Recovering Joint Probability from Pairwise Marginals

Shahana Ibrahim
Joint Work with Dr. Xiao Fu

Feb 2020

School of Electrical Engineering and Computer Science,
Oregon State University, Corvallis

Contents

- Motivation.
- Existing approach.
- Proposed method and our contributions.
- Theoretical analysis.
- Experimental results.
- Conclusion.

Joint PMF Learning

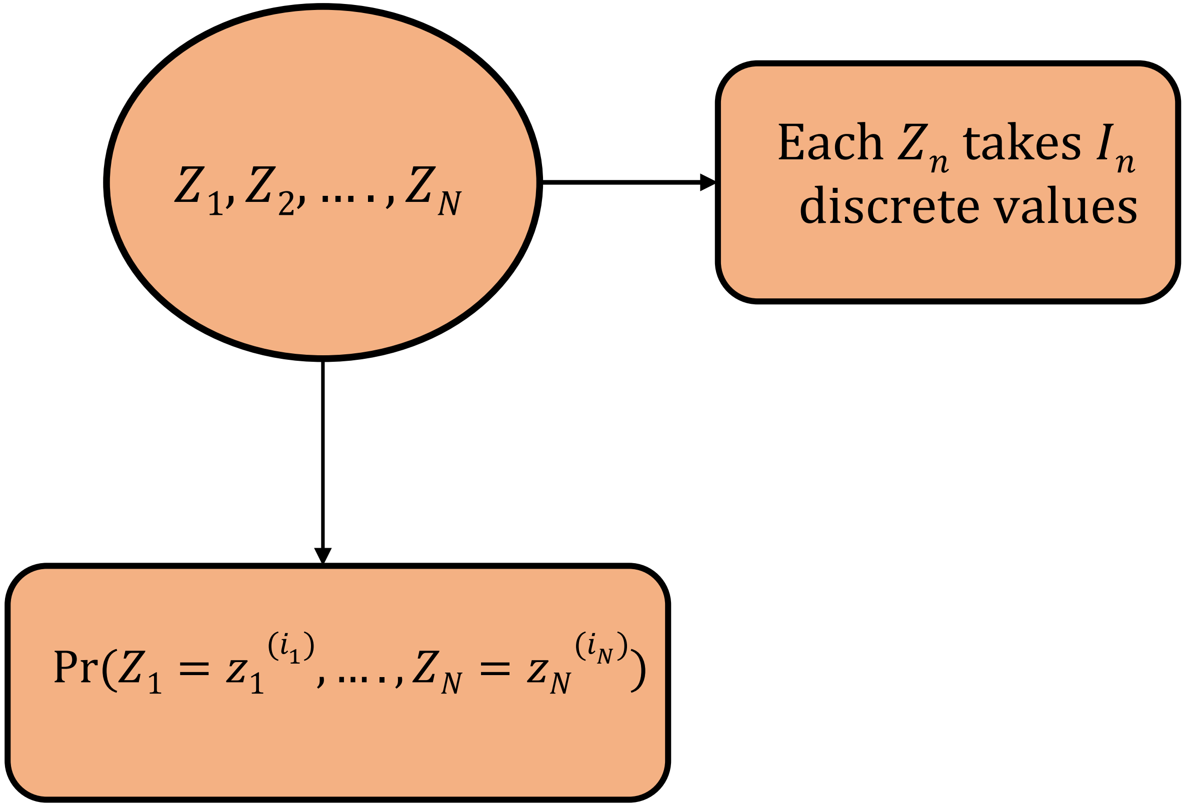
- Joint probability mass function (PMF) is considered as the **'gold standard'** in statistical machine learning.

- Joint PMF estimation has numerous applications:
  - **recommender systems**
  - **classification tasks**
  - **crowdsourcing**
  - **survey/database completion**

- In these applications, we are given with partial observations of the random variables.

- Knowing the joint PMF of the random variables can help us us predicting the missing data.

Joint PMF of N Random Variables



- Short hand notation for $\Pr(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)})$ is $\Pr(i_1, \ldots, i_N)$

Challenges in Joint PMF Learning

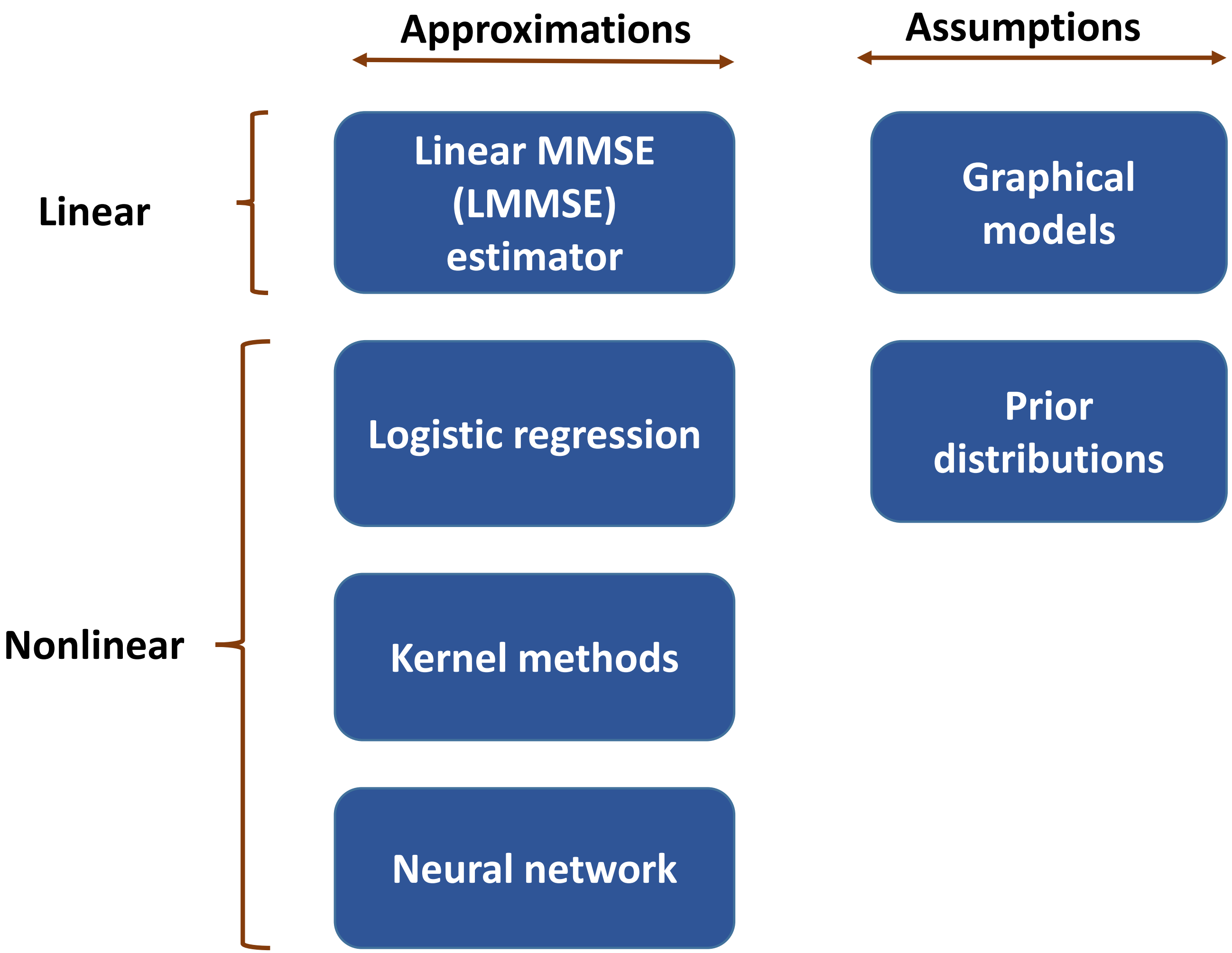
- Suppose we have 10 random variables each taking 10 different values.
- Then joint probability of these 10 random variables have 10^{10} entries!!!
- The 'naive' approach for joint PMF estimation is counting the occurences of the joint variable realizations which means we require $S \gg 10^{10}$ examples for a reasonable accuracy.
- This makes the 'naive' approach very inaccurate.

