min}(\overline{\boldsymbol{W}})\min\left(\frac{1}{2\sqrt{F-1}}, \frac{1}{4}\right)\left(1 + 80\kappa^2(\overline{\boldsymbol{W}})\right)^{-1}.$$

Then, SPA identifies an index set $\widehat{\boldsymbol{\Lambda}} = \{\widehat{l}_1, \ldots \widehat{l}_F\}$ such that

$$\max_{1\leq f\leq F}\min_{\widehat{l}_f\in\widehat{\Lambda}}\left\|\overline{\boldsymbol{W}}(:,f) - \overline{\boldsymbol{X}}(:,\widehat{l}_f)\right\|_2 \leq \left(\sigma_{\max}(\overline{\boldsymbol{W}})\varepsilon + \varphi\right)\left(1 + 80\kappa^2(\overline{\boldsymbol{W}})\right), \tag{14}$$

where $\kappa(\overline{\boldsymbol{W}}) = \frac{\sigma_{\max}(\overline{\boldsymbol{W}})}{\sigma_{\min}(\overline{\boldsymbol{W}})}$ is the condition number of $\overline{\boldsymbol{W}}$.

The proof of Lemma 1 is given in Sec. D.

The right hand side of (14) can be written as

$$\begin{aligned}\max_{1\leq f\leq F}\min_{\widehat{l}_f\in\widehat{\boldsymbol{\Lambda}}}\left\|\overline{\boldsymbol{W}}(:,f) - \overline{\boldsymbol{X}}(:,\widehat{l}_f)\right\|_2^2 &= \frac{1}{M}\sum_{m=1}^{M}\max_{1\leq f\leq F}\min_{\widehat{l}_f\in\widehat{\boldsymbol{\Lambda}}}\left\|\boldsymbol{A}_{\ell_m}(:,f) - \widehat{\boldsymbol{A}}_{\ell_m}(:,f)\right\|_2^2 \\ &\geq \frac{1}{M}\max_{1\leq f\leq F}\min_{\widehat{l}_f\in\widehat{\boldsymbol{\Lambda}}}\left\|\boldsymbol{A}_{\ell_m}(:,f) - \widehat{\boldsymbol{A}}_{\ell_m}(:,f)\right\|_2^2,\end{aligned} \tag{15}$$

for any $m \in \{1,\ldots,M\}$, where the first equality is due to $\overline{\boldsymbol{W}} = [\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_M}^\top]^\top/M$, in which $\widehat{\boldsymbol{A}}_{\ell_m}$ denotes the corresponding estimate of $\boldsymbol{A}_{\ell_m}$.

Since $\|\boldsymbol{W}(:,f)\|_1 = M$ for any f, $\overline{\boldsymbol{W}} = \boldsymbol{W}/M$. Therefore, we have

$$\sigma_{\max}(\overline{\boldsymbol{W}}) = \sigma_{\max}(\boldsymbol{W})/M, \quad \sigma_{\min}(\overline{\boldsymbol{W}}) = \sigma_{\min}(\boldsymbol{W})/M, \quad \kappa(\overline{\boldsymbol{W}}) = \frac{\sigma_{\max}(\overline{\boldsymbol{W}})}{\sigma_{\min}(\overline{\boldsymbol{W}})} = \kappa(\boldsymbol{W}). \tag{16}$$

Therefore, by combining (13),(15),(16) and Lemma 1, SPA estimates for any $m \in \{1,\ldots,M\}$ such that

$$\max_{1\leq f\leq F}\min_{\widehat{l}_f\in\widehat{\boldsymbol{\Lambda}}}\left\|\boldsymbol{A}_{\ell_m}(:,f) - \widehat{\boldsymbol{A}}_{\ell_m}(:,f)\right\|_2 \leq \left(\sigma_{\max}(\boldsymbol{W})\varepsilon + \frac{2M\sqrt{I}\phi}{\eta}\right)\left(1 + 80\kappa^2(\boldsymbol{W})\right), \tag{17}$$

if the below condition is satisfied:

$$\frac{\sigma_{\max}(\boldsymbol{W})}{M}\varepsilon + \frac{2\sqrt{I}\phi}{\eta} \leq \sigma_{\min}(\boldsymbol{W})\min\left(\frac{1}{2\sqrt{F-1}}, \frac{1}{4}\right)\left(1 + 80\kappa^2(\boldsymbol{W})\right)^{-1}. \tag{18}$$

Letting $\varepsilon = \frac{M\min\left(\frac{1}{2\sqrt{F-1}},\frac{1}{4}\right)}{2\kappa(\boldsymbol{W})(1+80\kappa^2(\boldsymbol{W}))}$, from (18), we get the condition on ϕ as follows:

$$\phi \leq \frac{\eta\sigma_{\max}(\boldsymbol{W})\varepsilon}{4M\sqrt{I}}. \tag{19}$$

From Proposition 1, we have $\phi = \frac{\sqrt{2}(1+\sqrt{\log(2/\delta)})}{p\sqrt{S}}$ with probability greater than $1-\delta$. By substituting ϕ on the left had side of (19), we get the number of realizations S required to get the estimation error bound (17) as below:

$$S \geq \frac{32M^2I(1+\sqrt{\log(2/\delta)})^2}{\sigma^2_{\max}(\boldsymbol{W})\eta^2\varepsilon^2p^2}.$$

Note that Lemma 1 holds if $\overline{\boldsymbol{H}}$ satisfies ε-separability condition. By combining Assumption 1 and Lemma 7, we get that $\overline{\boldsymbol{H}}$ satisfies ε-separability assumption with probability greater than $1-\rho$, if

$$(N-M)I = \Omega\left(\frac{\varepsilon^{-2(F-1)}}{F}\log\left(\frac{F}{\rho}\right)\right). \tag{20}$$

By substituting ϕ in (17) and using the fact that for any matrix $\boldsymbol{A} \in \mathbb{R}^{I\times F}$, the matrix 2-norm $\|\boldsymbol{A}\|_2 \leq \sqrt{F}\max_{1\leq f\leq F}\|\boldsymbol{A}(:,f)\|_2$, we get the result (4) in the theorem.

Finally, we combine the probabilities involved in the results used in our proof. For the concentration bound in Proposition 1 and ε-separability condition on $\overline{\boldsymbol{H}}$ given by Lemma 7 to jointly occur with probability greater than $1-2\delta$, we can assign $\rho = \delta$ in (20).

