Tensors, Learning, and 'Kolmogorov Extension' for Finite-alphabet Random Vectors

Nikos Kargas, Nicholas D. Sidiropoulos, *Fellow, IEEE*, and Xiao Fu, *Member, IEEE*

***Abstract*—Estimating the joint probability mass function (PMF) of a set of random variables lies at the heart of statistical learning and signal processing. Without structural assumptions, such as modeling the variables as a Markov chain, tree, or other graphical model, joint PMF estimation is often considered mission impossible – the number of unknowns grows exponentially with the number of variables. But who gives us the structural model? Is there a *generic*, 'non-parametric' way to control joint PMF complexity without relying on *a priori* structural assumptions regarding the underlying probability model? Is it possible to *discover* the operational structure without biasing the analysis up front? What if we only observe random subsets of the variables, can we still reliably estimate the joint PMF of all? This paper shows, perhaps surprisingly, that if the joint PMF of any three variables can be estimated, then the joint PMF of all the variables can be provably recovered under relatively mild conditions. The result is reminiscent of Kolmogorov's extension theorem – consistent specification of lower-dimensional distributions induces a unique probability measure for the entire process. The difference is that for processes of limited complexity (rank of the high-dimensional PMF) it is possible to obtain complete characterization from only three-dimensional distributions. In fact not all three-dimensional PMFs are needed; and under more stringent conditions even two-dimensional will do. Exploiting multilinear (tensor) algebra, this paper proves that such higher-dimensional PMF completion can be guaranteed – several pertinent identifiability results are derived. It also provides a practical and efficient algorithm to carry out the recovery task. Judiciously designed simulations and real-data experiments on movie recommendation and data classification are presented to showcase the effectiveness of the approach.**

***Index Terms*—Statistical learning, joint PMF estimation, tensor decomposition, rank, elementary probability, Kolmogorov extension, recommender systems, classification**

I. Introduction

Estimating a joint Probability Mass Function (PMF) of a set of random variables is of great interest in numerous applications in the fields of machine learning, data mining and signal processing. In many cases, we are given partial observations and/or statistics of the data, i.e., incomplete data, marginalized lower-dimensional distributions, or lower-order moments of the data, and our goal is to estimate the missing data. If the full joint PMF of all variables of interest were known, this would have been a straightforward task. A classical example is in recommender systems, where users rate only a small fraction of the total items (e.g., movies) and the objective is to make item recommendations to users according to predicted ratings. If the joint PMF of the item ratings is known, such recommendation is readily implementable based on the conditional expectation or mode of the unobserved ratings given the observed ratings. A closely related problem is top-K recommendation, where the goal is to predict the K items that a user is most likely to buy next. When the joint PMF of the items is known, it is easy to identify the K items with the highest individual or joint ('bundle') conditional probability given the observed user ratings. Another example is data classification. If the joint PMF of the features and the label is known, then given a test sample it is easy to infer the label according to the Maximum a *Posteriori* (MAP) principle. In fact, the joint PMF can be used to infer any of the features (or subsets of them), which is useful in imputing incomplete information in surveys or databases.

Despite its importance in signal and data analytics, estimating the joint PMF is often considered mission impossible in general, if no structure or relationship between the variables (e.g., a tree structure or a Markovian structure) can be assumed. This is true even when the problem size is merely moderate. The reason is that the number of unknown parameters is exponential in the number of variables. Consider a simple scenario of 10 variables taking 10 distinct values each. The number of parameters we need to estimate in this case is 10^{10}. The 'naive' approach for joint PMF estimation is counting the occurences of the joint variable realizations. In practice, however, when dealing with even moderately large sets of random variables, the probability of encountering any particular realization is very low. Therefore, only a small portion of the empirical distribution will be non-zero given a reasonable amount of data samples – this makes the approach very inaccurate.

In many applications, different workarounds have been proposed to circumvent this sample complexity problem. For example, in recommender systems, instead of trying to estimate the joint PMF of the ratings (which would be the estimation-theoretic gold standard), the most popular approach is based on low-rank matrix completion [2], [3], [4]. The idea is that the users can be roughly clustered into several types, and users of the same type would rate different movies similarly. Consequently, the user-rating matrix is approximately low rank

Original manuscript submitted to *IEEE Trans. on Signal Processing* November 30, 2017; revised April 24, 2018; accepted June 28, 2018. Supported in part by NSF IIS-1447788 and IIS-1704074. Conference version of part of this work appeared in *Information Theory and Applications Workshop* 2017 [1].

N. Kargas is with the Dept. of ECE, Univ. of Minnesota, Minneapolis, MN 55455; N. D. Sidiropoulos is with the Dept. of ECE, Univ. of Virginia, Charlottesville, VA 22904; X. Fu is with the School of EE and CS, Oregon State University, Corvallis, OR 97330. Author e-mails: karga005@umn.edu, nikos@virginia.edu, xiao.fu@oregonstate.edu

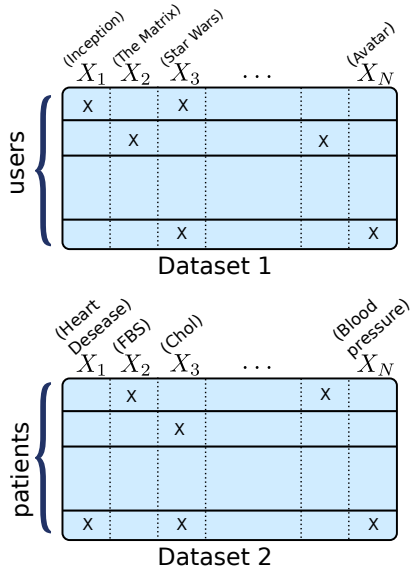


Fig. 1: Applications of joint PMF estimation. Top: recommender systems: given partially observed ratings of a user on movies, we would like to infer the unobserved ratings. Bottom: classification problems: given medical features of people, we would like to infer if a person has heart disease.

and this is used as prior information to infer the missing ratings. In classification, parsimonious function approximations are employed to model the relationship (or the conditional probability function) between the features and the label. Successful methods that fall into this category are support vector machines (linear function approximation), logistic regression (log-linear function approximation) and more recently kernels and neural networks (nonlinear function approximation) [5].

The above mentioned methods are nice and elegant, have triggered a tremendous amount of theoretical research and practical applications, and have been successful in many ways. However, these workarounds have not yet answered our question of interest: Can we ever reliably estimate the joint PMF of variables given limited data? This question is very well-motivated in practice, since knowledge of the joint PMF is indeed the gold standard: it enables optimal estimation under a variety of well-established criteria, such as mean-square error and minimum probability of error or Bayes risk. Knowing the joint PMF can facilitate a large variety of applications including recommender systems and classification in a unified and statistically optimal way, instead of resorting to often ad-hoc modeling tools.

This paper shows, perhaps surprisingly, that if the joint PMF of any three variables can be estimated, then the joint PMF of all the variables can be provably recovered under relatively mild conditions. The result is reminiscent of Kolmogorov's extension theorem – consistent specification of lower-dimensional distributions induces a unique probability measure for the entire process. The difference is that for processes of limited complexity (rank of the high-dimensional PMF) it is possible to obtain complete characterization from only three-dimensional distributions. In fact not all three-dimensional PMFs are needed; and under more stringent conditions even two-dimensional will do. The rank condition on the high-dimensional joint PMF has an interesting interpretation: loosely speaking, it means that the random variables are 'reasonably (in)dependent'. This makes sense, because estimation problems involving fully independent or fully dependent regressors and unknowns are contrived – it is the middle ground that is interesting. It is also important to note that the marginal PMFs of triples can be reliably estimated at far smaller sample complexity than the joint PMF of all variables. For example, for user-movie ratings, the marginal PMF of three given variables (movies) can be estimated by counting the co-occurrences of the given ratings (values of the variables) of the three given movies; but no user can rate *all* movies.

Contributions Our specific contributions are as follows:

• We propose a novel framework for joint PMF estimation given limited and possibly very incomplete data samples. Our method is based on a nice and delicate connection between the Canonical Polyadic Decomposition (CPD) [6], [7] and the naive Bayes model. The CPD model, sometimes referred to as the Parallel Factor Analysis (PARAFAC) model, is a popular analytical tool from multiway linear algebra. The CPD model has been used to model and analyze tensor data (data with more than two indices) in signal processing and machine learning, and it has found many successful applications, such as speech separation [8], blind CDMA detection [9], array processing [10], spectrum sensing and unmixing in cognitive radio [11], topic modeling [12], and community detection [13] – see the recent overview paper in [14]. Nevertheless, CPD has never been considered as a statistical learning tool for recovering a *general* joint PMF and our work is the first to establish the exciting connection [1].

• We present detailed identifiability analysis of the proposed approach. We first show that, any joint PMF can be represented by a naive Bayes model with a finite-alphabet latent variable – and the size of the latent alphabet (which happens to be the *rank* of the joint PMF tensor, as we will see) is bounded by a function of the alphabet sizes of the (possibly intermittently) observed variables. We further show that, if the latent alphabet size is under a certain threshold, then the joint PMF of *an arbitrary number* of random variables can be identified from three-dimensional marginal distributions. We prove this identifiability result by relating the joint PMF and marginal PMFs to the CPD model, which is known for its uniqueness even when the tensor rank is much larger than its outer dimensions.

• In addition to the novel formulation and identifiability results, we also propose an easily implementable joint PMF recovery algorithm. Our identification criterion can be considered as a coupled simplex-constrained tensor factorization problem, and we propose a very efficient alternating optimization-based algorithm to handle it. To deal with the probability simplex constraints that arise for PMF estimation,

[1]There are works that considered using CPD to model a joint PMF for some specific problems [12]. However, these works rely on specific physical interpretation of the associated model, which is sharply different to our setup – in which we employ the CPD model to explain a general joint PMF without assuming any physical model.

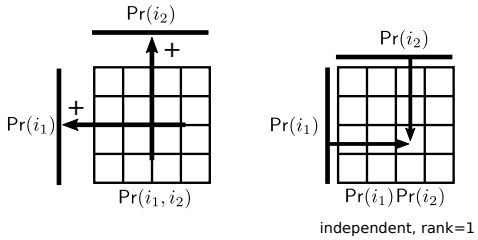


Fig. 2: It is impossible to recover the joint PMF from one-dimensional marginals without making strong assumptions.

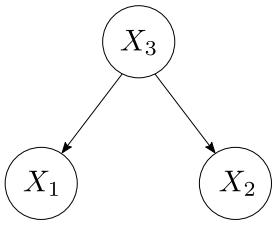


Fig. 3: Bayesian network of three variables.

the celebrated Alternating Direction Method of Multipliers (ADMM) algorithm is employed, resulting in lightweight iterations. Judiciously designed simulations and real experiments on movie recommendation and classification tasks are used to showcase the effectiveness of the approach.

Preliminary version of part of this work appeared at ITA 2017 [1]. This journal version includes new and stronger identifiability theorems and interpretations, detailed analysis of the theorems, and insightful experiments on a number of real datasets.

A. Notation

Bold, lowercase and uppercase letters denote vectors and matrices respectively. Bold, underlined, uppercase letters denote N-way ($N \geq 3$) tensors. Uppercase (lowercase) letters denote scalar random variables (realizations thereof, respectively). The outer product of N vectors is a N-way tensor with elements $(\mathbf{a}_1 \circ \mathbf{a}_2 \cdots \circ \mathbf{a}_N)(i_1, i_2, \ldots, i_N) = \mathbf{a}_1(i_1)\mathbf{a}_2(i_2)\cdots\mathbf{a}_N(i_N)$. The Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \otimes \mathbf{B}$. The Khatri-Rao (column-wise Kronecker) product of matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \odot \mathbf{B}$. The Hadamard (element-wise) product of matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \circledast \mathbf{B}$. We define vec($\underline{\mathbf{X}}$) the vector obtained by vertically stacking the elements of a tensor $\underline{\mathbf{X}}$ into a vector. Additionally, diag($\mathbf{x}$) $\in \mathbb{R}^{I \times I}$ denotes the diagonal matrix with the elements of vector $\mathbf{x} \in \mathbb{R}^I$ on its diagonal. The set of integers $\mathcal{S} = \{1, \ldots, N\}$ is denoted as $[N]$ and $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$.