Challenges in Joint PMF Learning

- Suppose we have 10 random variables each taking 10 different values.
- Then joint probability of these 10 random variables have 10^{10} entries!!!
- The 'naive' approach for joint PMF estimation is counting the occurences of the joint variable realizations which means we require $S \gg 10^{10}$ examples for a reasonable accuracy.
- This makes the 'naive' approach very inaccurate.

What are the workarounds?

Existing Alternatives for Joint PMF Learning



- These are effective surrogates, but do not directly address the fundamental challenge in estimating high-dimensional joint probability from limited samples.

- These are effective surrogates, but do not directly address the fundamental challenge in estimating high-dimensional joint probability from limited samples.

Can we ever reliably estimate the joint PMF of variables given limited data without any structural assumptions?

Joint PMF Learning via Tensor Decomposition

- Kargas et al. proposed a new framework for blindly estimating the joint probability mass function (PMF) of N discrete random variables [Kargas et al., 2018].

- The method is based on establishing a link between joint PMF and tensors.

- Joint PMF $\Pr(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)})$, where Z_n can take I_n different values can be represented as a N-th order tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with

$$\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \Pr(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)}).$$

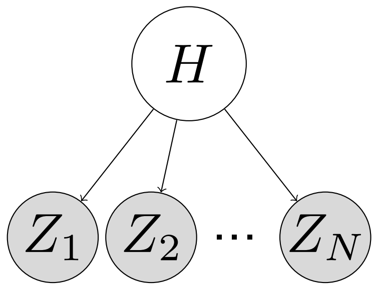
- If an N-th order tensor $\underline{\boldsymbol{X}}$ has CP rank F, then it can be **uniquely** expressed as,

$$\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \sum_{f=1}^{F} \boldsymbol{\lambda}(f) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n, f), \quad \boxed{\underline{\boldsymbol{X}} = [\![\boldsymbol{\lambda}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_N]\!].}$$

where $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$ and $\boldsymbol{\lambda} \in \mathbb{R}^F$.

Tensor Decomposition and Joint PMF

- The key point in [Kargas et al., 2018] is 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.

$$\begin{aligned}\mathsf{Pr}(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)}) &= \sum_{f=1}^{F} \mathsf{Pr}(H = f)\mathsf{Pr}(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)}|H = f) \\ &= \sum_{f=1}^{F} \mathsf{Pr}(H = f) \prod_{n=1}^{N} \mathsf{Pr}(Z_n = z_n^{(i_n)}|H = f)\end{aligned}$$

Tensor Decomposition and Joint PMF

- Putting together,

$$\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \text{Pr}(Z_1 = z_1^{(i_1)}, \ldots, Z_N = z_N^{(i_N)}). \qquad (1)$$