This completes the proof.

Appendix B
Proof of Theorem 2

The EM algorithm proposed in [3] is as follows. Let $\boldsymbol{d}_s \in \mathbb{R}^N$ be the s-th joint realization of N RVs. Let $f_s \in \{1,\ldots,F\}$ be the realization of the 'latent variable' H in the sth realization. Suppose $\{z_n^{(1)}, z_n^{(2)}, \ldots, z_n^{(I_n)}\}$ denotes the alphabet set of n-th RV, then n-th entry in $\boldsymbol{d}_s$ denoted as $\boldsymbol{d}_s(n) \in \{z_n^{(1)}, z_n^{(2)}, \ldots, z_n^{(I_n)}\}$. The expectation maximization algorithm proposed in [3] has the following E-step and M-step which are executed alternatively until convergence:

E-step: The parameter $\widehat{q}_{s,f}$ is updated for all s,f using the current estimates of $\widehat{\boldsymbol{A}}_n$'s and $\widehat{\boldsymbol{\lambda}}$

$$\widehat{q}_{s,f} = \frac{\exp(\log(\widehat{\boldsymbol{\lambda}}(f)) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})\log(\widehat{\boldsymbol{A}}_n(i,f)))}{\sum_{f'=1}^{F}\exp(\log(\widehat{\boldsymbol{\lambda}}(f')) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})\log(\widehat{\boldsymbol{A}}_n(i,f')))} \tag{21}$$

M-step: Using the estimated $\widehat{q}_{s,f}$, $\widehat{\boldsymbol{A}}_n$ and $\widehat{\boldsymbol{\lambda}}$ are updated as:

$$\widehat{\boldsymbol{A}}_n(i,f) \leftarrow \frac{\sum_{s=1}^{S}\widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})}{\sum_{i'=1}^{I_n}\sum_{s=1}^{S}\widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i')})},\ \forall\ i,f \tag{22a}$$

$$\widehat{\boldsymbol{\lambda}}(f) \leftarrow \frac{\sum_{s=1}^{S}\widehat{q}_{s,f}}{\sum_{f'=1}^{F}\sum_{s=1}^{S}\widehat{q}_{s,f'}},\ \forall\ f. \tag{22b}$$

The proof of convergence for the above defined iterates is inspired by the convergence proof of an EM algorithm handling the crowdsourcing problem in [10]. There, the EM algorithm assumes a uniform latent distribution, i.e., $\boldsymbol{\lambda}(f) = 1/F$ while formulating the maximum likelihood function. In our case, we do not assume uniform prior for $\boldsymbol{\lambda}$.

First, we define certain events as below:

$$\mathcal{E}_1: \quad \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})\log\left(\frac{\boldsymbol{A}_n(i,f_s)}{\boldsymbol{A}_n(i,f)}\right) \geq N\overline{D}_1/2, \quad \text{for all } s \text{ and } f\neq f_s$$

$$\mathcal{E}_2: \quad \left|\sum_{s=1}^{S}\mathbb{I}(f_s=f)\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)}) - S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f)\right| \leq St_{nif}, \quad \text{for all } n,i,f,$$

$$\mathcal{E}_3: \quad \left|\sum_{s=1}^{S}\mathbb{I}(f_s=f)\mathbb{I}(\boldsymbol{d}_s(n)\neq 0) - S\boldsymbol{\lambda}(f)p\right| \leq St_{nif}, \quad \text{for all } n,i,f,$$

$$\mathcal{E}_4: \quad \left|\sum_{s=1}^{S}\mathbb{I}(f_s=f) - S\boldsymbol{\lambda}(f)\right| \leq Sc_f, \quad \text{for all } f,$$

where $\boldsymbol{d}_s(n) \neq 0$ represents that n-th RV is observed with any value from its alphabet set in the s-th sample, and $t_{nif} > 0, c_f > 0$ are scalars which will be assigned specific value later in the proof.

First, we consider the E-step update given in (21). The parameter $\widehat{q}_{s,f}$ can be bounded using the following lemma:

Lemma 2 *Assume that the event $\mathcal{E}_1$ happens and also assume that $\boldsymbol{A}_n, \boldsymbol{\lambda}$ and the initial estimates satisfy $|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)| \leq \delta_1$, $\boldsymbol{A}_n(i,f) \geq \rho_1$, $|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| \leq \delta_2$ and $\boldsymbol{\lambda}(f) \geq \rho_2$ for all n, i, f. Then, if $\widehat{q}_{s,f}$ is updated by* (21)*, the below holds:*

$$|\widehat{q}_{s,f} - \mathbb{I}(f_s = f)| \leq \exp\left(-\left(N\overline{D} + 1 - \frac{1}{\rho_2(\rho_2 - \delta_2)} + N\left(1 - \frac{1}{\rho_1(\rho_1 - \delta_1)}\right)\right) + \log(F)\right), \ \forall f, s. \tag{23}$$

The proof of Lemma 2 is given in Sec. E Next lemma shows that once $\widehat{q}_{s,f}$ updated in E-step is bounded, the subsequent M-step updates to $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$ are bounded.

Lemma 3 *Assume that $\mathcal{E}_2 \bigcap \mathcal{E}_3 \bigcap \mathcal{E}_4$ holds. Suppose $\widehat{q}_{s,f}$ updated by* (21) *satisfies the following:*

$$|\widehat{q}_{s,f} - \mathbb{I}(f_s = f)| \leq \beta, \ \forall f, s, \tag{24}$$

where $\beta > 0$ is a scalar. Then $\widehat{\boldsymbol{A}}_n$ and $\widehat{\boldsymbol{\lambda}}$ updated by (22) *are bounded by:*

$$|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)| \leq \frac{2St_{nif} + 2S\beta}{S\boldsymbol{\lambda}(f)p - St_{nif} - S\beta}, \tag{25a}$$

$$|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| \leq \frac{Sc_f + S\beta + SF\beta}{S - SF\beta}. \tag{25b}$$

The proof of Lemma 3 is given in Sec. F

Next, we show that the estimation accuracy bounds for $\widehat{\boldsymbol{A}}_n$ and $\widehat{\boldsymbol{\lambda}}$ given in Lemma 3 are less than or equal to the initial estimation accuracy. For this, we have the following lemma:

Lemma 4 *Assume that $\mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3 \bigcap \mathcal{E}_4$ happens. Also assume that $\boldsymbol{A}_n, \boldsymbol{\lambda}$ and the initial estimates satisfy $|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)| \leq \delta_1 := \frac{4}{\rho_1(4+\overline{D})}$, $\boldsymbol{A}_n(i,f) \geq \rho_1$, $|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| \leq \delta_2 := \frac{4}{\rho_2(4+N\overline{D})}$, $\boldsymbol{\lambda}(f) \geq \rho_2$ for all n, i, f and $\overline{D} \geq \max\left\{\frac{8-4\rho_1^2}{\rho_1^2}, \frac{8-4\rho_2^2}{N\rho_2^2}\right\}$. Suppose that the following holds $\forall g \in \{\{t_{nif}\}_{n,i,f}, \{c_f\}_f\}$:*

$$2\exp(-\frac{N\overline{D}}{2} + \log(F)) \leq g \leq \frac{p\rho_2}{8F}\min\left(\frac{4}{\rho_1(4+\overline{D})}, \frac{4}{\rho_2(4+N\overline{D})}\right). \tag{26}$$

Then, by updating the parameters using (21) *and* (22) *at least once (i.e., after runing the EM algorithm for at least one iteration), we have the following:*

$$\left|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)\right| \leq \frac{4t_{nif}}{\boldsymbol{\lambda}(f)p} \leq \delta_1, \tag{27a}$$

$$\left|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)\right| \leq 8Fc_f \leq \delta_2. \tag{27b}$$

The proof of Lemma 4 is given in Sec G.

Next step is to find out the probabilities that the bounds in (27) hold true. Specifically, we need to characterize the probability for the event $\mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3 \bigcap \mathcal{E}_4$ to happen and the conditions under which (26) holds . Theorem 4 in [10] characterizes the probabilities of the occurrence of events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$. Specifically, we get the following results:

$$\mathsf{Pr}(\mathcal{E}_1) \geq 1 - SF\exp\left(\frac{N\overline{D}_1}{33\log(1/\rho_1)}\right), \tag{28a}$$

$$\mathsf{Pr}(\mathcal{E}_2) \geq 1 - \sum_{n=1}^{N}\sum_{f=1}^{F}\sum_{i=1}^{I_n} 2\exp\left(-\frac{St_{nif}^2}{3p\boldsymbol{\lambda}(f)}\right), \tag{28b}$$

$$\mathsf{Pr}(\mathcal{E}_3) \geq 1 - \sum_{n=1}^{N}\sum_{f=1}^{F}\sum_{i=1}^{I_n} 2\exp\left(-\frac{St_{nif}^2}{3p\boldsymbol{\lambda}(f)}\right). \tag{28c}$$

In order to characterize $\mathsf{Pr}(\mathcal{E}_4)$, we observe that $\sum_{s=1}^{S} \mathbb{I}(f_s = f)$ is sum of i.i.d. Bernoulli random variables with mean $S\boldsymbol{\lambda}(f)$. Therefore, using the Chernoff bound, we have

$$\mathsf{Pr}\left(\left|\sum_{s=1}^{S} \mathbb{I}(f_s = f) - S\boldsymbol{\lambda}(f)\right| \geq Sc_f\right) \leq 2\exp(-Sc_f^2/(3\boldsymbol{\lambda}(f))), \quad \forall f. \tag{29}$$

By taking the union bound over all $f \in \{1, \ldots, F\}$, we obtain

$$\mathsf{Pr}(\mathcal{E}_4) \geq 1 - \sum_{f=1}^{F} 2\exp(-Sc_f^2/(3\boldsymbol{\lambda}(f))). \tag{30}$$

Summing the probability bounds for $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ given by (28) and for $\mathcal{E}_4$ given by (30), one can see that $\mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3 \bigcap \mathcal{E}_4$ holds with probability at least

$$1 - SI\exp\left(\frac{N\overline{D}_1}{33\log(1/\rho_1)}\right) - \sum_{n=1}^{N}\sum_{f=1}^{F}\sum_{i=1}^{I_n} 4\exp\left(-\frac{St_{nif}^2}{3p\boldsymbol{\lambda}(f)}\right) - \sum_{f=1}^{F} 2\exp\left(-\frac{Sc_f^2}{3\boldsymbol{\lambda}(f)}\right).$$

To ensure that the estimation error bounds for $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$ given by (27) hold with probability greater than $1-\epsilon$, the following conditions has to be satisfied simultaneously:

$$N \geq \frac{33\log(1/\rho_1)\log(3SF/\epsilon)}{\overline{D}_1} \tag{31}$$

$$S \geq \frac{3p\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{t_{nif}^2} \tag{32}$$

$$S \geq \frac{3\boldsymbol{\lambda}(f)\log(6F/\epsilon)}{c_f^2}. \tag{33}$$

We can assign specific values to t_{nif} and c_f such that the above conditions are satisfied. Let

$$t_{nif} := \sqrt{\frac{3p\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{S}}, \tag{34a}$$

$$c_f := \sqrt{\frac{3\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{S}}. \tag{34b}$$

By this selection of t_{nif} and c_f, the conditions in (32) and (33) hold. To enforce the condition (26), the following equalities have to hold:

$$\sqrt{\frac{3p\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{S}} \geq 2\exp\left(-\frac{N\overline{D}}{2} + \log(F)\right)$$

$$\sqrt{\frac{3\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{S}} \leq \frac{p\rho_2}{8F}\min\left(\frac{4}{\rho_1(4+\overline{D})}, \frac{4}{\rho_2(4+N\overline{D})}\right) = \frac{p\rho_2\delta_{\min}}{8F},$$

where $\delta_{\min} = \min(\delta_1, \delta_2)$. The above can be implied by the following:

$$N \geq \frac{4\log(2SF^2/(3p\rho_2\log(12NFI/\epsilon)))}{\overline{D}} \tag{35}$$

$$S \geq \frac{192F^2\log(12NFI/\epsilon)}{p^2\rho_2^2\delta_{\min}^2}, \tag{36}$$

where we have used $1 \geq \boldsymbol{\lambda}(f) \geq \rho_2$.