II. Problem Statement

Consider a set of N random variables, i.e., $\{X_n\}_{n=1}^N$. Assume that each X_n can take I_n discrete values and only the joint PMFs of variable triples, i.e., $\mathsf{Pr}(X_j = i_j, X_k = i_k, X_\ell = i_\ell)$'s, are available. Can we identify the joint PMF of $\{X_n\}_{n=1}^N$, i.e., $\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N)$, from the three-dimensional marginals? This question lies at the heart of statistical learning. To see this, consider a classification problem and let $X_1 \ldots, X_{N-1}$ represent the set of observed features, and X_N the sought label. If $\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N)$ is known, then given a specific realization of the features, one can easily compute the posterior probability

$$\mathsf{Pr}(i_N | i_1 \ldots, i_{N-1}) = \frac{\mathsf{Pr}(i_1, \ldots, i_N)}{\sum_{i_N=1}^{I_N} \mathsf{Pr}(i_1, \ldots, i_{N-1}, i_N)},$$

and predict the label according the MAP principle (here $\mathsf{Pr}(i_1, \ldots, i_N)$ is shorthand for $\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N)$ and likewise $\mathsf{Pr}(i_N | i_1 \ldots, i_{N-1})$ for $\mathsf{Pr}(X_N = i_N | X_1 = i_1 \ldots, X_{N-1} = i_{N-1})$. In recommender systems, given a set of observed item ratings $X_1 \ldots, X_{N-1}$ one can compute the conditional expectation of an unobserved rating given the observed ones

$$\mathbb{E}(X_N | i_1, \ldots, i_{N-1}) = \sum_{i_N=1}^{I_N} i_N \mathsf{Pr}(i_N | i_1, \ldots, i_{N-1}).$$

At this point the reader may wonder why we consider recovery from three-dimensional joint PMFs and not from one- or two-dimensional PMFs. It is well-known that recovery from one-dimensional marginal PMFs is possible when all random variables are known to be independent. In this case, the joint PMF is equal to the product of the individual one-dimensional marginals. Interestingly, recovery from one-dimensional marginals is also possible when the random variables are known to be fully dependent i.e., one is completely determined by the other. In this case, the joint PMF can be recovered if each one-dimensional marginal is a unique permutation of the other.

However, complete (in)dependence is unrealistic in statistical estimation and learning practice. In general it is not possible to recover a joint PMF from one-dimensional marginals. An illustration for two variables is shown in Figure 2: $\mathsf{Pr}(i_1, i_2)$ can be represented as a matrix, and $\mathsf{Pr}(i_1)$, $\mathsf{Pr}(i_2)$ are 'projections' of the matrix along the row and column directions using the projector $\mathbf{1}^T$ and $\mathbf{1}$, respectively: $\mathsf{Pr}(i_1) = \sum_{i_2=1}^{I_2} \mathsf{Pr}(i_1, i_2)$ and $\mathsf{Pr}(i_2) = \sum_{i_1=1}^{I_1} \mathsf{Pr}(i_1, i_2)$. In this case, if we denote $\mathbf{P}$ the matrix such that $\mathbf{P}(i_1, i_2) = \mathsf{Pr}(X_1 = i_1, X_2 = i_2)$, then $\mathrm{rank}(\mathbf{P}) = r > 1$ if X_1 and X_2 are not independent. From basic linear algebra, one can see that knowing $\mathbf{1}^T\mathbf{P}$ and $\mathbf{P}\mathbf{1}$ is not enough for recovering $\mathbf{P}$ in general – since this is equivalent to solving a very underdetermined system of linear equations with $(I_1 + I_2) \times r$ variables but only $I_1 + I_2$ equations.

What if we know two-dimensional marginals? When the given random variables obey a probabilistic graphical model, and a genie reveals that model to us, then estimating a high-dimensional joint PMF from two-dimensional marginals may be possible. An example is shown in Figure 3. If we know *a priori* that random variables X_1 and X_2 are conditionally independent given X_3, one can verify that knowledge of $\mathsf{Pr}(X_1 = i_1, X_3 = i_3)$ and $\mathsf{Pr}(X_2 = i_2, X_3 = i_3)$ is sufficient to recover $\mathsf{Pr}(X_1 = i_1, X_2 = i_2, X_3 = i_3)$. However, this kind of approach hinges on knowing the probabilistic graph structure. Unfortunately, genies are hard to come by in real

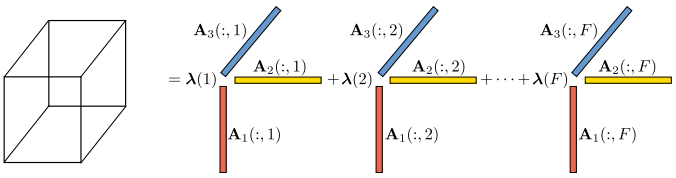


Fig. 4: Illustration of the rank decomposition of a three-way tensor.

life, and learning the graph structure from data is itself a very challenging problem in statistical learning [15].

In our problem setup, we do not assume any *a priori* knowledge of the graph structure, and in this sense we have a 'blind' joint PMF recovery problem. Interestingly, under certain conditions, this is no hindrance.

III. Preliminaries

Our framework is heavily based on low-rank tensor factorization and its nice identifiability properties. To facilitate our later discussion, we briefly introduce pertinent aspects of tensors in this section.

A. Rank Decomposition

An N-way tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a data array whose elements are indexed by N indices. A two-way tensor is a matrix, whose elements have two indices; i.e., $\mathbf{X}(i,j)$ denotes the (i,j)-th element of the matrix $\mathbf{X}$. If a matrix $\mathbf{X}$ has rank F, it admits a rank decomposition $\mathbf{X} = \sum_{f=1}^{F} \mathbf{A}_1(:,f) \circ \mathbf{A}_2(:,f) = \mathbf{A}_1\mathbf{A}_2^T$ where we have $\mathbf{A}_n = [\mathbf{A}_n(:,1), \ldots, \mathbf{A}_n(:,F)]$ and $\circ$ denotes the outer product of two vectors, i.e., $[\mathbf{x} \circ \mathbf{y}](i,j) = \mathbf{x}(i)\mathbf{y}(j)$. Similarly, if an N-way tensor $\underline{\mathbf{X}}$ has rank F, it admits the following rank decomposition:

$$\underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{A}_1(:,f) \circ \mathbf{A}_2(:,f) \circ \cdots \circ \mathbf{A}_N(:,f), \tag{1}$$

where $\mathbf{A}_n \in \mathbb{R}^{I_n \times F}$ and F is the smallest number for which such a decomposition exists. For convenience, we use the notation $\underline{\mathbf{X}} = [\![\mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$ to denote the decomposition. The above rank decomposition is also called the Canonical Polyadic Decomposition (CPD) or Parallel Factor Analysis (PARAFAC) model of a tensor. It is critical to note that every tensor admits a CPD, and that the rank F is not necessarily smaller than $I_1, \ldots, I_N$ – the latter is in sharp contrast to the matrix case [14]. In the matrix case, it is easy to see that $\mathbf{X}(i_1,i_2) = \sum_{f=1}^{F} \mathbf{A}_1(i_1,f)\mathbf{A}_2(i_2,f)$. Similarly, for an N-way tensor we have $\underline{\mathbf{X}}(i_1,i_2,\ldots,i_N) = \sum_{f=1}^{F} \prod_{n=1}^{N} \mathbf{A}_n(i_n,f)$. Sometimes one wishes to restrict the columns of $\mathbf{A}_n$'s to have unit norm (e.g., as in SVD). Therefore, the tensors can be represented as

$$\underline{\mathbf{X}} = \sum_{f=1}^{F} \boldsymbol{\lambda}(f)\mathbf{A}_1(:,f) \circ \mathbf{A}_2(:,f) \circ \cdots \circ \mathbf{A}_N(:,f), \tag{2}$$

or, equivalently

$$\underline{\mathbf{X}}(i_1,i_2,\ldots,i_N) = \sum_{f=1}^{F} \boldsymbol{\lambda}(f) \prod_{n=1}^{N} \mathbf{A}_n(i_n,f), \tag{3}$$

where $\|\mathbf{A}_n(:,f)\|_p = 1$ for a certain $p \geq 1$, $\forall\ n,f$, and $\boldsymbol{\lambda} = [\boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(F)]^T$ with $\|\boldsymbol{\lambda}\|_0 = F$ is employed to 'absorb' the norms of columns. An illustration of a three-way tensor and its CPD is shown in Figure 4. Under such cases, we denote the N-way tensor as $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$ – again, in this expression, we have automatically assumed that $\|\mathbf{A}_n(:,f)\|_p = 1$, $\forall\ n,f$ and a certain $p \geq 1$. We will refer to the decomposition of $\underline{\mathbf{X}}$ into nonnegative factors $\boldsymbol{\lambda} \in \mathbb{R}_+^F$, $\mathbf{A}_n \in \mathbb{R}_+^{I_n \times F}$ as nonnegative decomposition.

The following definitions will prove useful in the rest of the paper. We define the *mode-n matrix unfolding* of $\underline{\mathbf{X}}$ as the matrix $\mathbf{X}^{(n)}$ of size $\prod_{\substack{k=1 \\ k\neq n}}^{N} I_k \times I_n$. We have that $\underline{\mathbf{X}}(i_1,i_2,\ldots,i_N) = \mathbf{X}^{(n)}(j,i_n)$, where

$$j = 1 + \sum_{\substack{k=1 \\ k\neq n}}^{N} (i_k - 1)J_k \text{ with } J_k = \prod_{\substack{m=1 \\ m\neq n}}^{k-1} I_m.$$

In terms of the CPD factors, the mode-n matrix unfolding can be expressed as

$$\mathbf{X}^{(n)} = \left(\bigodot_{\substack{j=1 \\ j\neq n}}^{N} \mathbf{A}_j \right) \text{diag}(\boldsymbol{\lambda}) \mathbf{A}_n^T, \tag{4}$$

where $\bigodot_{\substack{j=1 \\ j\neq n}}^{N} \mathbf{A}_j = \mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \odot \cdots \odot \mathbf{A}_1$.

We can also express a tensor in a vectorized form $\underline{\mathbf{X}}(i_1,i_2,\ldots,i_N) = \mathbf{x}(j)$, where

$$j = 1 + \sum_{k=1}^{N} (i_k - 1)J_k \text{ with } J_k = \prod_{m=1}^{k-1} I_m.$$

In terms of the CPD factors, the vectorized form of a tensor can be expressed as

$$\text{vec}(\underline{\mathbf{X}}) = \left(\bigodot_{j=1}^{N} \mathbf{A}_j \right) \boldsymbol{\lambda}. \tag{5}$$

B. Uniqueness of Rank Decomposition

A distinctive feature of tensors is that they have *essentially* unique CPD under mild conditions – even when F is much larger than $I_1, \ldots, I_N$. To continue our discussion, let us first formally define what we mean by essential uniqueness of rank decomposition of tensors.

Definition 1. *(Essential uniqueness) For a tensor $\underline{\mathbf{X}}$ of (nonnegative) rank F, we say that a nonnegative decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$, $\boldsymbol{\lambda} \in \mathbb{R}_+^F$, $\mathbf{A}_n \in \mathbb{R}_+^{I_n \times F}$ is essentially unique if the factors are unique up to a common permutation. This means that if there exists another nonnegative decomposition $\underline{\mathbf{X}} = [\![\widehat{\boldsymbol{\lambda}}, \widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_N]\!]$, then, there exists a permutation matrix $\boldsymbol{\Pi}$ such that $\widehat{\mathbf{A}}_n = \mathbf{A}_n\boldsymbol{\Pi}, \forall n \in [N]$ and $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\Pi}^T\boldsymbol{\lambda}$.*

In other words, if a tensor has an essentially unique nonnegative CPD, then the only ambiguity is column permutation of the column-normalized factors $\{\mathbf{A}_n\}_{n=1}^N$, which simply amounts to a permutation of the rank-one 'chicken feet' outer products (rank-one tensors) in Fig. 4, that is clearly

unavoidable[2]. Regarding the essential uniqueness of tensors, let us consider the three-way case first. The following is arguably the most well-known uniqueness condition that was revealed by Kruskal in 1977.