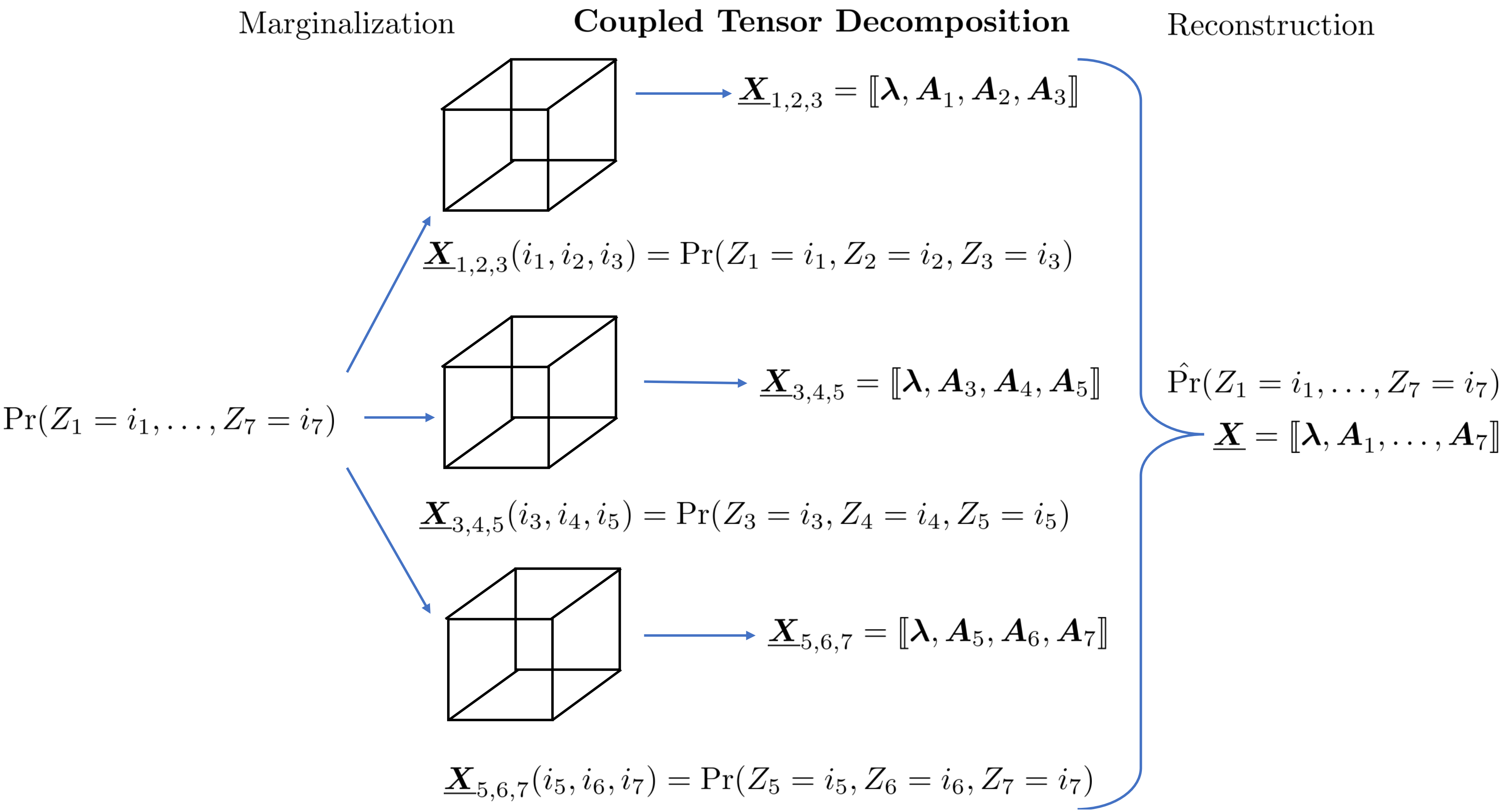
LHS of (1): $$\underline{\boldsymbol{X}}(i_1, \ldots, i_N) = \sum_{f=1}^{F} \boldsymbol{\lambda}(f) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n, f),$$

RHS of (1): $$\text{Pr}(Z_1 = i_1, \ldots, Z_K = i_N) = \sum_{f=1}^{F} \text{Pr}(H = f) \prod_{n=1}^{N} \text{Pr}(Z_n = z_n^{(i_n)} | H = f)$$

Decomposition of joint PMF tensor can identify the latent factors $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$,

$$\boldsymbol{A}_n(i_n, f) = \text{Pr}(Z_n = i_n | H = f), \quad \boldsymbol{\lambda}(f) = \text{Pr}(H = f). \qquad (2)$$

Joint PMF Learning from Third-order Marginals[1]



[1] [Kargas et al., 2018]

Challenges in the Existing Approach

- The result in [Kargas et al., 2018] is inspiring, but a couple of major hurdles exist for practical implementations.

- **High sample complexity:** Estimating three-dimensional marginals $\mathsf{Pr}(i_j, i_k, i_\ell)$ is not easy, since one needs many co-occurrences of three random variables.

- **High computational complexity:** Tensor decomposition is a hard computation problem [Hillar and Lim, 2013]—and the optimization problem involves many tensors.

Challenges in the Existing Approach

- The result in [Kargas et al., 2018] is inspiring, but a couple of major hurdles exist for practical implementations.

- **High sample complexity:** Estimating three-dimensional marginals $\text{Pr}(i_j, i_k, i_\ell)$ is not easy, since one needs many co-occurrences of three random variables.

- **High computational complexity:** Tensor decomposition is a hard computation problem [Hillar and Lim, 2013]—and the optimization problem involves many tensors.

Can we address these challenges?

Proposed Approach

- To advance the task of joint PMF recovery from marginal distributions, we propose a **pairwise marginal-based** approach.

Proposition 1: Consider discrete RVs $Z_1, \ldots, Z_N$. Assume $I_1 = \ldots = I_N = I$. Denote $p \in (0, 1]$ as the probability that an RV is observed. Let S be the number of available data samples. Assume that $\min\left(({}^2/_S)\log(2/\delta), 1\right) \leq p \leq 1$. Then, with probability at least $1 - \delta$,

$$\|\boldsymbol{X}_{jk} - \widehat{\boldsymbol{X}}_{jk}\|_{\mathrm{F}} \leq {}^{\sqrt{2}(1+\sqrt{\log(2/\delta)})}/_{(p\sqrt{S})}$$

$$\|\underline{\boldsymbol{X}}_{jk\ell} - \widehat{\underline{\boldsymbol{X}}}_{jk\ell}\|_{\mathrm{F}} \leq {}^{\sqrt{2I}(1+\sqrt{\log(2I/\delta)})}/_{(p^{3/2}\sqrt{S})}$$

hold for any distinct j, k, ℓ, where $\widehat{\boldsymbol{X}}_{jk}$ and $\widehat{\underline{\boldsymbol{X}}}_{jk\ell}$ represent the empirical estimate of $\boldsymbol{X}_{jk}$ and $\underline{\boldsymbol{X}}_{jk\ell}$ respectively, obtained via sample averaging.

- With the same amount of data, the second-order statistics can be estimated to a much higher accuracy, compared to the third-order ones.

Proposed Approach

- Consider any pairwise marginal, $\text{Pr}(i_j, i_k) = \sum_{f=1}^{F} \text{Pr}(f)\text{Pr}(i_j|f)\text{Pr}(i_k|f)$

- Since we can associate

$$\boldsymbol{X}_{jk}(i_j, i_k) = \text{Pr}(i_j, i_k),$$
$$\boldsymbol{A}_j(i_j|f) = \text{Pr}(i_j|f), \quad \boldsymbol{\lambda}(f) = \text{Pr}(f),$$
$$\boxed{\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top,} \quad \text{where } \boldsymbol{D}(\boldsymbol{\lambda}) = \text{Diag}(\boldsymbol{\lambda}).$$

- Hence, the key information for recovering the joint PMF (i.e., $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$) still shows up in the pairwise marginals.