Using the inequality $\log x > 1 - \frac{1}{x}, x > 0$, we can express the condition (35) as

$$N \geq \frac{4\log(2SF^2/(3p\rho_2(1-\epsilon/(12NFI))))}{\overline{D}}$$

Using the fact that $1 - \epsilon/(12NFI) > \epsilon/2$, we can further write the above condition as follows:

$$N \geq \frac{4\log(4SF^2/(3p\rho_2\epsilon))}{\overline{D}}. \tag{37}$$

Combing the two conditions (31) and (37), we have

$$N \geq \max\left(\frac{33\log(3SF/\epsilon)}{\rho_1\overline{D}_1}, \frac{4\log(4SF^2/(3p\rho_2\epsilon))}{\overline{D}}\right), \tag{38}$$

where we have used the fact that $\log(1/\rho_1) \leq (1/\rho_1) - 1 < 1/\rho_1$.

To summarize, if (36) and (38) hold and t_{nif} and c_f are chosen to be as in (34), then, with probability at least $1-\epsilon$, the following inequalities hold by Lemma 4:

$$|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)|^2 \leq \frac{16t_{nif}^2}{p^2\boldsymbol{\lambda}(f)^2} \leq \frac{48\log(12NFI/\epsilon)}{Sp\boldsymbol{\lambda}(f)}$$
$$|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)|^2 \leq 64F^2c_f^2 \leq \frac{192F^2\boldsymbol{\lambda}(f)\log(12NFI/\epsilon)}{S}$$

This completes the proof.

Appendix C
Proof Proposition 1

Let $\boldsymbol{d}_s \in \mathbb{R}^N$ denote the sth realization of the joint PMF $\mathsf{Pr}(Z_1,\ldots,Z_N)$. Recall that p is the probability of an RV being observed in any realization. Let $\{z_n^{(1)},\ldots,z_n^{(I_n)}\}$ denote the alphabet set of Z_n. For simplicity, we assume that 0 does not belong to the alphabets of $Z_1,\ldots,Z_N$, and we use the notation $\boldsymbol{d}_s(j)=0$ to represent that 'Z_j is not observed in the sth realization'.

For S realizations of the joint PMF, i.e., $\{\boldsymbol{d}_s\}_{s=1}^S$, the sample averaging expressions for estimating $\boldsymbol{X}_{jk}$ is defined as follows:

$$\widehat{\boldsymbol{X}}_{jk}(i_j,i_k) = \frac{1}{|\mathcal{S}_{jk}|}\sum_{s\in\mathcal{S}_{jk}} \mathbb{I}[\boldsymbol{d}_s(j)=z_j^{(i_j)}, \boldsymbol{d}_s(k)=z_k^{(i_k)}],$$

where the indicator function $\mathbb{I}[E]$ is one if the event E happens and zero otherwise; e.g.,

$$\mathbb{I}[\boldsymbol{d}_s(j)=z_j^{(i_j)}, \boldsymbol{d}_s(k)=z_k^{(i_k)}] = \begin{cases} 1, & \boldsymbol{d}_s(j)=z_j^{(i_j)}, \boldsymbol{d}_s(k)=z_k^{(i_k)} \\ 0, & \text{o.w.} \end{cases}.$$

In addition, $\mathcal{S}_{jk} = \{s \mid \mathbb{I}[\boldsymbol{d}_s(j)\neq 0, \boldsymbol{d}_s(k)\neq 0]\}$.

Let us construct a random variable $V_{j,s}$, where $V_{j,s}=1$ if Z_j is observed in $\boldsymbol{d}_s$; otherwise $V_{j,s}=0$. With this definition, we can see that the parameter S_{jk} in Lemma 5 is the sum of S i.i.d. Bernoulli random variable given by

$$S_{jk} = \sum_{s=1}^{S} \mathbb{I}[V_{j,s}=1 \text{ and } V_{k,s}=1], \tag{39}$$

with mean $\mathbb{E}[S_{jk}] = Sp^2$, since $V_{j,s}$ and $V_{k,s}$ are independent.

In order to characterize the random variable S_{jk}, we can use Chernoff lower tail bound such that for $0<t<1$,

$$\mathsf{Pr}(S_{jk} \leq (1-t)Sp^2) \leq e^{-Sp^2t^2/2}. \tag{40}$$

Eq. (40) also implies that

$$\mathsf{Pr}(S_{jk} \geq (1-t)Sp^2)) \geq 1 - e^{-Sp^2t^2/2}. \tag{41}$$

Combining Lemma 5 and (41), we have

$$\begin{aligned}\mathsf{Pr}\left[\|\widehat{\boldsymbol{X}}_{jk} - \boldsymbol{X}_{jk}\|_{\mathrm{F}} \leq \frac{(1+\sqrt{\log(1/\delta)})}{\sqrt{(1-t)Sp^2}}\right] &\\ = \mathsf{Pr}\left[\|\widehat{\boldsymbol{X}}_{jk} - \boldsymbol{X}_{jk}\|_{\mathrm{F}} \leq \frac{(1+\sqrt{\log(1/\delta)})}{\sqrt{S_{jk}}}, S_{jk} \geq (1-t)Sp^2\right] &\\ \geq 1 - \delta - e^{-Sp^2t^2/2}, &\end{aligned}$$

where we have applied the De Morgan's law and the union bound to obtain the last inequality.

Setting $t=1/2$, we have

$$\begin{aligned}\mathsf{Pr}\left[\|\widehat{\boldsymbol{X}}_{jk} - \boldsymbol{X}_{jk}\|_{\mathrm{F}} \leq \frac{\sqrt{2}(1+\sqrt{\log(1/\delta)})}{p\sqrt{S}}\right] &\\ \geq 1 - \delta - e^{-Sp^2/8}. &\end{aligned} \tag{42}$$

It follows that if $p^2 \geq \frac{8}{S}\log(1/\delta)$, the right hand side of (42) is greater than $1-2\delta$.