Lemma 1. *[16] Let* $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, *where* $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$. *If* $k_{\mathbf{A}_1} + k_{\mathbf{A}_2} + k_{\mathbf{A}_3} \geq 2F + 2$ *then rank*$(\underline{\mathbf{X}}) = F$ *and the decomposition of* $\underline{\mathbf{X}}$ *is essentially unique.*

Here, $k_{\mathbf{A}}$ denotes the Kruskal rank of the matrix $\mathbf{A}$ which is equal to the largest integer such that every subset of $k_{\mathbf{A}}$ columns are linearly independent. Lemma 1 implies the following generic result: The decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$ is essentially unique, almost surely, if

$$\min(I_1, F) + \min(I_2, F) + \min(I_3, F) \geq 2F + 2. \quad (6)$$

This is because $k_{\mathbf{A}_n} = \min(I_n, F)$ with probability one if the elements of $\mathbf{A}_n$ are generated following a certain absolutely continuous distribution. More relaxed and powerful uniqueness conditions have been proven in recent years.

Lemma 2. *[17], [18] Let* $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, *where* $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$, $I_1 \leq I_2 \leq I_3$, $I_1 \geq 3$ *and* $F \leq I_3$. *Then, rank*$(\underline{\mathbf{X}}) = F$ *and the decomposition of* $\underline{\mathbf{X}}$ *is essentially unique, almost surely, if and only if* $F \leq (I_1 - 1)(I_2 - 1)$.

Lemma 3. *[17] Let* $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, *where* $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$, $I_1 \leq I_2 \leq I_3$. *Let* α, β *be the largest integers such that* $2^\alpha \leq I_1$ *and* $2^\beta \leq I_2$. *If* $F \leq 2^{\alpha+\beta-2}$ *then the decomposition of* $\underline{\mathbf{X}}$ *is essentially unique almost surely. The condition also implies that if* $F \leq \frac{(I_1+1)(I_2+1)}{16}$, *then* $\underline{\mathbf{X}}$ *has a unique decomposition almost surely.*

There are many more different uniqueness conditions for CPD. The take-home point here is that the CPD model is essentially generically unique even if F is much larger than I_1, I_2, I_3 – so long it is less than maximal possible rank. For example, in Lemma 3, F can be as large as $\mathcal{O}(I_1 I_2)$ (but not equal to $I_1 I_2$), and the CPD model is still unique.

Remark 1. We should mention that the above identifiability results are derived for tensors under a noiseless setup[3]. In addition, although the results are stated for real factor matrices, they are very general and also cover nonnegative $\mathbf{A}_n$'s due to the fact that the nonnegative orthant has positive measure. It follows that if a tensor is generated using random nonnegative factor matrices then under the noiseless setup, a plain CPD can recover the true nonnegative factors. On the other hand, in practice, instead of considering *exact* tensor decomposition, often *low-rank tensor approximation* is of interest, because of limited sample size and other factors. The best low-rank tensor approximation might not even exist in this case; fortunately, adding structural constraints on the latent factors can mitigate this, see [19]. In this work, our interest lies in revealing the fundamental limits of joint PMF estimation. Therefore, our analysis will be leveraging exact decomposition results, e.g., Lemmas 2-3. However, since the formulated problem naturally involves nonnegative latent factors, our computational framework utilizes this structural prior knowledge to enhance performance in practice.

[2]Generally, there is also column scaling / counter-scaling ambiguity [14]: a red column can be multiplied by γ and the corresponding yellow column divided by γ without any change in the outer product. There is no scaling ambiguity for *nonnegative* column-normalized representation $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$, where there is obviously no sign ambiguity and all scaling is 'absorbed' in $\boldsymbol{\lambda}$.

[3]In this context, noise will typically come from insufficient sample averaging in empirical frequency estimation.

IV. Naive Bayes Model: A Rank-decomposition Perspective

We will show that *any* joint PMF admits a naive Bayes model *representation*, i.e., it can be generated from a latent variable model with just one hidden variable. The naive Bayes model postulates that there is a hidden discrete random variable H taking F possible values, such that given $H = h$ the discrete random variables $\{X_n\}_{n=1}^N$ are conditionally independent. It follows that the joint PMF of $\{X_n\}_{n=1}^N$ can be decomposed as

$$\mathsf{Pr}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \mathsf{Pr}(f) \prod_{n=1}^{N} \mathsf{Pr}(i_n|f), \quad (7)$$

where $\mathsf{Pr}(f) := \mathsf{Pr}(H = f)$ is the prior distribution of the latent variable H and $\mathsf{Pr}(i_n|f) := \mathsf{Pr}(X_n = i_n|H = f)$ are the conditional distributions (Fig. 5). The naive Bayes model in (7) is also referred to as the latent class model [20] and is the simplest form of a Bayesian network [15]. It has been employed in diverse applications such as classification [21], density estimation [22] and crowdsourcing [23], just to name a few.

An interesting observation is that the naive Bayes model can be interpreted as a special nonnegative polyadic decomposition. This was alluded to in [24], [25] but not exploited for identifying the joint PMF from lower-dimensional marginals, as we do. Consider the element-wise representation in (3) and compare it with (7): each column of the factor matrices can represent a conditional PMF and the vector $\boldsymbol{\lambda}$ contains the prior probabilities of the latent variable H, i.e.,

$$\mathbf{A}_n(i_n, f) = \mathsf{Pr}(i_n|f), \quad \boldsymbol{\lambda}(f) = \mathsf{Pr}(f). \quad (8)$$

This is a *special* nonnegative polyadic decomposition model because it restricts $\mathbf{1}^T\boldsymbol{\lambda} = 1$. There is a subtle point however: the maximal rank F in a CPD (*canonical* polyadic decomposition) model is bounded, but the number of latent states (latent alphabet size) for the naive Bayes model may exceed this bound. Even if the number of latent states is under the maximal rank bound, a naive Bayes model may be *reducible*, in the sense that there exists a naive Bayes model with fewer latent states that generates the same joint PMF. The net result is that *every* joint PMF admits a naive Bayes model *interpretation* with bounded F, and every naive Bayes model is or can be reduced to a special CPD model. We have the following result.

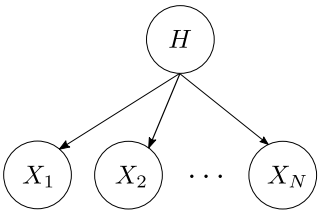


Fig. 5: Naive Bayes model.

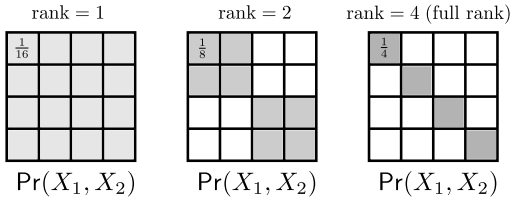


Fig. 6: Rank and independence.

Proposition 1. *The maximum F needed to represent an* arbitrary *PMF as a naive Bayes model is bounded by the following inequality*

$$F \leq \min_k \left(\prod_{\substack{n=1 \\ n \neq k}}^{N} I_n \right). \tag{9}$$

Proof: Let $\underline{\mathbf{X}} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$ denote a joint PMF of three random variables i.e., $\underline{\mathbf{X}}(i_1, i_2, i_3) = \mathsf{Pr}(X_1 = i_1, X_2 = i_2, X_3 = i_3)$. We define the following matrices

$$\begin{aligned}
\mathbf{A}_1 &:= [\underline{\mathbf{X}}(:,:,1), \cdots, \underline{\mathbf{X}}(:,:,I_3)], \\
\mathbf{A}_2 &:= [\mathbf{I}_{I_2 \times I_2}, \cdots, \mathbf{I}_{I_2 \times I_2}] = \mathbf{1}_{I_3}^T \otimes \mathbf{I}_{I_2 \times I_2}, \\
\mathbf{A}_3 &:= \mathbf{I}_{I_3 \times I_3} \otimes \mathbf{1}_{I_2}^T,
\end{aligned}$$

where $\mathbf{A}_1 \in \mathbb{R}_+^{I_1 \times I_2 I_3}, \mathbf{A}_2 \in \mathbb{R}_+^{I_2 \times I_2 I_3}, \mathbf{A}_3 \in \mathbb{R}_+^{I_3 \times I_2 I_3}$ and have used MATLAB notation $\underline{\mathbf{X}}(:,:,i_3)$ to denote the frontal slabs of the tensor $\underline{\mathbf{X}}$. Additionally, $\mathbf{I}_{I_n \times I_n}$ denotes the identity matrix of size $I_n \times I_n$ and $\mathbf{1}_{I_n}$ is a vector of all 1's of size I_n. Then every frontal slab of the tensor $\underline{\mathbf{X}}$ can be synthesized as $\underline{\mathbf{X}}(:,:,i_3) = \mathbf{A}_1 \mathrm{diag}(\mathbf{A}_3(i_3,:))\mathbf{A}_2^T$. Upon normalizing the columns of matrix $\mathbf{A}_1$ such that they sum to one and absorbing the scaling in $\boldsymbol{\lambda}$, i.e., $\mathbf{A}_1 = \widehat{\mathbf{A}}_1 \mathrm{diag}(\boldsymbol{\lambda})$ we can decompose the tensor as $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \widehat{\mathbf{A}}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$. The number of columns of each factor is $I_2 I_3$. Due to role symmetry, by permuting the modes of the tensor it follows that we need at most $\min(I_1 I_2, I_2 I_3, I_1 I_3)$ columns for each factor for exact decomposition.

The result is easily generalized to a four-way tensor $\underline{\mathbf{X}} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3 \times I_4}$ by noticing that each slab $\underline{\mathbf{X}}(:,:,:,i_4)$ is a three-way tensor and thus can be decomposed as $[\![\boldsymbol{\lambda}_{i_4}, \widehat{\mathbf{A}}_{1,i_4}, \mathbf{A}_{2,i_4}, \mathbf{A}_{3,i_4}]\!]$ as before. We define

$$\begin{aligned}
\boldsymbol{\lambda} &= [\boldsymbol{\lambda}_1^T, \cdots, \boldsymbol{\lambda}_{I_4}^T]^T, \\
\widehat{\mathbf{A}}_1 &= [\widehat{\mathbf{A}}_{1,1}, \cdots, \widehat{\mathbf{A}}_{1,I_4}], \quad \mathbf{A}_2 = [\mathbf{A}_{2,1}, \cdots, \mathbf{A}_{2,I_4}], \\
\mathbf{A}_3 &= [\mathbf{A}_{3,1}, \cdots, \mathbf{A}_{3,I_4}], \quad \mathbf{A}_4 = \mathbf{I}_{I_4} \otimes \mathbf{1}_{I_2 I_3}^T.
\end{aligned}$$

The four-way tensor can therefore be decomposed as $[\![\boldsymbol{\lambda}, \widehat{\mathbf{A}}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4]\!]$. Due to symmetry, the number of columns of each factor is at most $\min(I_1 I_2 I_3, I_2 I_3 I_4, I_1 I_3 I_4, I_1 I_2 I_4)$. By the same argument it follows that for a N-way tensor the bound on the nonnegative rank is $\min_k(\prod_{\substack{n=1 \\ n \neq k}}^{N} I_n)$. ∎

The proof of Proposition 1 employs the same type of argument used to prove the upper bound on tensor rank. The main difference is in the normalization – latent nonnegativity follows from data nonnegativity "for free" since the latent factors used for constructing the CPD are either fibers drawn from the joint PMF itself, or from identity matrices or Kronecker products thereof. While the proof is fairly straightforward for someone versed in tensor analysis, the implication of this proposition to probability theory is significant: it asserts that *every* joint PMF can be represented by a naive Bayes model with a *bounded* number of latent states $|\mathcal{H}|$. In fact, the connection between a naive Bayes model and CPD was utilized to approach some machine learning problems such as community detection and Gaussian Mixture Model (GMM) estimation in [13]. However, in those cases, the hidden variable has a specific physical meaning (e.g., $H = f$ represents the fth community in community detection) and thus connection was established using a specific data generative model. Here, we emphasize that even when there is no physically meaningful H or presumed generative model, *one can always represent an arbitrary joint PMF, possibly corresponding to a very complicated probabilistic graphical model, as a "simple" naive Bayes model with a bounded number of latent states F*. This result is very significant, also because it spells out that the latent structure of a probabilistic graphical model cannot be identified by simply assuming few hidden nodes; one has to limit the number of hidden node states as well.