Proposed Approach

- Consider any pairwise marginal, $\text{Pr}(i_j, i_k) = \sum_{f=1}^{F} \text{Pr}(f)\text{Pr}(i_j|f)\text{Pr}(i_k|f)$

- Since we can associate

$$\boldsymbol{X}_{jk}(i_j, i_k) = \text{Pr}(i_j, i_k),$$
$$\boldsymbol{A}_j(i_j|f) = \text{Pr}(i_j|f), \quad \boldsymbol{\lambda}(f) = \text{Pr}(f),$$
$$\boxed{\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top,} \quad \text{where } \boldsymbol{D}(\boldsymbol{\lambda}) = \text{Diag}(\boldsymbol{\lambda}).$$

- Hence, the key information for recovering the joint PMF (i.e., $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$) still shows up in the pairwise marginals.

However, there are some challenges to be addressed in pairwise-marginal based approach.

Identifiability of Matrix Factorization

- Key idea used for the triple-based approach in [Kargas et al., 2018] is that tensors admit unique CPD, under mild conditions.

- Pairwise distributions such as $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^T$ are matrices, and low-rank matrix decomposition is in general $nonunique$.

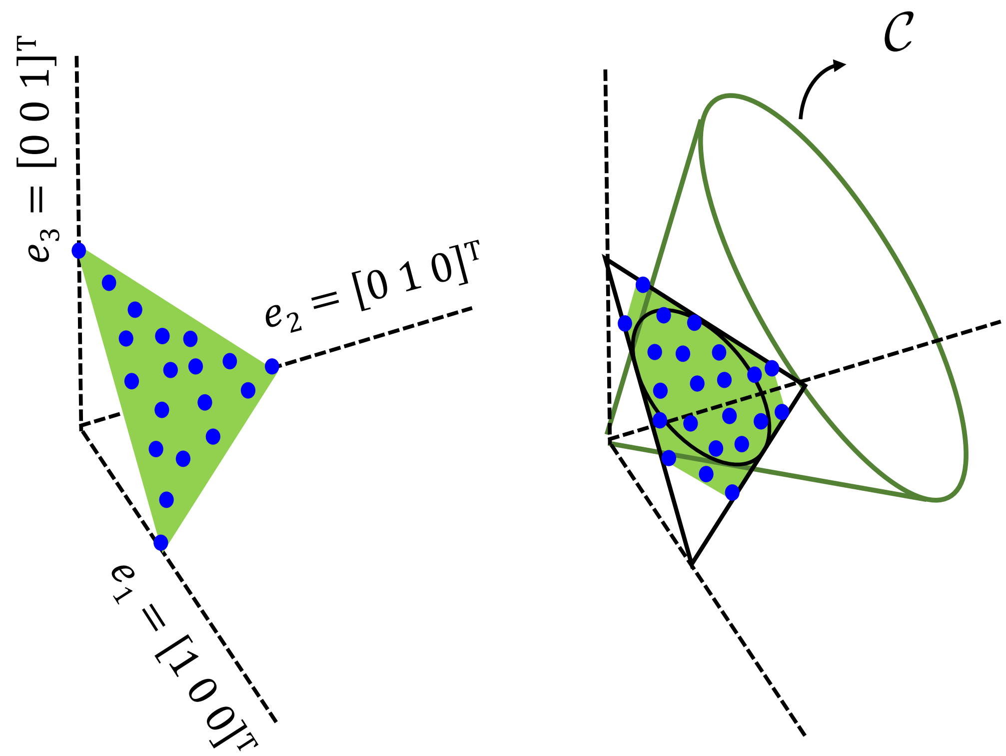
- A natural way in our case would be to employ **NMF (nonnegative matrix factorization)** tools, since the latent factors are all nonnegative.

Seperability and Sufficiently Scattered

- Assume that the nonnegative matrix $\boldsymbol{X}$ is generated by the product of two latent matrices, i.e., $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^\top$, where $\boldsymbol{W} \in \mathbb{R}^{L\times F}$ and $\boldsymbol{H} \in \mathbb{R}^{K\times F}$, $\boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0$.

Seperability: [Donoho and Stodden, 2003] If $\boldsymbol{H} \geq \mathbf{0}$, and $\boldsymbol{\Lambda} = \{l_1, \ldots, l_F\}$ such that $\boldsymbol{H}(\boldsymbol{\Lambda}, :) = \boldsymbol{\Sigma}$ holds, where $\boldsymbol{\Sigma} = \mathrm{Diag}(\alpha_1, \ldots, \alpha_F)$ and $\alpha_f > 0$, then, $\boldsymbol{H}$ satisfies the *separability condition*. When $\boldsymbol{\Lambda} = \{l_1, \ldots, l_F\}$ satisfies $\|\boldsymbol{H}(l_f, :) - \boldsymbol{e}_f\|_2 \leq \varepsilon$ for $f = 1, \ldots, F$, $\boldsymbol{H}$ is called ε-separable.

Sufficiently scattered: [Huang et al.,2014] Assume that $\boldsymbol{H} \geq \mathbf{0}$ and $\mathcal{C} \subseteq \mathrm{cone}\{\boldsymbol{H}^\top\}$ where $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^F \mid \mathbf{x}^\top\mathbf{1} \geq \sqrt{F-1}\|\mathbf{x}\|_2\}$ is a second-order cone. In addition, assume that cone$\{\boldsymbol{H}^\top\} \not\subseteq$ cone$\{\boldsymbol{Q}\}$ for any orthonormal $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ except for the permutation matrices. Then, $\boldsymbol{H}$ is called *sufficiently scattered*.



- If one of $\boldsymbol{W}$ and $\boldsymbol{H}$ satisfies the separability condition and the other has full column rank, **we can provably identify $\boldsymbol{W}$ and $\boldsymbol{H}$** up to scaling and permutation ambiguities [Gillis and Vavasis, 2014, Arora et al., 2013].

- If $\boldsymbol{W}$ and $\boldsymbol{H}$ are both sufficiently scattered, then **the model $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}^\top$ is unique** up to scaling and permutation ambiguities [Huang et al., 2014].