APPENDIX D
PROOF OF LEMMA 1

From the assumption that $\overline{\boldsymbol{H}}$ satisfies ε-separability, there exists a set of indices $\boldsymbol{\Lambda} = \{l_1, \ldots, l_F\}$ such that

$$\overline{\boldsymbol{H}}(\boldsymbol{\Lambda}, :) = \boldsymbol{I}_F + \boldsymbol{E},$$

$\boldsymbol{E} \in \mathbb{R}^{F \times F}$ is the error matrix with $\|\boldsymbol{E}(l, :)\|_2 \leq \varepsilon$. and $\boldsymbol{I}_F$ is the identity matrix of size $F \times F$

Now we can write the normalized data model given in (11) as

$$\begin{aligned}\overline{\boldsymbol{X}} &= \overline{\boldsymbol{W}}\boldsymbol{H}^\top + \overline{\boldsymbol{N}} \\ &= \overline{\boldsymbol{W}}[\boldsymbol{I}_F + \boldsymbol{E}^\top, (\boldsymbol{H}^*)^\top] + \overline{\boldsymbol{N}} \\ &= \overline{\boldsymbol{W}}[\boldsymbol{I}_F, (\boldsymbol{H}^*)^\top] + [\overline{\boldsymbol{W}}\boldsymbol{E}^\top, \boldsymbol{0}] + \overline{\boldsymbol{N}},\end{aligned}$$

where the zero matrix $\boldsymbol{0}$ has the same dimension as that of $\boldsymbol{H}^*$. By defining the noise matrix $\boldsymbol{N} \in \mathbb{R}^{L \times K}$ such that $\boldsymbol{N} := [\overline{\boldsymbol{W}}\boldsymbol{E}^\top, \boldsymbol{0}] + \overline{\boldsymbol{N}}$, we have $\overline{\boldsymbol{X}} = \overline{\boldsymbol{W}}[\boldsymbol{I}_F \quad (\boldsymbol{H}^*)^\top] + \boldsymbol{N}$. Then, for any $q \in \{1, \ldots, K\}$, the following inequality holds:

$$\begin{aligned}\|\boldsymbol{N}(:, q)\|_2 &\leq \|\overline{\boldsymbol{W}}\|_2 \|\boldsymbol{E}(q, :)\|_2 + \|\overline{\boldsymbol{N}}(:, q)\|_2 \\ &\leq \sigma_{\max}(\overline{\boldsymbol{W}})\varepsilon + \varphi, \end{aligned} \tag{43}$$

where the first inequality is by the Cauchy-Schwartz inequality and by the assumptions in the lemma.

Then, we invoke Lemma 3 to characterize the estimation accuracy of $\overline{\boldsymbol{W}}$. Combining (43) and Lemma 3, we get the final result of the lemma given by (14) if

$$\left(\sigma_{\max}(\overline{\boldsymbol{W}})\varepsilon + \varphi\right) \leq \sigma_{\min}(\overline{\boldsymbol{W}}) \min\left(\frac{1}{2\sqrt{F-1}}, \frac{1}{4}\right)\left(1 + 80\kappa^2(\overline{\boldsymbol{W}})\right)^{-1}.$$

APPENDIX E
PROOF OF LEMMA 2

Consider the update to $\widehat{q}_{s,f}$ in the E-step given by (21). For any $f \neq f_s$,

$$\begin{aligned}\widehat{q}_{s,f} &\leq \frac{\exp(\log(\widehat{\boldsymbol{\lambda}}(f)) + \sum_{n=1}^{N}\sum_{i=1}^{I_n} \mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log(\widehat{\boldsymbol{A}}_n(i, f)))}{\exp(\log(\widehat{\boldsymbol{\lambda}}(f_s)) + \sum_{n=1}^{N}\sum_{i=1}^{I_n} \mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log(\widehat{\boldsymbol{A}}_n(i, f_s)))} \\ &= \frac{\widehat{\boldsymbol{\lambda}}(f)\prod_{n=1}^{N}\prod_{i=1}^{I_n}(\widehat{\boldsymbol{A}}_n(i, f))^{\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)})}}{\widehat{\boldsymbol{\lambda}}(f_s)\prod_{n=1}^{N}\prod_{i=1}^{I_n}(\widehat{\boldsymbol{A}}_n(i, f_s))^{\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)})}} \\ &= \frac{\widehat{\boldsymbol{\lambda}}(f)}{\widehat{\boldsymbol{\lambda}}(f_s)} \prod_{n=1}^{N}\prod_{i=1}^{I_n}\left(\frac{\widehat{\boldsymbol{A}}_n(i, f)}{\widehat{\boldsymbol{A}}_n(i, f_s)}\right)^{\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)})} \\ &= 1/\exp\underbrace{\left(\log\left(\widehat{\boldsymbol{\lambda}}(f_s)/\widehat{\boldsymbol{\lambda}}(f)\right) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log\left(\widehat{\boldsymbol{A}}_n(i, f_s)/\widehat{\boldsymbol{A}}_n(i, f)\right)\right)}_{B_f}. \end{aligned} \tag{44}$$

Then it follows that

$$\begin{aligned}B_f &= \log\left(\widehat{\boldsymbol{\lambda}}(f_s)/\widehat{\boldsymbol{\lambda}}(f)\right) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log\left(\widehat{\boldsymbol{A}}_n(i, f_s)/\widehat{\boldsymbol{A}}_n(i, f)\right) \\ &= \log\left(\frac{\boldsymbol{\lambda}(f_s)}{\boldsymbol{\lambda}(f)}\frac{\widehat{\boldsymbol{\lambda}}(f_s)}{\boldsymbol{\lambda}(f_s)}\frac{\boldsymbol{\lambda}(f)}{\widehat{\boldsymbol{\lambda}}(f)}\right) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log\left(\frac{\boldsymbol{A}_n(i, f_s)}{\boldsymbol{A}_n(i, f)}\frac{\widehat{\boldsymbol{A}}_n(i, f_s)}{\boldsymbol{A}_n(i, f_s)}\frac{\boldsymbol{A}_n(i, f)}{\widehat{\boldsymbol{A}}_n(i, f)}\right) \\ &= \log\left(\frac{\boldsymbol{\lambda}(f_s)}{\boldsymbol{\lambda}(f)}\right) + \log\left(\frac{\widehat{\boldsymbol{\lambda}}(f_s)}{\boldsymbol{\lambda}(f_s)}\right) - \log\left(\frac{\widehat{\boldsymbol{\lambda}}(f)}{\boldsymbol{\lambda}(f)}\right) + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)}) \log\left(\frac{\boldsymbol{A}_n(i, f_s)}{\boldsymbol{A}_n(i, f)}\right) \\ &\quad + \sum_{n=1}^{N}\sum_{i=1}^{I_n}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i)})\left[\log\left(\frac{\widehat{\boldsymbol{A}}_n(i, f_s)}{\boldsymbol{A}_n(i, f_s)}\right) - \log\left(\frac{\widehat{\boldsymbol{A}}_n(i, f)}{\boldsymbol{A}_n(i, f)}\right)\right]. \end{aligned} \tag{45}$$