We should remark that although any joint PMF admits a naive Bayes representation, this does not mean that such representation is unique. Clearly, F needs to be strictly smaller than the upper bound in (9) to guarantee uniqueness (cf. Lemmas 1-3). Fortunately, many joint PMFs that we encounter in practice are relatively low-rank tensors, since random variables in the real world are only moderately dependent. This leads to an interesting connection between linear dependence/independence and statistical dependence/independence. To explain, let us consider the simplest case where $N = 2$. In this case, we have

$$\mathsf{Pr}(i_1, i_2) = \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_1|f)\mathsf{Pr}(i_2|f). \tag{10}$$

The two-way model corresponds to Nonnegative Matrix Factorization (NMF) and is related to Probabilistic Latent Semantic Indexing (PLSI) [26], [27]. For the two-way model, independence of the variables implies that the probability matrix is rank-1. On the other hand, when the variables

TABLE I: Rel. error for different joint PMFs of 3 variables.

	Rank (F)		
	5	10	15
INCOME	2.1×10^{-2}	5.5×10^{-3}	5.1×10^{-3}
MUSHROOM	4.3×10^{-2}	2.4×10^{-2}	1.9×10^{-2}
MOVIELENS	1.8×10^{-2}	7.5×10^{-3}	4.1×10^{-3}

are fully dependent i.e., the value of one variable exactly determines the value of the other, the probability matrix is full-rank. However, low-rank does not necessarily mean that the variables are close to being independent as shown in Figure 6. There, a low rank probability matrix (rank $= 2$) can also model highly dependent random variables. In practice, we expect that random variables will be neither independent nor fully dependent and we are interested in cases where the rank of the joint PMF is lower (and ideally much lower) than the upper bound given in Proposition 1.

As a sanity check, we conducted preliminary experiments on some real-life data. As anticipated, we verified that many joint PMFs are indeed low-rank tensors in practice. Table I shows interesting results: The joint PMF of three movies over 5 rating values was first estimated, using data from the MovieLens project. The joint PMF is then factored using a nonnegative CPD model with different rank values. One can see that with rank as low as 5, the modeling error in terms of the relative error $\|\mathbf{X} - \widehat{\mathbf{X}}\|_F / \|\mathbf{X}\|_F$ is quite small, meaning that the low-rank modeling is fairly accurate. The same applies to two more datasets drawn from the UCI repository.

V. Joint PMF Recovery

A. General Procedures

The key observation that enables our approach is that the marginal distribution of any subset of random variables is also a nonnegative CPD model. This is a direct consequence of the law of total probability. Marginalizing with respect to the k-th random variable we have that

$$\begin{aligned}\sum_{i_k=1}^{I_k} \mathsf{Pr}(i_1, \ldots, i_N) &= \sum_{f=1}^{F} \sum_{i_k=1}^{I_k} \mathsf{Pr}(f) \prod_{n=1}^{N} \mathsf{Pr}(i_n|f) \\ &= \sum_{f=1}^{F} \mathsf{Pr}(f) \prod_{\substack{n=1 \\ n\neq k}}^{N} \mathsf{Pr}(i_n|f) \sum_{i_k=1}^{I_k} \mathsf{Pr}(i_k|f) \\ &= \sum_{f=1}^{F} \mathsf{Pr}(f) \prod_{\substack{n=1 \\ n\neq k}}^{N} \mathsf{Pr}(i_n|f), \end{aligned} \tag{11}$$

since $\sum_{i_n=1}^{I_n} \mathsf{Pr}(i_n|f) = 1$.

Consider the model in (7) and assume that the marginal distributions $\mathsf{Pr}(X_j = i_j, X_k = i_k, X_l = i_l)$, denoted $\mathsf{Pr}(i_j, i_k, i_l)$ for brevity, $\forall j, k, l \in [N],\ l > k > j$ are available and perfectly known. Then, there exists an exact decomposition of the form

$$\mathsf{Pr}(i_j, i_k, i_l) = \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_j|f)\mathsf{Pr}(i_k|f)\mathsf{Pr}(i_l|f). \tag{12}$$

The marginal distributions of triples of random variables satisfy $\underline{\mathbf{X}}_{jkl} = [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l]\!]$, where $\{\mathbf{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ are defined as in (8) and they satisfy $\mathbf{A}_l \geq \mathbf{0}$, $\mathbf{A}_k \geq \mathbf{0}$, $\mathbf{A}_j \geq \mathbf{0}$, $\mathbf{1}^T\mathbf{A}_l = \mathbf{1}^T$, $\mathbf{1}^T\mathbf{A}_k = \mathbf{1}^T$, $\mathbf{1}^T\mathbf{A}_j = \mathbf{1}^T$, $\boldsymbol{\lambda} > \mathbf{0}$, $\mathbf{1}^T\boldsymbol{\lambda} = 1$. Based on the connection between the naive Bayes model of lower-dimensional marginals and the joint PMF, we propose the following steps to recover the complete joint PMF from three-dimensional marginals.

Procedure: Joint PMF Recovery From Triples
[S1] Estimate $\underline{\mathbf{X}}_{jk\ell}$ from data;
[S2] Jointly factor $\underline{\mathbf{X}}_{jkl} = [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l]\!]$ to estimate $\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l\ \forall\ j, k, l$ using a CPD model with rank F;
[S3] Synthesize the joint PMF $\underline{\mathbf{X}}$ via $\mathsf{Pr}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \mathsf{Pr}(f) \prod_{n=1}^{N} \mathsf{Pr}(i_n|f)$, w/ $\mathsf{Pr}(i_n|f) = \mathbf{A}_n(i_n, f)$, $\mathsf{Pr}(f) = \boldsymbol{\lambda}(f)$.

One can see from step [S2], that if the individual factorization of at least one $\underline{\mathbf{X}}_{jkl}$ is unique, then the joint PMF is readily identifiable via [S3]. This is already very interesting. However, as we will show in Sec. VI, we may identify the joint PMF even when the marginal tensors do not have unique CPD. The reason is that many marginal tensors share factors and we can exploit this to come up with much stronger identifiability results.

B. Algorithm: Coupled Matrix/Tensor Factorization

Before we discuss theoretical results such as identifiability of the joint PMF using three or higher-dimensional marginals, we first propose an implementation of [S2] in the proposed procedure. For brevity, we assume we have estimates of three-dimensional marginal distributions, i.e., we are given empirical estimates $\widehat{\mathsf{Pr}}(X_j = i_j, X_k = i_k, X_l = i_l),\ \forall j, k, l \in [N],\ l > k > j$, which we put in a tensor $\underline{\mathbf{X}}_{jkl}$ i.e., $\underline{\mathbf{X}}_{jkl}(i_j, i_k, i_l) = \widehat{\mathsf{Pr}}(X_j = i_j, X_k = i_k, X_l = i_l)$.

The method can be easily generalized to any type of low-dimensional marginal distributions. Under the assumption of a low-rank CPD model, every empirical marginal distribution of three random variables can be approximated as follows

$$\widehat{\mathsf{Pr}}(i_j, i_k, i_l) \approx \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_j|f)\mathsf{Pr}(i_k|f)\mathsf{Pr}(i_l|f). \tag{13}$$

Therefore, in order to compute an estimate of the full joint PMF, we propose solving the following optimization problem

$$\begin{aligned} \min_{\{\mathbf{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \quad & \sum_j \sum_{k>j} \sum_{l>k} \frac{1}{2} \left\| \underline{\mathbf{X}}_{jkl} - [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l]\!] \right\|_F^2 \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{1}^T\boldsymbol{\lambda} = 1, \\ & \mathbf{A}_n \geq \mathbf{0},\ n = 1, \ldots, N, \\ & \mathbf{1}^T\mathbf{A}_n = \mathbf{1}^T,\ n = 1, \ldots, N. \end{aligned} \tag{14}$$

The optimization problem in (14) is an instance of coupled tensor factorization. Coupled tensor/matrix factorization is usually used as a way of combining various datasets that share dimensions and corresponding factor matrices [28], [29]. Notice that in the case where we have estimates of two-dimensional marginals, the optimization problem in (14) corresponds to coupled matrix factorization. The optimization

Algorithm 1 Coupled Tensor Factorization Approach

Input: A discrete valued dataset $\mathbf{D} \in \mathbb{R}^{M \times N}$
Output: Estimates of $\{\mathbf{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$
1: Estimate $\underline{\mathbf{X}}_{j,k,l} \quad \forall j,k,l \in [N],\ l > k > j$ from data.
2: Initialize $\{\mathbf{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ such that the probability simplex constraints are satisfied.
3: **repeat**
4: **for all** $n \in [N]$ **do**
5: Solve optimization problem (16)
6: **end for**
7: Solve optimization problem (17)
8: **until** convergence criterion satisfied

problem *per se* is very challenging and deserves developing sophisticated algorithms for handling it: first, when the number of random variables (N) gets large, there is a large number of optimization variables (i.e., $\{\mathbf{A}_n\}_{n=1}^N$) to be determined in (14) – and each $\mathbf{A}_n$ is an $I_n \times F$ matrix where I_n (the alphabet size of the n-th random variable) can be large. In addition, the probability simplex constraints impose some extra computational burden. Nevertheless, we found that, by carefully re-arranging terms, the formulated problem can be recast in convenient form and handled in a way that is reminiscent of the classical alternating least squares algorithm with constraints.

The idea is that we cyclically update variables $\{\mathbf{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ while fixing the remaining variables at their last updated values. Assume that we fix estimates $\boldsymbol{\lambda}, \mathbf{A}_n, \forall n \in [N] \setminus \{j\}$. Then, the optimization problem with respect to $\mathbf{A}_j$ becomes

$$\begin{aligned} \min_{\mathbf{A}_j} \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \quad & \frac{1}{2} \left\| \underline{\mathbf{X}}_{jkl} - [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l]\!] \right\|_F^2 \\ \text{subject to} \quad & \mathbf{A}_j \geq \mathbf{0},\ \mathbf{1}^T \mathbf{A}_j = \mathbf{1}^T. \end{aligned} \tag{15}$$

Note that we have dropped the terms that do not depend on $\mathbf{A}_j$. By using the *mode*-1 *matrix unfolding* of each tensor $\underline{\mathbf{X}}_{jkl}$, the problem can be equivalently written as

$$\begin{aligned} \min_{\mathbf{A}_j} \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \quad & \frac{1}{2} \left\| \mathbf{X}_{jkl}^{(1)} - (\mathbf{A}_l \odot \mathbf{A}_k) \text{diag}(\boldsymbol{\lambda}) \mathbf{A}_j^T \right\|_F^2 \\ \text{subject to} \quad & \mathbf{A}_j \geq \mathbf{0},\ \mathbf{1}^T \mathbf{A}_j = \mathbf{1}^T, \end{aligned} \tag{16}$$

which is a least-squares problem with respect to matrix $\mathbf{A}_j$ under probability simplex constraints on its columns. The optimization problem has the same form for each factor $\mathbf{A}_n$ due to role symmetry. In order to update $\boldsymbol{\lambda}$ we solve the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \sum_{j} \sum_{k > j} \sum_{l > k} \quad & \frac{1}{2} \left\| \text{vec}(\underline{\mathbf{X}}_{jkl}) - (\mathbf{A}_l \odot \mathbf{A}_k \odot \mathbf{A}_j) \boldsymbol{\lambda} \right\|_2^2 \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{1}^T \boldsymbol{\lambda} = 1. \end{aligned} \tag{17}$$

Both Problems (16) and (17) are linearly constrained quadratic programs, and can be solved to optimality by many standard solvers. Here, we propose to employ the *Alternating Direction Method of Multipliers* (ADMM) to solve these two sub-problems because of its flexibility and effectiveness in handling large-scale tensor decomposition [30], [31]. Details of the ADMM algorithm for solving Problems (16)-(17) can be found in the Appendix B. The whole procedure is listed in Algorithm 1. As mentioned, the algorithm is easily modified to cover the cases where higher-dimensional marginals or pairwise marginals are given, and thus these cases are omitted.