Seprability and Sufficiently Scattered

- Our goal is to identify $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$ from the available pairwise marginals $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top$'s using NMF model.

$$\boldsymbol{X}_{jk} = \underbrace{\boldsymbol{A}_j}_{\boldsymbol{W}} \underbrace{\boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top}_{\boldsymbol{H}^\top} \tag{3}$$

- Note that F is the inner dimension of $\boldsymbol{A}_j \in \mathbb{R}^{I_j \times F}, \boldsymbol{A}_k \in \mathbb{R}^{I_k \times F}$ and the dimension of $\boldsymbol{D}(\boldsymbol{\lambda}) \in \mathbb{R}^{F \times F}$.

- Since F could be much larger than the I_j's. i.e., $F \gg \min\{I_j, I_k\}$ in general, separability or sufficiently scattered cannot be achieved.

When can NMF be unique?

- Intuitively, if one has many rows in $\boldsymbol{H} \geq \mathbf{0}$, then there will be some rows approaching the extreme rays of the nonnegative cone.

- This concept was formalized [Ibrahim et al., 2019]:

Lemma 1: Let $\rho > 0, \varepsilon > 0$, and assume that the rows of $\boldsymbol{H} \in \mathbb{R}^{L \times F}$ are generated within the $(F-1)$-probability simplex uniformly at random (and then nonnegatively scaled). If $L \geq \Omega\left(\frac{\varepsilon^{-2(F-1)}}{F}\log\left(\frac{F}{\rho}\right)\right)$, then, with probability greater than or equal to $1-\rho$, there exist rows of $\boldsymbol{H}$ indexed by $l_1, \ldots l_F$ such that $\|\boldsymbol{H}(l_f,:) - \boldsymbol{e}_f^\top\|_2 \leq \varepsilon,\ f = 1, \ldots, F$.

- **Also, [Ibrahim et al., 2019] proposes that more rows in H increases the probability that H is sufficiently scattered, and the probability is higher than that of H being separable, under the same L.**

Proposed Approach

- Consider a splitting of the indices of the N variables, i.e., $\mathcal{S}_1 = \{\ell_1, \ldots, \ell_M\}$ and $\mathcal{S}_2 = \{\ell_{M+1}, \ldots, \ell_N\}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, \ldots, N\}$, $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$.

- Then, we construct the following matrix:

$$\begin{aligned}\widetilde{\boldsymbol{X}} &= \begin{bmatrix} \boldsymbol{X}_{\ell_1 \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_1 \ell_N} \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_{\ell_M \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_M \ell_N} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix} \boldsymbol{D}(\boldsymbol{\lambda})}_{\boldsymbol{W}} \underbrace{\left[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top\right]}_{\boldsymbol{H}^\top}.\end{aligned} \tag{4}$$

- The idea is to construct $\widetilde{\boldsymbol{X}}$ such that $F \leq \min\{MI, (N-M)I\}$ so that $\boldsymbol{W}$ and $\boldsymbol{H}$ may satisfy the conditions for NMF identifiability.

Proposed Approach

- Consider a splitting of the indices of the N variables, i.e., $\mathcal{S}_1 = \{\ell_1, \ldots, \ell_M\}$ and $\mathcal{S}_2 = \{\ell_{M+1}, \ldots, \ell_N\}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \{1, \ldots, N\}$, $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$.

- Then, we construct the following matrix:

$$\begin{aligned} \widetilde{\boldsymbol{X}} &= \begin{bmatrix} \boldsymbol{X}_{\ell_1 \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_1 \ell_N} \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_{\ell_M \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_M \ell_N} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix} \boldsymbol{D}(\boldsymbol{\lambda})}_{\boldsymbol{W}} \underbrace{[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top]}_{\boldsymbol{H}^\top}. \end{aligned} \tag{5}$$

- The idea is to construct $\widetilde{\boldsymbol{X}}$ such that $F \leq \min\{MI, (N-M)I\}$ so that $\boldsymbol{W}$ and $\boldsymbol{H}$ may satisfy the conditions for NMF identifiability.

However, there are a couple of caveats.

Proposed Approach

- Finding a suitable splitting of $\mathcal{S}_1, \mathcal{S}_2$ such that $\boldsymbol{W}$ and $\boldsymbol{H}$ are sufficiently scattered is highly nontrivial [Huang et al.,2014].

- To address this challenge, we consider the following coupled NMF problem:

$$\underset{\{\boldsymbol{A}_n\}_{n=1}^{N}\ \boldsymbol{\lambda}}{\text{minimize}} \sum_{j,k\in\boldsymbol{\Omega}} \text{dist}\left(\boldsymbol{X}_{jk} \mid\mid \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top\right)$$

$$\text{subject to } \mathbf{1}^\top \boldsymbol{A}_j = \mathbf{1}^\top,\ \boldsymbol{A}_j \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geq \mathbf{0}$$

where $\boldsymbol{\Omega}$ contains the index set of (j,k)'s such that $j < k$ and the joint PMF $\Pr(i_j, i_k)$ is accessible.

Proposed Approach

- Finding a suitable splitting of $\mathcal{S}_1, \mathcal{S}_2$ such that $\boldsymbol{W}$ and $\boldsymbol{H}$ are sufficiently scattered is highly nontrivial [Huang et al.,2014].

- To address this challenge, we consider the following coupled NMF problem:

$$\underset{\{\boldsymbol{A}_n\}_{n=1}^N\ \boldsymbol{\lambda}}{\text{minimize}} \sum_{j,k\in\boldsymbol{\Omega}} \text{dist}\left(\boldsymbol{X}_{jk} \mid\mid \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top\right) \tag{7a}$$

$$\text{subject to } \mathbf{1}^\top \boldsymbol{A}_j = \mathbf{1}^\top,\ \boldsymbol{A}_j \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geq \mathbf{0} \tag{7b}$$

where $\boldsymbol{\Omega}$ contains the index set of (j,k)'s such that $j < k$ and the joint PMF Pr(i_j, i_k) is accessible.

Next, our task is to analyze under what conditions (7) can identify $\boldsymbol{A}_j$'s and $\boldsymbol{\lambda}$.