To proceed, we can bound all the terms in (45). First we have

$$\begin{aligned}\log\left(\frac{\widehat{\boldsymbol{\lambda}}(f_s)}{\boldsymbol{\lambda}(f_s)}\right)-\log\left(\frac{\widehat{\boldsymbol{\lambda}}(f)}{\boldsymbol{\lambda}(f)}\right) &\geq \log\left(\rho_2-\delta_2\right)+\log\left(\rho_2\right)\\ &=\log\left(\rho_2(\rho_2-\delta_2)\right)\\ &\geq 1-\frac{1}{\rho_2(\rho_2-\delta_2)},\end{aligned} \tag{46}$$

where the first inequality uses the results $\widehat{\boldsymbol{\lambda}}(f)\geq\rho_2-\delta_2$, $\boldsymbol{\lambda}(f)\geq\rho_2$ and the facts $\widehat{\boldsymbol{\lambda}}(f),\boldsymbol{\lambda}(f)\leq 1$. The last inequality is due to the fact that $\log(x)>1-\frac{1}{x}$ for $x>0$.

Similarly, we can bound

$$\log\left(\frac{\widehat{\boldsymbol{A}}_n(i,f_s)}{\boldsymbol{A}_n(i,f_s)}\right)-\log\left(\frac{\widehat{\boldsymbol{A}}_n(i,f)}{\boldsymbol{A}_n(i,f)}\right)\geq 1-\frac{1}{\rho_1(\rho_1-\delta_1)}. \tag{47}$$

Assuming that $\mathcal{E}_1$ happens, we have

$$\begin{aligned}B_f &\geq \frac{N\overline{D}_2}{2}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+\frac{N\overline{D}_1}{2}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\\ &=\frac{N(\overline{D}_1+\overline{D}_2)}{2}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right),\end{aligned}$$

where the first inequality is obtained by using the definitions of $\overline{D}_2$ and event $\mathcal{E}_1$, equations (46) and(47). Defining $\overline{D}:=\frac{\overline{D}_1+\overline{D}_2}{2}$, we can have

$$B_f\geq N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right). \tag{48}$$

Combining (48) with (44), we have for every $f\neq f_s$,

$$\widehat{q}_{s,f}\leq 1/\exp B_f\leq\exp\left(-\left(N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\right)\right). \tag{49}$$

Using (49) and $\sum_{f=1}^{F}\widehat{q}_{s,f}=1$, we have

$$\widehat{q}_{s,f_s}=1-\sum_{f\neq f_s}\widehat{q}_{s,f}\geq 1-F\exp\left(-\left(N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\right)\right). \tag{50}$$

The inequalities in (49) and (50) can be summarized as follows:

$$|\widehat{q}_{s,f}-\mathbb{I}(f_s=f)|\leq\exp\left(-\left(N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\right)+\log(F)\right),\ \forall f,s, \tag{51}$$

where we have used the fact that $\exp(\cdot)$ is an increasing function.

Appendix F
Proof of Lemma 3

The first result given in (27a) follows similar steps as in Lemma 9 from [10] which is detailed below using the notations in our case:

According to the M-step update (22), we can write

$$\widehat{\boldsymbol{A}}_n(i,f)=\frac{A}{B},$$

where $A:=\sum_{s=1}^{S}\widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})$ and $B:=\sum_{i'=1}^{I_n}\sum_{s=1}^{S}\widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i')})$.

Assuming that the event $\mathcal{E}_2$ holds true, we can have

$$\begin{aligned}|A-S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f)| &\leq |\sum_{s=1}^{S}\mathbb{I}(f_s=f)\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})-S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f)|\\ &\quad+|\sum_{s=1}^{S}\widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})-\sum_{s=1}^{S}\mathbb{I}(f_s=f)\mathbb{I}(\boldsymbol{d}_s(n)=z_n^{(i)})|\\ &\leq St_{nif}+S\beta,\end{aligned} \tag{52}$$

where the last result is obtained from the definition of $\mathcal{E}_2$ and (24).

Assuming that the event $\mathcal{E}_3$ holds true, we can have

$$\begin{aligned}|B - S\boldsymbol{\lambda}(f)p| &\leq |\sum_{s=1}^{S} \mathbb{I}(f_s = f)\mathbb{I}(\boldsymbol{d}_s(n) \neq 0) - S\boldsymbol{\lambda}(f)p| \\ &\quad + |\sum_{i'=1}^{I_n}\sum_{s=1}^{S} \widehat{q}_{s,f}\mathbb{I}(\boldsymbol{d}_s(n) = z_n^{(i')}) - \sum_{s=1}^{S}\mathbb{I}(f_s = f)\mathbb{I}(\boldsymbol{d}_s(n) \neq 0)| \\ &\leq St_{nif} + S\beta \end{aligned} \tag{53}$$

where the last inequality is obtained by assuming that event $\mathcal{E}_3$ holds true and using (24).

Combining the bounds for A and B, we can get

$$\begin{aligned}\left|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)\right| &= \left|\frac{A}{B} - \boldsymbol{A}_n(i,f)\right| \\ &= \left|\frac{S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f) + A - S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f)}{S\boldsymbol{\lambda}(f)p + B - S\boldsymbol{\lambda}(f)p} - \boldsymbol{A}_n(i,f)\right| \\ &= \left|\frac{A - S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f) - \boldsymbol{A}_n(i,f)(B - S\boldsymbol{\lambda}(f)p)}{S\boldsymbol{\lambda}(f)p + B - S\boldsymbol{\lambda}(f)p}\right| \\ &\leq \frac{|A - S\boldsymbol{\lambda}(f)p\boldsymbol{A}_n(i,f)| + \boldsymbol{A}_n(i,f)\,|(B - S\boldsymbol{\lambda}(f)p)|}{|S\boldsymbol{\lambda}(f)p + B - S\boldsymbol{\lambda}(f)p|} \\ &\leq \frac{2St_{nif} + 2S\beta}{S\boldsymbol{\lambda}(f)p - St_{nif} - S\beta},\end{aligned}$$

where the last inequality is by the fact that $\boldsymbol{A}_n(i,f) \leq 1$ and the bounds from (52) and (53).