VI. Joint PMF Identifiability Analysis

In this section, we study the conditions under which we can identify $\Pr(i_1, \ldots, i_N)$ from marginalized lower-dimensional distributions. For brevity, we focus on three-dimensional as lower-dimensional distributions, and even though many more results are possible, we concentrate here on the case $I_n = I\ \forall n \in [N]$ for ease of exposition and manuscript length considerations. Similar type of analysis applies when $I_1, \ldots, I_N$ are different, however the analysis should be customized to properly address particular cases. Our aim here is to convey the spirit of what is possible in terms of identifiability results, as we cannot provide an exhaustive treatment (there are combinatorially many cases, clearly).

Obviously, if $\underline{\mathbf{X}}_{jkl}$ is individually identifiable for each combination of j,k,l, then, $\Pr(i_j|f)$, $\Pr(i_k|f)$, $\Pr(i_l|f)$, and $\Pr(f)$ are identifiable. This means that given three-dimensional marginal distributions, $\Pr(i_1, \ldots, i_N)$ is generically identifiable if $F \leq \frac{3I-2}{2}$ assuming that $I_n = I\ \forall n \in [N]$. This can be readily shown by invoking Lemma 1, equation (6), and the link between the naive Bayes model and tensor factorization discussed in Sec. IV. Note that $F \leq \frac{3I-2}{2}$ is already not a bad condition, since in many cases we have approximately low-rank tensors in practice. However, since we have many factor-coupled $\underline{\mathbf{X}}_{jkl}$'s, this identifiability condition can be significantly improved. We have the following theorems.

Theorem 1. *Assume that* $\Pr(i_n|f),\ \forall n \in [N]$ *are drawn from an absolutely continuous distribution, that* $I_1 = \ldots = I_N = I$*, and that the joint PMF* $\Pr(i_1, \ldots, i_N)$ *can be represented using a naive Bayes model of rank* F*. If* $N \leq I$ *then,* $\Pr(i_1, \ldots, i_N)$ *is almost surely (a.s) identifiable from the* $\Pr(i_j, i_k, i_l)$*'s if*

$$F \leq I(N-2)$$

If $N > I$ *then,* $\Pr(i_1, \ldots, i_N)$ *is a.s. identifiable from the* $\Pr(i_j, i_k, i_l)$*'s if*

$$F \leq \left(\lfloor \frac{\sqrt{NI-1}}{I} \rfloor I - 1 \right)^2$$

Proof: The proof is relegated to Appendix A. ■

Theorem 2. *Assume that* $\Pr(i_n|f),\ \forall n \in [N]$ *are drawn from an absolutely continuous distribution, that* $I_1 = \ldots = I_N = I$*, and that the joint PMF* $\Pr(i_1, \ldots, i_N)$ *can be represented using a naive Bayes model of rank* F*. Let* α *be the largest integer such that* $2^\alpha \leq \lfloor \frac{N}{3} \rfloor I$*. Then,* $\Pr(i_1, \ldots, i_N)$ *is a.s. identifiable from the* $\Pr(i_j, i_k, i_l)$*'s if*

$$F \leq 4^{\alpha - 1}$$

which is implied by

$$F \leq \frac{(\lfloor \frac{N}{3} \rfloor I + 1)^2}{16}.$$

TABLE II: Rank bounds for generic identifiability ($I = 3$).

	Number of Variables (N)				
	6	10	20	40	80
Triples	4	7	27	105	410
Quadruples	10	36	179	729	2916

TABLE III: Rank bounds for generic identifiability ($N = 6$).

	Alphabet size (I)				
	6	10	20	40	80
Triples	24	40	105	410	1620
Quadruples	45	131	544	2220	8966

Proof: The proof is relegated to Appendix A. ∎

The rank bounds in Theorems 1-2 are nontrivial, albeit far from the maximal attainable rank for the cases considered. Recalling that higher-order tensors are identifiable for higher ranks, a natural question is whether knowledge of four- or higher-dimensional marginals can further enhance identifiability of the complete joint PMF. The next theorem shows that the answer is affirmative.

Theorem 3. *Assume that* $\mathsf{Pr}(i_n|f),\ \forall n \in [N]$ *are drawn from an absolutely continuous distribution, that* $I_1 = \ldots = I_N = I$*, and that the joint PMF* $\mathsf{Pr}(i_1, \ldots, i_N)$ *can be represented using a naive Bayes model of rank* F*. Further assume that* $\mathcal{S} = [N]$ *can be partitioned into* 4 *disjoint subsets denoted by* $\mathcal{S}_1, \ldots, \mathcal{S}_4$ *such that the four-dimensional marginals* $\mathsf{Pr}(i_j, i_k, i_l, i_m),\ \forall j \in \mathcal{S}_1, \forall k \in \mathcal{S}_2, \forall l \in \mathcal{S}_3, \forall m \in \mathcal{S}_4$ *are available. Then, the joint PMF* $\mathsf{Pr}(i_1, \ldots, i_N)$ *is a.s. identifiable if*

$$F \leq I^2 |\mathcal{S}_3||\mathcal{S}_4|,$$
$$2F(F-1) \leq I^2 |\mathcal{S}_1||\mathcal{S}_2|(I|\mathcal{S}_1| - 1)(I|\mathcal{S}_2| - 1).$$

Proof: The proof is relegated to Appendix A. ∎

The conditions of Theorem 3 are satisfied for much higher rank than those of Theorems 1-2 as shown in Tables II-III. The results related to the four-dimensional marginals are obtained following Theorem 3 via checking all possible partitions. The caveat is that one may need many more samples to reliably estimate the four-dimensional marginals. Nevertheless, the theorems that we present in this section offer insights regarding the choice of lower-dimensional marginals to work with – such choice depends on the size of the alphabet of each variable (I) and the number of variables (N) as well as the amount of available data samples.

Remark 2. The above results rely on Lemmas 2, 3 and concern the identifiability of a generic choice of parameters; i.e., the parameters are assumed to be drawn randomly from a jointly continuous distribution. At this point one may wonder whether this is a realistic assumption in practice. For example, in some latent model identification problems a hidden variable has specific physical meaning and an observed variable may not depend on the state of the hidden variable for one or more of its values. Consider a Hidden Markov Model (HMM) where we denote the observed variable at time t as X_t and the hidden state is S_t. The conditional distribution $\widetilde{\mathbf{A}}_t(i, s) := \mathsf{Pr}(X_t = i|S_t = s)$ may be the same for two different values s_1 and s_2 of the hidden state S_t. In such a case, the Kruskal rank of matrix $\widetilde{\mathbf{A}}_t$ would be equal to 1, thereby rendering the deterministic identifiability condition (Lemma 1) useless. Do note, however, that in our setting the latent variable H does not necessarily have a physical interpretation; the CPD is just a convenient 'universal' parametrization of the joint PMF. Therefore the conditional distribution of an observed variable may be the same for two values of the hidden state, but it may still depend on the value of the 'virtual' global latent variable H, and hence recovery of the the joint PMF using lower-dimensional marginals could still be possible.

TABLE IV: Mean relative factor and tensor error when lower-dimensional marginals are perfectly known.

Rank		MRE_{fact}	MRE_{ten}
$F = 5$	Pairs	0.277	0.148
	Triples	1.18×10^{-7}	4.58×10^{-8}
	Quadruples	3.39×10^{-8}	1.19×10^{-8}
$F = 10$	Pairs	0.440	0.187
	Triples	3.58×10^{-7}	8.70×10^{-8}
	Quadruples	1.26×10^{-7}	2.58×10^{-8}
$F = 15$	Pairs	0.466	0.184
	Triples	6.77×10^{-7}	1.52×10^{-7}
	Quadruples	1.78×10^{-7}	3.57×10^{-8}

VII. Numerical Results

In this section, we employ judiciously designed synthetic data simulations to showcase the effectiveness of the proposed joint PMF recovery methods. We also apply the approach to real-data problems such as classification and recommender systems to demonstrate its usefulness in real machine learning tasks.

A. Synthetic-Data Simulations

We first evaluate the proposed approach using synthetic data. We consider a case where $N = 5$ random variables are present, and each variable can take $I_n = 10$ discrete values. We assume that the joint PMF of the 5 random variables can be represented by a naive Bayes model whose latent variable H can take F values, where F is set to be $\{5, 10, 15\}$. We generate matrices $\mathbf{A}_n \in \mathbb{R}_+^{I_n \times F}$, that model the conditional probabilities i.e., $\mathbf{A}_n(i_n, f) = \mathsf{Pr}(i_n|f)$. A vector $\boldsymbol{\lambda} \in \mathbb{R}_+^F$ is also generated for the latent random variable H such that $\boldsymbol{\lambda}(f) = \mathsf{Pr}(f)$. The elements of each $\mathbf{A}_n$ and the vector $\boldsymbol{\lambda}$ are drawn independently from a uniform distribution between zero and one, and each column is normalized to sum to 1. The ground-truth joint PMF is then constructed following the naive Bayes model, i.e., $\mathsf{Pr}(i_1, \ldots, i_5) = \underline{\mathbf{X}}(i_1, \ldots, i_5) = \sum_{f=1}^{F} \boldsymbol{\lambda}(f) \prod_{n=1}^{5} \mathbf{A}_n(i_n, f)$. We assume that the observable data are two-, three- and four-dimensional marginals of the joint PMF. Under such settings, we can verify if the proposed procedure and algorithm can effectively recover the joint PMF, if there is no modeling error and the joint PMF does have low rank. We run 20 Monte Carlo simulations and compute the

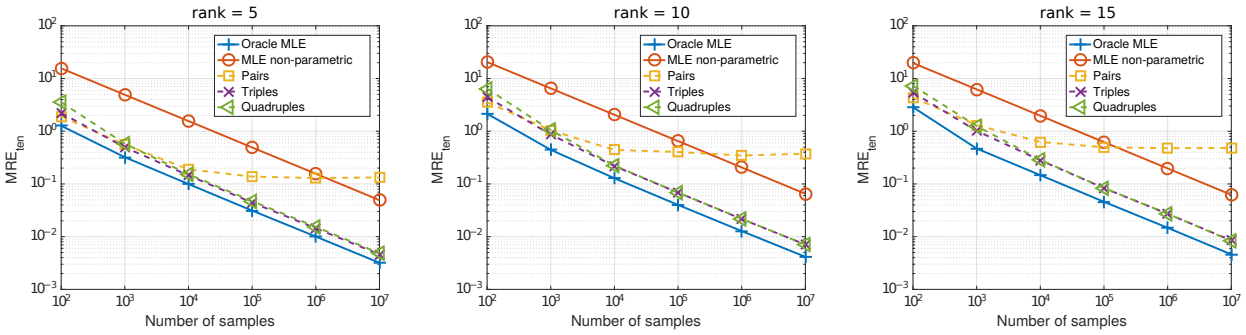


Fig. 7: Mean relative error of the estimated joint PMF under different number of available samples.