Theorem 1 - Recoverability

Theorem 1: Assume that that Pr(i_j, i_k)'s for $j, k \in \boldsymbol{\Omega}$ are available and that Pr$(f) \neq 0$ for $f = 1, \ldots, F$. Suppose that there exists $\mathcal{S}_1 = \{\ell_1, \ldots, \ell_M\}$ and $\mathcal{S}_2 = \{\ell_{M+1}, \ldots, \ell_Q\}$ such that $Q \leq N$ and $\mathcal{S}_1 \cup \mathcal{S}_2 \subseteq \{1, \ldots, N\}$, $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$. Also assume the following conditions hold:

- the matrices $[\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_M}^\top]^\top$ and $[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_Q}^\top]^\top$ are *sufficiently scattered*;
- all pairwise marginal distributions Pr(i_j, i_k)'s for $j \in \mathcal{S}_1$ and $k \in \mathcal{S}_2$ are available;
- every T-concatenation of $\boldsymbol{A}_n$'s, i.e., $[\boldsymbol{A}_{n_1}^\top, \ldots, \boldsymbol{A}_{n_T}^\top]^\top$, is a full column rank matrix, if $I_{n_1} + \ldots + I_{n_T} \geq F$;
- for every $j \notin \mathcal{S}_1 \cup \mathcal{S}_2$ there exists a set of $r_t \in \mathcal{S}_1 \cup \mathcal{S}_2$ for $t = 1, \ldots, T$ such that Pr(i_j, i_{r_t}) or Pr(i_{r_t}, i_j) are available.

Then, solving Problem (7) recovers Pr$(i_j|f)$ and Pr(f) for $j = 1, \ldots, N$, $f = 1, \ldots, F$, thereby the joint PMF Pr$(i_1, \ldots, i_N)$.

- The criterion spares one the effort for first finding $\mathcal{S}_1$ and $\mathcal{S}_2$ and then constructing the matrix $\widetilde{\boldsymbol{X}}$.

- Theorem 1 does not impose any restrictions on F, and thus can be very general.

- The criterion spares one the effort for first finding $\mathcal{S}_1$ and $\mathcal{S}_2$ and then constructing the matrix $\widetilde{\boldsymbol{X}}$.

- Theorem 1 does not impose any restrictions on F, and thus can be very general.

Our analysis shows that a stronger identifiability guarantee can be derived if F is below a certain threshold.

Theorem 2 : Enhanced Recoverability

Theorem 2: Assume that Pr$(f) \neq 0$ for $f = 1, \ldots, F$, and that Pr(i_j, i_k)'s for all j, k are available and Pr$(i_k, i_j) =$ Pr(i_j, i_k). If
i) $\boldsymbol{Z} = [\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_N^\top]^\top \in \mathbb{R}^{NI \times F}$ is separable or sufficiently scattered
ii) $F \leq (N-1)I - 1$,
then, solving the problem in (7) recovers Pr$(i_j|f)$ and Pr(f) for $j = 1, \ldots, N,\ f = 1, \ldots, F$, thereby the joint PMF Pr$(i_1, \ldots, i_N)$.

- In Theorem 1, the recoverability of the joint PMF depends on if $\boldsymbol{W} = [\boldsymbol{A}_{\ell_1}^\top, \ldots, \boldsymbol{A}_{\ell_M}^\top]^\top$ and $\boldsymbol{H} = [\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top]^\top$ are sufficiently scattered.
- However, under Theorem 2, the recoverability of the joint PMF depends on $\boldsymbol{Z}$ being scattered/seperable.
- Having more rows increases the probability of being separable/sufficiently scattered, thus stronger guarantee for identifibaility.

Algorithm for Coupled NMF

- Recall the coupled NMF problem

$$\underset{\{\boldsymbol{A}_n\}_{n=1}^N \, \boldsymbol{\lambda}}{\text{minimize}} \sum_{j,k \in \boldsymbol{\Omega}} \text{dist}\left(\boldsymbol{X}_{jk} \mid\mid \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top\right)$$
$$\text{subject to } \mathbf{1}^\top \boldsymbol{A}_j = \mathbf{1}^\top, \; \boldsymbol{A}_j \geq \mathbf{0}, \; \mathbf{1}^\top \boldsymbol{\lambda} = 1, \; \boldsymbol{\lambda} \geq \mathbf{0}$$

  where $\boldsymbol{\Omega}$ contains the index set of (j,k)'s such that $j < k$ and the joint PMF $\Pr(i_j, i_k)$ is accessible.

- To handle this, we propose a simple procedure based on *block coordinate descent* (BCD).

- To be specific, we cyclically minimize the constrained optimization problem w.r.t.

$\boldsymbol{A}_k$, when fixing $\boldsymbol{A}_j$ for all $j \neq k$ and $\boldsymbol{\lambda}$.

$$\underset{\boldsymbol{A}_k}{\text{minimize}} \sum_{j \in \boldsymbol{\Omega}_k} \text{dist}\left(\boldsymbol{X}_{jk} \;||\; \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top\right) \tag{9a}$$

$$\text{subject to } \mathbf{1}^\top \boldsymbol{A}_k = \mathbf{1}^\top,\ \boldsymbol{A}_k \geq \mathbf{0}, \tag{9b}$$

where $\boldsymbol{\Omega}_k$ is the index set of j such that $\Pr(i_j, i_k)$ is available.

- In our work, we adopt the KL divergence since it is natural for measuring distance between PMFs.

- Many off-the-shelf convex optimization tools can be employed to solve the above, e.g., mirror descent.

- We show that with a carefully designed initialization scheme, accurately recovering joint PMFs from pairs is viable.

Gram–Schmidt-like Initialization

- We also propose a simple algebraic algorithm for identifying $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$.

- Recall the splitting of random variables and construction of matrix $\widetilde{\boldsymbol{X}}$.

$$
\begin{aligned}
\widetilde{\boldsymbol{X}} &= \begin{bmatrix} \boldsymbol{X}_{\ell_1 \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_1 \ell_N} \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_{\ell_M \ell_{M+1}} & \cdots & \boldsymbol{X}_{\ell_M \ell_N} \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix}}_{\boldsymbol{W}} \boldsymbol{D}(\boldsymbol{\lambda}) \underbrace{\left[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top \right]}_{\boldsymbol{H}^\top}.
\end{aligned} \tag{10}
$$

- Let us assume $\boldsymbol{W}$ is full rank and $\boldsymbol{H}$ is separable.