For the second result, consider the M-step update for $\widehat{\boldsymbol{\lambda}}$ given by (22). One can write $\widehat{\boldsymbol{\lambda}}(f) = C/D$ where

$$C = \sum_{s=1}^{S}\widehat{q}_{s,f}, \quad D = \sum_{f'=1}^{F}\sum_{s=1}^{S}\widehat{q}_{s,f'}.$$

Assume that the event $\mathcal{E}_4$ happens, using (24), we have

$$\begin{aligned}|C - S\boldsymbol{\lambda}(f)| &\leq \left|\sum_{s=1}^{S}\mathbb{I}(f_s = f) - S\boldsymbol{\lambda}(f)\right| + \left|\sum_{s=1}^{S}\widehat{q}_{s,f} - \sum_{s=1}^{S}\mathbb{I}(f_s = f)\right| \\ &\leq Sc_f + S\beta.\end{aligned}$$

In addition, we have

$$|D - S| \leq \left|\sum_{f'=1}^{F}\sum_{s=1}^{S}\mathbb{I}(f_s = f') - S\right| + \left|\sum_{f'=1}^{F}\sum_{s=1}^{S}\widehat{q}_{s,f'} - \sum_{f'=1}^{F}\sum_{s=1}^{S}\mathbb{I}(f_s = f')\right| \leq SF\beta.$$

Combining the bounds for C and D, we obtain

$$\begin{aligned}|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| &= \left|\frac{C}{D} - \boldsymbol{\lambda}(f)\right| \\ &= \left|\frac{(C - S\boldsymbol{\lambda}(f)) + S\boldsymbol{\lambda}(f)}{(D - S) + S} - \boldsymbol{\lambda}(f)\right| \\ &= \left|\frac{(C - S\boldsymbol{\lambda}(f_s)) - \boldsymbol{\lambda}(f)(D - S)}{(D - S) + S}\right| \\ &\leq \frac{Sc_f + S\beta + SF\beta}{S - SF\beta},\end{aligned}$$

where the last inequality is by using triangle inequality and the fact that $\boldsymbol{\lambda}(f) \leq 1$.

APPENDIX G
PROOF OF LEMMA 4

Consider the below term in (23) from Lemma 2:

$$N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)= N\left(\underbrace{\frac{\overline{D}}{2}+1-\frac{1}{\rho_1(\rho_1-\delta_1)}}_{E}\right)+\underbrace{\frac{N\overline{D}}{2}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}}_{F}.$$

In order to bound the term E as below

$$E:=\frac{\overline{D}}{2}+1-\frac{1}{\rho_1(\rho_1-\delta_1)}\geq\frac{\overline{D}}{4},$$

δ_1 has to be bounded such that

$$\begin{aligned}\delta_1&\leq\rho_1-\frac{4}{\rho_1(4+\overline{D})}\\&=\frac{4\rho_1^2+\overline{D}\rho_1^2-4}{\rho_1(4+\overline{D})}.\end{aligned}\tag{54}$$

Similarly, in order to bound F as below

$$F:=\frac{N\overline{D}}{2}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}\geq\frac{N\overline{D}}{4},$$

δ_2 has to be bounded such that

$$\begin{aligned}\delta_2&\leq\rho_2-\frac{4}{\rho_2(4+N\overline{D})}\\&=\frac{4\rho_2^2+N\overline{D}\rho_2^2-4}{\rho_2(4+N\overline{D})}.\end{aligned}\tag{55}$$

Without loss of generality, we can fix values to δ_1 and δ_2 such that the conditions (54) and (55) get satisfied respectively. Since $\overline{D}\geq\max\left\{\frac{8-4\rho_1^2}{\rho_1^2},\frac{8-4\rho_2^2}{N\rho_2^2}\right\}$, $4\rho_1^2+\overline{D}\rho_1^2\geq 8$ and $4\rho_2^2+N\overline{D}\rho_2^2\geq 8$ hold true. Therefore

$$\delta_1:=\frac{4}{\rho_1(4+\overline{D})},\qquad\delta_2:=\frac{4}{\rho_2(4+N\overline{D})}\tag{56}$$

satisfy the conditions (54) and (55). Note that since $\overline{D}\geq\frac{8-4\rho_1^2}{\rho_1^2}$, $\delta_1:=\frac{4}{\rho_1(4+\overline{D})}\leq\frac{\rho_1}{2}\leq 1$ and is a valid assignment. Similarly, we can observe that $\delta_2:=\frac{4}{\rho_2(4+N\overline{D})}\leq\frac{\rho_2}{2}\leq 1$. Therefore, using the assigned values of δ_1 and δ_2, we have

$$N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\geq\frac{N\overline{D}}{2}.\tag{57}$$

Then, we have

$$\begin{aligned}|\widehat{q}_{s,f}-\mathbb{I}(f_s=f)|&\leq\exp\left(-\left(N\overline{D}+1-\frac{1}{\rho_2(\rho_2-\delta_2)}+N\left(1-\frac{1}{\rho_1(\rho_1-\delta_1)}\right)\right)+\log(F)\right)\\&\leq\exp\left(-\frac{N\overline{D}}{2}+\log(F)\right)\\&\leq t_{\min}/2,\end{aligned}\tag{58}$$

where $t_{\min}=\min\{\{t_{nif}\},\{c_f\}\}$ and the first inequality is by Lemma 2, the second inequality is by using (57) and the third inequality is from the condition (26).

From (26), we have $t_{nif}\leq\frac{p\rho_2}{8}\leq\frac{p\boldsymbol{\lambda}(f)}{8}$. Combining this with Lemma 3, we get

$$\left|\widehat{\boldsymbol{A}}_n(i,f)-\boldsymbol{A}_n(i,f)\right|\leq\frac{2St_{nif}+2S\beta}{(7/8)S\boldsymbol{\lambda}(f)p-S\beta},$$

where β is defined such that $|\widehat{q}_{s,f} - \mathbb{I}(f_s = f)| \leq \beta,\ \forall f, s$.