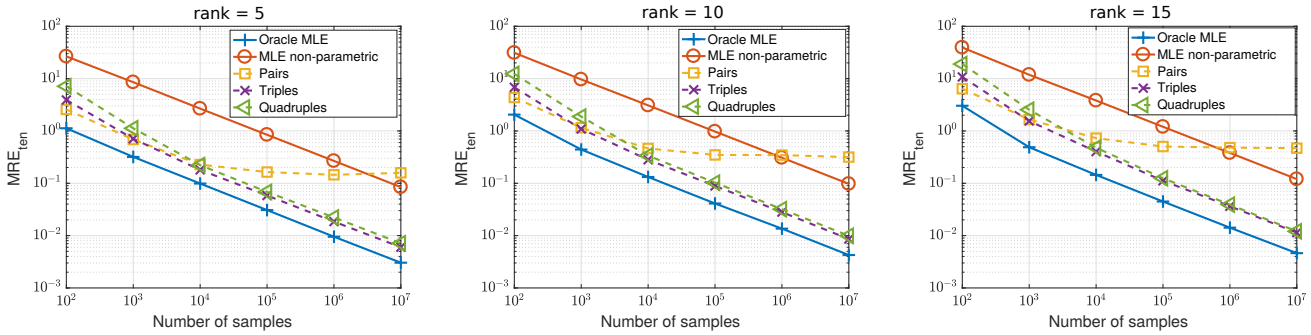


Fig. 8: Mean relative error of the estimated joint PMF under different number of available samples.

mean relative error of the factors as well as the mean relative error of the recovered tensor which are defined as follows

$$\text{MRE}_{\text{fact}} = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{A}_n - \widehat{\mathbf{A}}_n\mathbf{\Pi}\|_F}{\|\mathbf{A}_n\|_F}\right),$$

$$\text{MRE}_{\text{ten}} = \mathbb{E}\left(\frac{\|\underline{\mathbf{X}} - \widehat{\underline{\mathbf{X}}}\|_F}{\|\underline{\mathbf{X}}\|_F}\right),$$

where $\mathbf{\Pi}$ is a permutation matrix to fix the permutation ambiguity, and $\widehat{\underline{\mathbf{X}}}$, $\widehat{\mathbf{A}}_n$ are the estimated joint PMF and the corresponding conditional PMFs.

Table IV shows the mean relative errors for estimating the conditional PMFs and joint PMF using the different types of input under different choices of rank. Consistent with our analysis, one can see that using marginal distributions of triples or quadruples of random variables (i.e., three- and four-dimensional marginals) we are able to recover the joint PMF of 5 random variables. Here recovery with high accuracy has been demonstrated; exact recovery is also possible in certain cases, see [14]. However, using pairs (i.e., two-dimensional marginals) is not as promising. Recall that our identifiability result is built upon the identifiability of third- and higher-order CPD models. These nice identifiability results in general do not hold for matrices – which explains the sharp performance difference between using the pairs and higher-dimensional marginals. Pairs can work, however, when the conditional probability matrices are sufficiently sparse, and under more stringent constraints on the rank F. We defer detailed discussion to a follow up paper, due to lack of space in this one.

The above simulation serves as sanity check – if the available data and the model perfectly match with each other and we have noiseless marginal distributions, the proposed approach can indeed recover the joint PMF. In practice, we usually do not have exact estimates of the lower-dimensional marginal distributions. Next, we provide a set of more realistic simulations where we estimate the marginal PMFs using sample averages from the observed data.

1) Fully-observed data: We follow the same way of generating the ground-truth joint PMF as in the previous simulation. Then, drawing from the joint PMF, we generate a synthetic dataset of M five-dimensional data points. The data is generated as follows: for each data point, we first draw a sample h_m according to $\boldsymbol{\lambda}$; i.e., a realization of the hidden variable H. Then the data point (vector) $\mathbf{s}_m = [\mathbf{s}_m(1), \ldots, \mathbf{s}_m(N)]^T$ is generated by drawing its elements independently from $\{\mathbf{A}_n\}(:, h_m)_{n=1}^N$, i.e., $\mathbf{s}_m(n)$ is drawn from $\{\mathbf{A}_n\}(:, h_m)$. This is equivalent to synthesizing the five-way joint PMF tensor and drawing an outcome from it (cf. the naive Bayes interpretation).

We use the generated 5-dimensional data points to estimate lower-dimensional marginals and run our ADMM algorithm to recover the full joint PMF. We repeat for a total of 10 Monte Carlo simulations. Figure 7 shows the tensor mean relative error of the estimated joint PMF under different dataset sizes M. We also include the performance of two additional methods for estimating the joint PMF. Given the full data together with 'oracle' observations of the hidden variable H, we perform Maximum Likelihood Estimation (MLE) of the naive Bayes parameters, denoted as oracle MLE, which is done simply by frequency counting

$$\mathsf{Pr}(f)_{ML} = \frac{\text{count}(f)}{M},\ \mathsf{Pr}(i_n|f)_{ML} = \frac{\text{count}(i_n, f)}{\text{count}(f)},$$

TABLE V: Misclassification error on different UCI datasets.

	Binary					Multiclass	
Method	INCOME	CREDIT	HEART	MUSHROOM	VOTES	CAR	NURSERY
CP (Pairs)	0.177±0.004	0.134±0.019	0.151±0.023	0.010±0.007	0.046±0.024	0.128±0.021	0.101±0.009
CP (Triples)	0.175±0.003	0.129±0.018	0.147±0.031	0.006±0.002	0.043±0.024	0.089±0.016	0.069±0.011
CP (Quadruples)	**0.171**±0.003	**0.123**±0.018	**0.138**±0.029	0.002±0.001	0.042±0.020	0.074±0.015	0.061±0.007
SVM (Linear)	0.179±0.004	0.146±0.027	0.170±0.053	**0**±0	**0.038**±0.025	0.065±0.006	0.063±0.004
SVM (RBF)	0.174±0.004	0.136±0.018	0.187±0.055	**0**±0	0.079±0.024	**0.026**±0.008	**0.006**±0.001
Naive Bayes	0.209±0.005	0.140±0.018	0.166±0.026	0.044±0.005	0.096±0.022	0.151±0.016	0.097±0.007

TABLE VI: Dataset information.

Dataset	N	F
INCOME	8	$[1, 20]$
CREDIT	9	$[1, 20]$
HEART	9	$[1, 10]$
MUSHROOM	22	$[1, 20]$
VOTES	17	$[1, 10]$
CAR	7	$[1, 15]$
NURSERY	9	$[1, 15]$

where $\mathrm{count}(i_n, f)$ is the number of times that $X_n = i_n$ and $H = f$ appear together in the dataset and $\mathrm{count(f)}$ the number of times H takes the value f. We also include the MLE of a non-parametric approach in which we use the empirical 5-dimensional distribution as our estimate i.e.,

$$\mathsf{Pr}(i_1, \ldots, i_N)_{ML} = \frac{\mathrm{count}(i_1, \ldots, i_N)}{M}.$$

One can see that the estimation performance of our method is similar under different rank values and approaches that of oracle MLE using three- and four-dimensional marginals. In addition, as expected, the recovery accuracy steadily improves as the size of the available dataset increases. On the other hand, when using two-dimensional marginals the performance improves until it reaches a plateau at approximately $M = 10^5$. The ability to recover the joint PMF using pairs of random variables is obviously limited by identifiability of matrix factorization.

2) Missing data: We repeat the above experiment when some of the dataset entries are missing. We randomly hide 20% of the data and compute estimates of two- three- and four-dimensional marginals using only the available data. We run the ADMM-based algorithm and repeat for 10 Monte Carlo simulations. The estimation performance of our method is again similar under different rank values and approaches that of oracle MLE. As expected, we observe a slight decrease in performance which is due to the less accurate estimation of the lower-dimensional marginals. In this case, the MLE non-parametric method takes into account only the fully observed samples.

Note that when empirical estimates of the lower-dimensional distributions are used and the number of samples is limited, three-dimensional distributions may give lower relative error compared to the four-dimensional ones. This shows that in some cases using lower-dimensional distributions can be more beneficial than higher-order ones in terms of parameter estimation accuracy. Actually, this is not very surprising since it can be shown that empirical lower-dimensional marginals are always more accurate than higher-dimensional ones when estimated given the same data [32].

B. Real-Data Experiments

In real applications, the ground-truth joint PMF and the conditional PMFs are not known. Nevertheless, we can evaluate the method on a variety of standard machine learning tasks to observe its effectiveness. In this subsection, we test the proposed approach on two different tasks, namely, data classification and recommender systems. Note that both tasks can be easily accomplished if the joint PMF of pertinent variables (e.g., features and labels in classification) is known and thus are suitable for evaluating our method. Note that the rank of the joint PMF tensor, or, $F = |\mathcal{H}|$, cannot be known as in the simulations. Fortunately, this is a single discrete variable that can be easily tuned, e.g., via observing validation errors as in machine learning.

1) Classification Task: We evaluate the performance of our approach on 7 different datasets from the UCI machine learning repository [33]. Five of the selected datasets correspond to binary classification and two to multi-class classification. For each dataset, we represent the training samples using its discrete features so that the PMF-based approach can be applied. We split each dataset such that 70% of the data samples is used for training, 10% used for validation and 20% for testing.

For each dataset, we let X_N be the label and $X_1, \ldots, X_{N-1}$ be the selected features. We estimate lower-dimensional marginal distributions of pairs, triples and quadruples of variables using the samples in the training set. Then, we use the marginals to estimate $\mathsf{Pr}(i_n|f)$, and $\mathsf{Pr}(f)$. After applying the proposed approach for estimating the joint PMF, we predict for each data point of the test set the corresponding label using the MAP rule. The MAP estimator of the label $l(\mathbf{s}_m)$ of the m-th observation $\mathbf{s}_m = [\mathbf{s}_m(1), \ldots, \mathbf{s}_m(N-1)]^T$ in the test set can be written as

$$\widehat{l}_{\text{map}}(\mathbf{s}_m) = \underset{i_N \in \{1,\ldots,I_N\}}{\arg\max} \ \mathsf{Pr}(i_N|\mathbf{s}_m(1), \ldots, \mathbf{s}_m(N-1)),$$

where I_N is the number of classes. Equivalently, using the Bayes rule the above can be found by

$$\widehat{l}_{\text{map}}(\mathbf{s}_m) = \underset{i_N \in \{1,\ldots,I_N\}}{\arg\max} \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_N|f) \prod_{n=1}^{N-1} \mathsf{Pr}(\mathbf{s}_m(n)|f).$$

For each dataset, we run 10 Monte Carlo simulations with randomly partitioned training/validation/test sets and observe the average result. As mentioned, we do not know *a priori*

TABLE VII: RMSE and MAE of different algorithms on MovieLens (Ratings are in the range [1-5]).

	MovieLens Dataset 1		**MovieLens Dataset 2**		**MovieLens Dataset 3**	
Method	RMSE	MAE	RMSE	MAE	RMSE	MAE
CP (Pairs)	0.802±0.003	0.608±0.003	0.795±0.002	0.611±0.002	0.897±0.003	0.702±0.002
CP (Triples)	0.783±0.002	0.591±0.002	**0.785 ±0.002**	**0.599 ±0.002**	0.887±0.002	0.691±0.002
CP (Quadruples)	**0.778±0.002**	**0.588±0.002**	0.786±0.002	0.600±0.002	**0.884 ±0.002**	**0.689 ±0.002**
Global Average	0.945±0.001	0.693±0.001	0.906±0.002	0.653±0.002	0.996±0.002	0.798±0.001
User Average	0.879±0.002	0.679±0.001	0.830±0.003	0.625±0.002	1.010±0.002	0.768±0.002
Movie Average	0.886±0.002	0.705±0.001	0.889±0.002	0.673±0.002	0.942±0.002	0.754±0.001
BMF	0.797±0.002	0.623±0.002	0.792±0.002	0.604±0.002	0.904±0.003	0.701±0.003

what is an appropriate rank for our model. Therefore, for each dataset, we fit models of different rank values and choose the one which minimizes the misclassification error of the validation set as in standard machine learning practice.

We use 3 different classical classifiers from the MATLAB Statistics and Machine Learning Toolbox as baselines; linear SVM, kernel SVM with radial basis function and a naive Bayes classifier. For SVM classifiers, we use both the original data encoding as well as the one-hot encoding which usually is more suitable for discrete data and report the best result among the two. Note that the baseline naive Bayes approach is very different from ours: the baseline method assumes that the features are independent given the label, while we assume that the label and the features are independent given an unknown latent variable. The former is a very strong assumption that is rarely satisfied by real data, but our assumption holds for an arbitrary set of random variables provided F is large enough, as we showed in Proposition 1.