- Under the separability condition, we have $\boldsymbol{H}(\boldsymbol{\Lambda},:) = \boldsymbol{\Sigma} = \mathrm{Diag}(\alpha_1, \ldots, \alpha_F)$ and

$$\boxed{\boldsymbol{W}\boldsymbol{\Sigma} = \widetilde{\boldsymbol{X}}(\boldsymbol{\Lambda},:).} \tag{11}$$

- i.e, Estimation of $\boldsymbol{W}$ is an index identification task and can be achieved by using **Successive projection algorithm (SPA)** [Araújo et al.,2001]

- SPA is very scalable- a Gram-Schmitt-like algorithm, which only consists of norm comparison and orthogonal projection.

- SPA is robust to noise and slight violation of separability.

- $\boldsymbol{A}_{\ell_n} \in \mathbb{R}^{I_{\ell_n} \times F}, n \in \{1, \ldots, M\}$ can be identified upto column permutations $(\widehat{\boldsymbol{A}}_{\ell_n} = \boldsymbol{A}_{\ell_n}\boldsymbol{\Pi})$ since

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix} \boldsymbol{D}(\boldsymbol{\lambda}), \boldsymbol{1}^\top \boldsymbol{A}_k = \boldsymbol{1}^\top, \;\; \boldsymbol{A}_k \geq \boldsymbol{0} \tag{12}$$

- $\boldsymbol{A}_{\ell_n}$ for $n \in \{M+1, \ldots, N\}$ can be identified upto column permutations, since $\boldsymbol{H}$ matrix can be estimated using (constrained) least squares, $\arg\min_{\boldsymbol{H} \geq 0} \|\widetilde{\boldsymbol{X}} - \boldsymbol{W}\boldsymbol{H}^\top\|_F^2$
- $\boldsymbol{\lambda}$ can be identified as $\widehat{\boldsymbol{\lambda}} = (\boldsymbol{H} \odot \widetilde{\boldsymbol{W}})^\dagger \text{vec}(\widetilde{\boldsymbol{X}}) = \boldsymbol{\Pi}\boldsymbol{\lambda}$, since

$$\widetilde{\boldsymbol{X}} = \underbrace{\begin{bmatrix} \boldsymbol{A}_{\ell_1} \\ \vdots \\ \boldsymbol{A}_{\ell_M} \end{bmatrix}}_{\widetilde{\boldsymbol{W}}} \boldsymbol{D}(\boldsymbol{\lambda}) \underbrace{\left[\boldsymbol{A}_{\ell_{M+1}}^\top, \ldots, \boldsymbol{A}_{\ell_N}^\top\right]}_{\boldsymbol{H}^\top}. \tag{13}$$

- Named as CNMF-SPA – scalable algorithm, a good choice for initialization.

Theorem 3 - Accuracy of CNMF-SPA

Theorem 3: Let p and S be the probability of each RV being observed in one realization of $\Pr(Z_1, \ldots, Z_N)$ and the number of total realizations. Suppose that $I_n = I$ for all n. Assume that $\|\widehat{\boldsymbol{X}}_{ij}(:,q)\|_1 \geq \eta > 0$ for any q, i, j, and that the rows of $\boldsymbol{A}_n$'s are generated from the probability simplex uniformly at random and then positively scaled. Also assume that $\min(\frac{2}{S}\log(4/\delta), 1) \leq p \leq 1$, $N = M + \Omega(\frac{M\kappa^3(\boldsymbol{W})}{I\sqrt{F}}\log\left(\frac{F}{\delta}\right))$ and $F = O\left(\frac{\eta p \sqrt{S}}{MI\kappa^2(\boldsymbol{W})\sqrt{\log(1/\delta)}} \min\left(\frac{\sigma_{\min}(\boldsymbol{W})}{\sqrt{M}}, \frac{\sigma_{\max}(\boldsymbol{H})}{4\sqrt{N-M}}\right)\right)$. Then, applying CNMF-SPA on $\widetilde{\boldsymbol{X}}$ with $\mathcal{S}_1 = \{1, \ldots, M\}$ and $\mathcal{S}_2 = \{M+1, \ldots, N\}$ outputs

$$\|\boldsymbol{A}_n - \widehat{\boldsymbol{A}}_n\|_2 = O\left(\kappa^3(\boldsymbol{W}) MF\sqrt{L}\eta^{-1}\zeta\right), \ \forall n,$$

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\|_2 = O\left(\kappa^3(\boldsymbol{W})\kappa(\boldsymbol{H}) MF\sqrt{MK}\eta^{-1}\zeta\right),$$

with probability at least $1-\delta$, where $L = MI$, $K = (N-M)I$, $\boldsymbol{W}$ and $\boldsymbol{H}$ follow the definition in (13) and $\zeta = \max\left(\frac{\sqrt{I\log(2/\delta)}}{\eta p\sqrt{S}}, \frac{\sigma_{\min}(\boldsymbol{W})}{\kappa^2(\boldsymbol{W})M\sqrt{F}}\right)$.

Experiments: Synthetic Data

- We consider $N = 5$ RV's where each variable takes $I = 10$ discrete values.
- The columns of the conditional PMF matrices (factor matrices) $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$ and the prior probability vector $\boldsymbol{\lambda} \in \mathbb{R}^F$ are generated with $F = 5$.
- The ε-separability condition on $\boldsymbol{H}$ is ensured with $\varepsilon = 0.1$.
- We generate S realizations of the joint PMF by randomly hiding each variable realization with observation probability $p = 0.5$.

Experiments: Synthetic Data

Table 1: MSE & MRE for $N = 5, F = 5, I = 10, p = 0.5$

Algorithms	Metric	$S = 10^3$	$S = 10^4$	$S = 10^5$	$S = 10^6$
CNMF-SPA	MSE	0.0703	0.0257	0.0213	0.0207
CNMF-OPT	MSE	**0.0520**	0.0234	0.0210	**0.0206**
CNMF-SPA-EM	MSE	0.0580	**0.0228**	**0.0209**	**0.0206**
RAND-EM	MSE	0.0923	0.0415	0.0447	0.0476
CTD	MSE	0.1644	0.0253	0.0212	0.0207
CNMF-SPA	MRE	0.7897	0.3171	0.1104	0.0338
CNMF-OPT	MRE	**0.6797**	0.2316	0.0769	0.0235
CNMF-SPA-EM	MRE	0.6847	**0.2095**	**0.0711**	**0.0217**
RAND-EM	MRE	0.8304	0.3952	0.2926	0.3179
CTD	MRE	0.9137	0.2993	0.0959	0.0313

- **CNMF-SPA-EM** : EM algorithm proposed in [Yeredor and Haardt,2019] initialized using CNMF-SPA, **CTD** : Coupled Tensor Decomposition based algorithm proposed in [Kargas et al.,2018].