From (58), the scalar β can be assigned a value such that $\beta = t_{\min}/2$. Then,

$$\begin{aligned}\left|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)\right| &\leq \frac{2St_{nif} + St_{\min}}{(7/8)S\boldsymbol{\lambda}(f)p - St_{\min}/2} \\ &\leq \frac{3St_{nif}}{(7/8)S\boldsymbol{\lambda}(f)p - St_{\min}/2},\end{aligned}$$

where the last step is obtained by using $t_{\min} \leq t_{nif}$. By using the condition $t_{\min} \leq \frac{p\rho_2}{8} \leq \frac{p\boldsymbol{\lambda}(f)}{8}$ from (26), we get

$$\left|\widehat{\boldsymbol{A}}_n(i,f) - \boldsymbol{A}_n(i,f)\right| \leq \frac{3St_{nif}}{(7/8)S\boldsymbol{\lambda}(f)p - S\boldsymbol{\lambda}(f)p/16} \leq \frac{4t_{nif}}{\boldsymbol{\lambda}(f)p}.$$

Since

$$\frac{4t_{nif}}{\boldsymbol{\lambda}(f)p} \leq \frac{4t_{nif}}{p\rho_2} \leq \frac{1}{2F}\min\left(\frac{4}{\rho_1(4+\overline{D})}, \frac{4}{\rho_2(4+N\overline{D})}\right) \leq \frac{4}{\rho_1(4+\overline{D})} = \delta_1,$$

the estimation error of the newly updated $\boldsymbol{A}_n$ as given in (22) is at least no worse than the initial estimation error δ_1.

Next, consider the result (27b) from Lemma 3. Assigning $\beta = t_{\min}/2$,

$$\begin{aligned}|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| &\leq \frac{Sc_f + S\beta + SF\beta}{S - SF\beta} \leq \frac{Sc_f + St_{\min}/2 + SFt_{\min}/2}{S - SFt_{\min}/2} \\ &\leq \frac{2Sc_f + St_{\min} + SFt_{\min}}{2S - SFt_{\min}} \leq \frac{2Sc_f + Sc_f + SFc_f}{2S - SFt_{\min}} \\ &\leq \frac{4SFc_f}{2S - SFt_{\min}} \leq \frac{4SFc_f}{S - SFt_{\min}} \leq 8Fc_f,\end{aligned}$$

where we have used the fact that $t_{\min} \leq 1/2F$ according to (26).

The above inequality also implies that

$$|\widehat{\boldsymbol{\lambda}}(f) - \boldsymbol{\lambda}(f)| \leq 8Fc_f \leq p\rho_2 \min\left(\frac{4}{\rho_1(4+\overline{D})}, \frac{4}{\rho_2(4+N\overline{D})}\right) \leq \frac{4}{\rho_2(4+N\overline{D})} = \delta_2.$$

That is, the estimation error of the newly updated $\boldsymbol{\lambda}$ as given in (22) is at least no worse than the initial estimation error δ_2.

APPENDIX H
LEMMATA

In this section, we present a collection of lemmata that are used in our proofs.

Lemma 5 *[13] Let $\delta \in (0,1]$ and let $\widehat{\boldsymbol{X}}_{jk}$ be the empirical average of S_{jk} independent co-occurrences of random variables Z_j and Z_k. Then the following holds*

$$\mathsf{Pr}\left[\|\widehat{\boldsymbol{X}}_{jk} - \boldsymbol{X}_{jk}\|_{\mathrm{F}} \leq \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{S_{jk}}}\right] \geq 1 - \delta, \tag{59}$$

Theorem 3 *[4] Under the described NMF model in Eq.* (5)*, assume that $\|\boldsymbol{N}(:,q)\|_2 \leq \epsilon$ for all $q \in \{1,\ldots,K\}$, if the below holds:*

$$\epsilon \leq \sigma_{\min}(\boldsymbol{W})\min\left(\frac{1}{2\sqrt{F-1}}, \frac{1}{4}\right)\left(1 + 80\kappa^2(\boldsymbol{W})\right)^{-1},$$

then SPA identifies an index set $\widehat{\boldsymbol{\Lambda}} = \{\widehat{l}_1, \ldots \widehat{l}_F\}$ such that

$$\max_{1\leq f\leq F}\min_{\widehat{l}_f \in \widehat{\Lambda}}\left\|\boldsymbol{W}(:,f) - \widetilde{\boldsymbol{X}}(:,\widehat{l}_f)\right\|_2 \leq \epsilon\big(1 + 80\kappa^2(\boldsymbol{W})\big), \tag{60}$$

where $\kappa(\boldsymbol{W}) = \frac{\sigma_{\max}(\boldsymbol{W})}{\sigma_{\min}(\boldsymbol{W})}$ is the condition number of $\boldsymbol{W}$.

Lemma 6 *[9] Consider a vector $\boldsymbol{x} \in \mathbb{R}^L$ and the corresponding estimate of the vector $\widehat{\boldsymbol{x}}$ such that $\widehat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{n}$ where $\boldsymbol{n}$ represents the noise vector and $\boldsymbol{x}, \widehat{\boldsymbol{x}} \geq 0$. Assume that $\|\widehat{\boldsymbol{x}}\|_1 \geq \eta$ where $\eta > 0$ and $\|\boldsymbol{n}\|_1 < \|\boldsymbol{x}\|_1$. Suppose, the vector $\widehat{\boldsymbol{x}}$ is normalized with respect to its ℓ_1 norm. The normalized version can be represented as*

$$\frac{\widehat{\boldsymbol{x}}}{\|\widehat{\boldsymbol{x}}\|_1} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_1} + \widetilde{\boldsymbol{n}},$$

where $\|\widetilde{\boldsymbol{n}}\|_1 \leq \frac{2\|\boldsymbol{n}\|_1}{\eta}$.

Lemma 7 *[9] Let* $\rho > 0, \varepsilon > 0$*, and assume that the rows of* $\boldsymbol{H} \in \mathbb{R}^{K\times F}$ *are generated within the* $(F-1)$*-probability simplex uniformly at random (and then nonnegatively scaled). If the number of rows satisfies*

$$K = \Omega\left(\frac{\varepsilon^{-2(F-1)}}{F}\log\left(\frac{F}{\rho}\right)\right), \tag{61}$$

then, with probability greater than or equal to $1-\rho$*, there exist rows of* $\boldsymbol{H}$ *indexed by* $l_1, \ldots l_F$ *such that*

$$\|\boldsymbol{H}(l_f,:) - \boldsymbol{e}_f^\top\|_2 \leq \varepsilon, \;\; f = 1, \ldots, F.$$