Table V shows the classification errors obtained on the datasets. One can see that our approach outperforms the naive Bayes classifier which assumes that the features are independent given the label. Several observations are in order. First, using higher-dimensional marginals, the proposed approach gives better classification results compared to using lower-dimensional ones. This is consistent with our analysis in Sec. VI – higher-dimensional marginals lead to stronger overall identifiability of the joint PMF. One can see that for all the datasets under test, using four-dimensional marginal distributions gives the best classification accuracy compared to the three- and two-dimensional ones. Second, for the five binary classification experiments, the proposed method works better than (on three datasets) or comparable to the baselines. This is quite surprising since our method does not directly optimize a classification criterion as SVM does. The result suggests that the proposed method indeed captures the essence of the joint distribution and the recovered joint PMF can be utilized to make inference in practice. Third, for the multiclass datasets, the proposed method yields accuracy that is less than the SVM methods. This also makes sense: when X_N has a value set whose cardinality grows from 2 to 5, the joint PMFs of X_N and X_i, X_j for $i, j < N$ require more samples to estimate accurately. This also shows an interesting sample complexity-accuracy trade-off of the proposed method. Nevertheless, the method still works comparably well with the linear SVM, which supports the usefulness of the joint PMF estimation method.

2) Recommender Systems: We also evaluate the method for the task of recommender systems using the MovieLens dataset [34]. MovieLens is a dataset that contains ratings on 5-star scale, with half-star increments, by a number of users. In order to test our algorithm we select three different subsets of the full dataset and round the ratings to the next integer. Initially, three different categories (action, animation and romance) are selected. From each category we extract a user-by-rating submatrix by keeping the 20 most rated movies and form the 3 datasets for our experiments. Note that the constructed three datasets have many missing values, since not all users watched and rated all movies. The task of recommender systems is to recommend unwatched movies to users based on prediction of the user's rating given the available data.

We aim at estimating the joint PMF of the movie ratings. In this case, each random variable X_n represents a movie, and it takes values from $\{1, \ldots, 5\}$, i.e., the ratings. Consequently, the joint PMF is a twenty-way tensor which has 5^{20} elements. The 3 partially observed datasets are used in order to estimate lower-dimensional marginal distributions of pairs, triples and quadruples of the variables (movies). We use the estimated PMF to compute the expected value of users' ratings that we do not observe given the ones we observe. More specifically, let $\mathbf{s}_m = [\mathbf{s}_m(1), \ldots, \mathbf{s}_m(N)]$ be the ratings of the m-th user and $\mathbf{s}_m(N) = 0$ i.e., the user has a missing rating. The conditional expectation of the movie's rating is given by

$$\widehat{s}_N = \sum_{i_N=1}^{I_N} i_N \mathsf{Pr}(i_N | \mathbf{s}_m(1), \ldots, \mathbf{s}_m(N-1)).$$

As a baseline algorithm, we use the Biased Matrix Factorization (BMF) method [2], which is a commonly used method in recommender systems. The BMF method is essentially low-rank matrix completion with modifications. Additionally, we present results obtained by global average of the ratings, the user average, and the item average as baselines for predicting the missing entries. For each dataset we randomly hide 20% ratings that we use as a test set, 10% ratings that we use as a validation set and the remaining dataset is used as a training set. We run 10 Monte Carlo simulations using our approach and the BMF algorithm. We select the parameters of both methods based on the RMSE of the validation set.

Table VII shows the performance of the two algorithms in terms of the RMSE and Mean Absolute Error (MAE). One can see that, for the three datasets under test, the proposed method and the BMF method output clearly lower RMSEs and MAEs relative to the naive methods using averaging. In addition, the

proposed method slightly outperforms BMF on all of the three datasets. Note that BMF is considered a state-of-art method for movie recommendation, and it incorporates application-specific custom features, such as user bias and movie bias to achieve good performance. The proposed method, on the other hand, only uses basic probability to handle the same task – it is completely application-blind. This suggests that the joint PMF modeling and the proposed algorithm are both quite effective. Last, we also observe accuracy improvement when we increase the dimension of the marginal distributions used in the approach. Again, this performance may come from the identifiability gain as we analyzed in Theorem 3.

VIII. Conclusions

In this work, we have taken a fresh look at one of the most fundamental problems in statistical learning – joint PMF estimation. Due to the curse of dimensionality, naive count-based estimation is mission impossible in most cases. One popular approach has historically been to assume a plausible structural model, such as a Markov chain, tree, or other probabilistic graphical model, and do inference using this model. We showed that a very different 'non-parametric' tensor-based approach is possible, and it features several key benefits. Foremost among them is guaranteed identifiability of the high-dimensional joint PMF from low-dimensional marginals, which can be reliably estimated using counting from much fewer examples, even if there are (many) samples missing from each example. This ability to infer a unique higher-dimensional joint PMF by specifying lower-dimensional ones is reminiscent of Kolmogorov extension, which is intuitively very pleasing.

We have also proven two more results that appear fundamental and close to the heart of probability and learning theory: i) *every* joint PMF can be *interpreted* as a naive Bayes model; and ii) probabilistic graphical models, which are very popular in statistical and computer science, are never identifiable if one simply limits the number of hidden nodes; one needs to bound the number of hidden states as well. Our non-parametric approach can reveal the true hidden structure, instead of assuming it; and this alleviates the risk of up-front bias in the analysis.

On the practical side, our approach is appealing since lower-dimensional marginals can be more reliably estimated from a limited amount of partially observable data. We have also provided a practical and easily implementable algorithm that is based on factor-coupled tensor factorization to handle the recovery problem. Simulations and judicious experiments with real data have shown that the performance of the proposed approach is consistent with our analysis, and approaches or exceeds that of state-of-art application-specific solutions that have come out after years of intensive research, which is satisfying.

Appendix A
Identifiability Results

A. Proof of Theorem 1

Each three-dimensional marginal satisfies $\underline{\mathbf{X}}_{jkl} = [\![\boldsymbol{\lambda}, \mathbf{A}_j, \mathbf{A}_k, \mathbf{A}_l]\!]$, where $\mathbf{A}_l \geq \mathbf{0}$, $\mathbf{A}_k \geq \mathbf{0}$, $\mathbf{A}_j \geq \mathbf{0}$, $\mathbf{1}^T\mathbf{A}_l = \mathbf{1}^T$, $\mathbf{1}^T\mathbf{A}_k = \mathbf{1}^T$, $\mathbf{1}^T\mathbf{A}_j = \mathbf{1}^T$, $\boldsymbol{\lambda} > \mathbf{0}$, $\mathbf{1}^T\boldsymbol{\lambda} = 1$. Consider a partition of the set $\mathcal{S} = [N]$ into three disjoint sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and define the following factors

$$\begin{aligned}\widehat{\mathbf{A}}_1 &= [\mathbf{A}_{u_1}^T, \cdots, \mathbf{A}_{u_{|\mathcal{S}_1|}}^T]^T \\ \widehat{\mathbf{A}}_2 &= [\mathbf{A}_{v_1}^T, \cdots, \mathbf{A}_{v_{|\mathcal{S}_2|}}^T]^T \\ \widehat{\mathbf{A}}_3 &= [\mathbf{A}_{w_1}^T, \cdots, \mathbf{A}_{w_{|\mathcal{S}_3|}}^T]^T\end{aligned} \tag{18}$$

with $u_t \in \mathcal{S}_1$, $v_t \in \mathcal{S}_2$, $w_t \in \mathcal{S}_3$. Then, we can construct a single virtual nonnegative CPD model

$$\widehat{\underline{\mathbf{X}}}^{(1)} = (\widehat{\mathbf{A}}_3 \odot \widehat{\mathbf{A}}_2)\text{diag}(\boldsymbol{\lambda})\widehat{\mathbf{A}}_1^T, \tag{19}$$

where $\widehat{\mathbf{A}}_1 \in \mathbb{R}_+^{I|\mathcal{S}_1| \times F}$, $\widehat{\mathbf{A}}_2 \in \mathbb{R}_+^{I|\mathcal{S}_2| \times F}$, $\widehat{\mathbf{A}}_3 \in \mathbb{R}_+^{I|\mathcal{S}_3| \times F}$ and $\widehat{\underline{\mathbf{X}}} \in \mathbb{R}_+^{I|\mathcal{S}_1| \times I|\mathcal{S}_2| \times I|\mathcal{S}_3|}$. We have used a subset of the available information to synthesize a virtual single nonnegative CPD model of size $I_1 \times I_2 \times I_3$, with $I_k := I|\mathcal{S}_k|$. Therefore, we can apply identifiability results of three-way tensors. We observe that the sizes of the different modes of the virtual tensor depend on the way we partition the variables. We distinguish between two cases and apply Lemma 2.

1) $(N \leq I)$: We partition the variables into three sets such that $I_1 = I, I_2 = I$ and $I_3 = (N-2)I$. Clearly we have that $I_3 > I_2, I_1$ and $I_3 < (I_1 - 1)(I_2 - 1)$. According to Lemma 2 tensor $\widehat{\underline{\mathbf{X}}}$ admits unique decomposition for $F \leq \min\left(I_3, (I_1-1)(I_2-1)\right) = (N-2)I$.
2) $(N > I)$: In this case we can partition the variables into three sets such that $I_3 = (I_1 - 1)(I_2 - 1)$. We can always have that $I_1 = I_2$. The equality is satisfied when $|\mathcal{S}_1| = |\mathcal{S}_2| = \frac{\sqrt{(NI-1)}}{I}$. According to Lemma 2 tensor $\widehat{\underline{\mathbf{X}}}$ admits unique decomposition for $F \leq \min(I_3, (I_1-1)(I_2-1)) \leq (\lfloor \frac{\sqrt{(NI-1)}}{I} \rfloor I - 1)^2$.

B. Proof of Theorem 2

As above, but this time choosing $|\mathcal{S}_1| = |\mathcal{S}_2| = \lfloor \frac{N}{3} \rfloor$, $|S3| = N - 2|S1|$ and invoking Lemma 3.

C. Proof of Theorem 3

Consider a partitioning the variables into four disjoint sets similar to the three-way case. We obtain a single nonnegative CPD of the following form

$$\underline{\mathbf{X}}^{(1)} = (\widehat{\mathbf{A}}_4 \odot \widehat{\mathbf{A}}_3 \odot \widehat{\mathbf{A}}_2)\text{diag}(\boldsymbol{\lambda})\widehat{\mathbf{A}}_1^T,$$

which is a fourth-order tensor $\widehat{\underline{\mathbf{X}}} \in \mathbb{R}_+^{I|\mathcal{S}_1| \times I|\mathcal{S}_2| \times I|\mathcal{S}_3| \times I|\mathcal{S}_4|}$. A fourth-order tensor can be viewed as a third-order tensor of size $I|\mathcal{S}_1| \times I|\mathcal{S}_2| \times I^2|\mathcal{S}_4||\mathcal{S}_3|$ with a specially structured factor matrix. We can write the mode-1 unfolding of the tensor $\widehat{\underline{\mathbf{X}}}$ as

$$\widehat{\underline{\mathbf{X}}}^{(1)} = (\bar{\mathbf{A}}_3 \odot \widehat{\mathbf{A}}_2)\text{diag}(\boldsymbol{\lambda})\widehat{\mathbf{A}}_1^T, \tag{20}$$

where $\bar{\mathbf{A}}_3 = \widehat{\mathbf{A}}_4 \odot \widehat{\mathbf{A}}_3$. Lemmas 2-3 cannot be applied in this case because of the Khatri-Rao structure of one factor. We will use the following result

Lemma 4. *[35] Let* $\underline{\mathbf{X}} = [\![\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, *where* $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$. *If rank*$(\mathbf{A}_3) = F$ *and*

rank$\big(\mathbf{M}_2(\mathbf{A}_1) \odot \mathbf{M}_2(\mathbf{A}_2)\big) = \binom{F}{2}$*, then rank*$(\underline{\mathbf{X}}) = F$ *and the decomposition is essentially unique.*

$\mathbf{M}_2(\mathbf{A})$ denotes the $\binom{I}{2} \times \binom{F}{2}$ compound matrix containing all 2×2 minors of $\mathbf{A}$. The generic version of Lemma 4 states that if $F \leq I_3$ and $2F(F-1) \leq I_1(I_1-1)I_2(I_2-1)$, then the decomposition of $\underline{\mathbf{X}}$ is essentially unique, a.s. [36]. We know that a Khatri-Rao product of two matrices is full rank almost surely [37, Corollary 1]. For the three-way tensor in (20) we have $I_1 = I|\mathcal{S}_1|$, $I_2 = I|\mathcal{S}_2|$ and $I_3 = I^2|\mathcal{S}_3||\mathcal{S}_4|$. Applying the generic version of Lemma 4 we obtain the desired result.

APPENDIX B
ALGORITHM

We reformulate optimization problem (16) by introducing an auxiliary variable $\widehat{\mathbf{A}}_j$. The problem can be equivalently written as

$$\begin{aligned} &\min_{\mathbf{A}_j, \widehat{\mathbf{A}}_j} \; f(\widehat{\mathbf{A}}_j) + r(\mathbf{A}_j) \\ &\text{subject to } \mathbf{A}_j = \widehat{\mathbf{A}}_j^T \end{aligned} \tag{21}$$

where

$$f(\widehat{\mathbf{A}}_j) = \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \frac{1}{2} \left\| \mathbf{X}^{(1)}_{jkl} - (\mathbf{A}_l \odot \mathbf{A}_k) \text{diag}(\boldsymbol{\lambda}) \widehat{\mathbf{A}}_j \right\|_F^2,$$

$r(\mathbf{A})$ is the indicator function for the probability simplex constraints $\mathcal{C} = \{\mathbf{A} \mid \mathbf{A} \geq 0, \mathbf{1}^T\mathbf{A} = \mathbf{1}^T\}$

$$r(\mathbf{A}) = \begin{cases} 0, & \mathbf{A} \in \mathcal{C} \\ \infty, & \mathbf{A} \notin \mathcal{C} \end{cases}.$$

Optimization problem (21) can be readily solved by applying the ADMM algorithm. We solve for $\mathbf{A}_j$ by performing the following updates

$$\begin{aligned} \widehat{\mathbf{A}}_j^{(\tau+1)} &= (\mathbf{G}_j + \rho\mathbf{I})^{-1} \left(\mathbf{V}_j + \rho \left(\mathbf{A}_j^{(\tau)} + \mathbf{U}_j^{(\tau)} \right)^T \right), \\ \mathbf{A}_j^{(\tau+1)} &= \mathcal{P}_{\mathcal{C}} \left(\mathbf{A}_j^{(\tau)} - \widehat{\mathbf{A}}_j^{(\tau+1)T} + \mathbf{U}_j^{(\tau)} \right), \\ \mathbf{U}_j^{(\tau+1)} &= \mathbf{U}_j^{(\tau)} + \mathbf{A}_j^{(\tau+1)} - \widehat{\mathbf{A}}_j^{(\tau+1)T}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{G}_j &= (\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \circledast \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \mathbf{Q}_{lk}^T \mathbf{Q}_{lk}, \\ \mathbf{V}_j &= \text{diag}(\boldsymbol{\lambda}) \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \mathbf{Q}_{lk}^T \mathbf{X}^{(1)}_{jkl}, \\ \mathbf{Q}_{lk} &= \mathbf{A}_l \odot \mathbf{A}_k. \end{aligned}$$

$\mathcal{P}_{\mathcal{C}}$ is the projection operator onto the convex set $\mathcal{C}$ and τ denotes the iteration index. Various methods exist for projecting onto the probability simplex, e.g., see [38]. Note that in order to efficiently compute matrix $\mathbf{G}_j$ we use a property of the Khatri-Rao product; $\mathbf{Q}_{lk}^T\mathbf{Q}_{lk} = (\mathbf{A}_l^T\mathbf{A}_l) \circledast (\mathbf{A}_k^T\mathbf{A}_k)$. Efficient algorithms also exist for the computation of matrix $\mathbf{V}_j$ which is a sum of Matricized Tensor Times Khatri-Rao Product (MTTKRP) terms [39], [40]. Similarly, we can derive updates for $\boldsymbol{\lambda}$. At each ADMM iteration we perform the following updates

$$\begin{aligned} \widehat{\boldsymbol{\lambda}}^{(\tau+1)} &= (\mathbf{G} + \rho\mathbf{I})^{-1} \left(\mathbf{V} + \rho \left(\boldsymbol{\lambda}^{(\tau)} + \mathbf{u}^{(\tau)} \right) \right), \\ \boldsymbol{\lambda}^{(\tau+1)} &= \mathcal{P}_{\mathcal{C}} \left(\boldsymbol{\lambda}^{(\tau)} - \widehat{\boldsymbol{\lambda}}^{(\tau+1)} + \mathbf{u}^{(\tau)} \right), \\ \mathbf{u}^{(\tau+1)} &= \mathbf{u}^{(\tau)} + \boldsymbol{\lambda}^{(\tau+1)} - \widehat{\boldsymbol{\lambda}}^{(\tau+1)}. \end{aligned}$$

where

$$\begin{aligned} \mathbf{G} &= \sum_j \sum_{k>j} \sum_{l>k} \mathbf{Q}_{lkj}^T \mathbf{Q}_{lkj}, \\ \mathbf{V} &= \sum_j \sum_{k>j} \sum_{l>k} \mathbf{Q}_{lkj}^T \text{vec}(\underline{\mathbf{X}}_{jkl}), \\ \mathbf{Q}_{lkj} &= \mathbf{A}_l \odot \mathbf{A}_k \odot \mathbf{A}_j. \end{aligned}$$

REFERENCES

[1] N. Kargas and N. D. Sidiropoulos, "Completing a joint PMF from projections: a low-rank coupled tensor factorization approach," in *Proc. IEEE ITA*, Feb. 2017.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[3] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[4] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1257–1264.

[5] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[6] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sep. 1970.

[7] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers Phonetics*, vol. 16, pp. 1–84, 1970.

[8] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1193–1207, Aug. 2010.

[9] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, Mar. 2000.

[10] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.

[11] X. Fu, N. D. Sidiropoulos, J. H. Tranter, and W.-K. Ma, "A factor analysis framework for power spectra separation and multiple emitter localization," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6581–6594, Aug. 2015.

[12] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2773–2832, Aug. 2014.

[13] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, "A tensor approach to learning mixed membership community models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2239–2312, Jan. 2014.

[14] 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, Jul. 2017.

[15] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[16] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[17] L. Chiantini and G. Ottaviani, "On generic identifiability of 3-tensors of small rank," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 3, pp. 1018–1037, 2012.

[18] I. Domanov and L. D. Lathauwer, "Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 4, pp. 1567–1589, 2015.

[19] Y. Qi, P. Comon, and L.-H. Lim, "Semialgebraic geometry of nonnegative tensor rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 4, pp. 1556–1580, 2016.

[20] N. L. Zhang, "Hierarchical latent class models for cluster analysis," *J. Mach. Learn. Res.*, vol. 5, no. 6, pp. 697–723, Jun. 2004.

[21] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 841–848.

[22] D. Lowd and P. Domingos, "Naive Bayes models for probability estimation," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 529–536.

[23] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[24] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 792–799.

[25] L.-H. Lim and P. Comon, "Nonnegative approximations of nonnegative tensors," *J. Chemometr.*, vol. 23, no. 7-8, pp. 432–441, Jul. 2009.

[26] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

[27] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2005, pp. 601–602.

[28] A. Beutel, P. P. Talukdar, A. Kumar, C. Faloutsos, E. E. Papalexakis, and E. P. Xing, "FlexiFact: Scalable flexible factorization of coupled tensors on hadoop," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 109–117.

[29] E. Acar, T. G. Kolda, and D. M. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," in *Proc. KDD Workshop Min. Learn. Graphs*, 2011.

[30] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5052–5065, 2016.

[31] X. Fu, K. Huang, W.-K. Ma, N. D. Sidiropoulos, and R. Bro, "Joint tensor factorization and outlying slab suppression with applications," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6315–6328, 2015.

[32] K. Huang, X. Fu, and N. D. Sidiropoulos, "Learning Hidden Markov Models from Pairwise Co-occurrences with Applications to Topic Modeling," *ArXiv preprint arXiv:1802.06894*, Feb. 2018.

[33] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[34] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Inter. Intel. Systems*, vol. 5, no. 4, pp. 1–19, Dec. 2016.

[35] T. Jiang and N. D. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2625–2636, Sep. 2004.

[36] A. Stegeman, J. M. T. Berge, and L. D. Lathauwer, "Sufficient conditions for uniqueness in CANDECOMP/PARAFAC and INDSCAL with random component matrices," *Psychometrika*, vol. 71, no. 2, pp. 219–229, 2006.

[37] T. Jiang, N. D. Sidiropoulos, and J. M. F. ten Berge, "Almost-sure identifiability of multidimensional harmonic retrieval," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1849–1859, Sep. 2001.

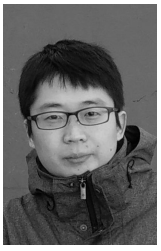
[38] W. Wang and M. A. Carreira-Perpinán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," *ArXiv preprint arXiv:1309.1541*, 2013.

[39] B. W. Bader and T. G. Kolda, "Efficient matlab computations with sparse and factored tensors," *SIAM J. Sci. Comp.*, vol. 30, no. 1, pp. 205–231, 2008.

[40] S. Smith, N. Ravindran, N. D. Sidiropoulos, and G. Karypis, "SPLATT: Efficient and parallel sparse tensor-matrix multiplication," in *IEEE Inter. Par. Dist. Process. Symposium*, May 2015, pp. 61–70.

Nikos Kargas received the Diploma and M.Sc. degrees in electronic and computer engineering from the Technical University of Crete (TUC), Greece, in 2013 and 2015, respectively. Since 2015, he has been working towards his Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota. He is affiliated with the Signal and Tensor Analytics Research group under the supervision of Professor Nikos Sidiropoulos. His research interests include signal processing, machine learning, and optimization.

Nicholas D. Sidiropoulos (F'09) received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Greece, and M.S. and Ph.D. degrees in Electrical Engineering from the University of Maryland at College Park, in 1988, 1990 and 1992, respectively. He served as assistant professor at the University of Virginia, associate professor at the University of Minnesota, professor at TU Crete, Greece and the University of Minnesota, and is currently professor and chair of the ECE Department at the University of Virginia. His research spans topics in signal processing theory and algorithms, optimization, communications, and factor analysis — with a long-term interest in tensor decomposition and its applications. His current focus is primarily on signal and tensor analytics for learning from big data. He received the NSF/CAREER award in 1998, and the IEEE Signal Processing (SP) Society Best Paper Award in 2001, 2007, and 2011. He served as IEEE SP Society Distinguished Lecturer (2008-2009), and as Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007-2008). He received the 2010 IEEE SP Society Meritorious Service Award, and the 2013 Distinguished Alumni Award from the Dept. of ECE, University of Maryland. He is a Fellow of IEEE (2009) and a Fellow of EURASIP (2014).



Xiao Fu (S'12-M'15) is an Assistant Professor in the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, United States. He received his Ph.D. degree in Electronic Engineering from The Chinese University of Hong Kong (CUHK), Hong Kong, 2014. He was a Postdoctoral Associate in the Department of Electrical and Computer Engineering, University of Minnesota - Twin Cities, Minneapolis, MN, United States, from 2014 to 2017. His research interests include the broad area of signal processing and machine learning, with a recent emphasis on tensor/matrix factorization. He received a Best Student Paper Award at ICASSP 2014 and was a finalist of the Best Student Paper Competition at IEEE SAM 2014. He also co-authored a paper that received a Best Student Paper Award at IEEE CAMSAP 2015.