Experiments: Synthetic Data

Table 2: MSE & MRE for $N = 15, F = 10, I = 10, p = 0.5$

Algorithms	**Metric**	$S = 10^3$	$S = 10^4$	$S = 10^5$	$S = 10^6$
CNMF-SPA	MSE	0.1183	0.1030	0.1063	0.1041
CNMF-OPT	MSE	**0.0218**	**0.0042**	**0.0022**	0.0020
CNMF-SPA-EM	MSE	0.0894	0.0110	0.0056	**0.0018**
RAND-EM	MSE	0.0376	0.0112	0.0149	0.0069
CTD	MSE	0.0329	0.0359	0.0404	0.0355

Experiments: Recommender Systems

- We test the approaches using the **MovieLens 20M** dataset [Harper and Konstan, 2015]. Ratings ranges in $\{1, 2, \ldots, 5\}$.
- We choose different movie genres, namely, action, animation and romance subsets and each subset contains 30 popular movies. Hence, for every subset, $N = 30$.
- We create the validation and testing sets by randomly hiding 20% and 30% of the dataset.
- The remianing 50% is used for training (learning joint PMF in our approach).
- We predict the rating for a movie N, by user k via computing $\mathbb{E}[i_N | r_k(1), \ldots, r_k(N-1)]$ (i.e., using the MMSE estimator), where $r_k(i)$ denotes the rating of movie i by user k.

Recommender Systems

Table 3: MovieLens Action Movies set

Algorithm	RMSE	MAE	Time (s)
CNMF-SPA	0.8497±0.0114	0.6663±0.0059	0.031
CNMF-OPT	0.8167±0.0035	0.6321±0.0040	70.018
CNMF-SPA-EM	**0.7840±0.0025**	**0.5991±0.0031**	2.424
CTD	0.8770±0.0088	0.6649±0.0076	52.253
BMF	0.8011±0.0012	0.6260±0.0013	46.637
Global Average	0.9468±0.0018	0.6956±0.0017	–
User Average	0.8950±0.0010	0.6825±0.0010	–
Movie Average	0.8847±0.0018	0.6982±0.0012	–

Recommender Systems

Table 4: MovieLens Animation Movies set

Algorithm	RMSE	MAE	Time (s)
CNMF-SPA	0.8705±0.0095	0.6798±0.0060	0.028
CNMF-OPT	**0.8124±0.0031**	**0.6241±0.0041**	61.018
CNMF-SPA-EM	0.8170±0.0075	0.6317±0.0086	2.424
CTD	0.8300±0.0053	0.6335±0.0029	48.253
BMF	0.8408±0.0023	0.6553±0.0015	46.637
Global Average	0.9371±0.0021	0.7042±0.0014	–
User Average	0.8850±0.0009	0.6632±0.0011	–
Movie Average	0.9027±0.0019	0.6900±0.0013	–

Recommender Systems

Table 5: MovieLens Romance Movies set

Algorithm	RMSE	MAE	Time (s)
CNMF-SPA	0.9280±0.0066	0.7376±0.0076	0.032
CNMF-OPT	0.9076±0.0014	0.7123±0.0029	60.762
CNMF-SPA-EM	**0.9057±0.0052**	**0.7106±0.0049**	1.881
CTD	0.9498±0.0085	0.7416±0.0054	47.010
BMF	0.9337±0.0007	0.7463±0.0009	31.823
Global Average	1.0019±0.0007	0.8078±0.0008	–
User Average	1.0195±0.0007	0.7862±0.0008	–
Movie Average	0.9482±0.0007	0.7599±0.0007	–

Experiments: Classification

- We use several UCI datasets in the classification tasks.
- We split each dataset into training, validation and testing sets in the ratio of $50 : 20 : 30$.
- We estimate the joint PMF of the features and the label using the training set, and then predict the labels on the testing data by constructing an MAP predictor.
- For each dataset, we perform 20 trials with randomly partitioned training/testing/validation sets.

Table 6: UCI Dataset Votes

Algorithm	**Accuracy (%)**	**Time (sec.)**
CNMF-SPA	88.39+/-2.61	0.005
CNMF-OPT	**95.28+/-3.84**	4.963
CNMF-SPA-EM	92.13+/-3.13	0.016
CTD	90.76+/-3.16	2.056
SVM	94.42+/-2.19	0.021
Linear Regression	95.11+/-1.77	0.020
Neural Net	93.05+/-3.30	0.106
SVM-RBF	90.38+/-3.74	0.009
Naive Bayes	88.93+/-2.76	0.018

Classification

Table 7: UCI Dataset Car

Algorithm	Accuracy (%)	Time (s)
CNMF-SPA	69.88±1.52	0.008
CNMF-OPT	85.29±2.37	2.306
CNMF-SPA-EM	**86.27**±**2.09**	0.014
CTD	84.92±2.12	0.845
SVM	84.07±1.59	0.315
Linear Regression	81.13±2.14	0.083
Neural Net	83.89±2.90	0.570
SVM-RBF	76.25±2.56	1.039
Naive Bayes	84.09±2.50	0.048

Classification

Table 8: UCI Dataset Credit

Algorithm	Accuracy (%)	Time (s)
CNMF-SPA	86.38±2.25	0.009
CNMF-OPT [Proposed]	**86.41±2.69**	4.985
CNMF-SPA-EM	85.79±2.07	0.012
CTD	86.13±2.41	3.774
SVM	85.99±2.04	0.176
Linear Regression	86.37±2.17	0.073
Neural Net	85.94±2.11	0.515
SVM-RBF	82.89±2.77	0.022
Naive Bayes	85.50±2.42	0.046

Conclusion

- We proposed a **new framework for recovering joint PMF** of any number of discrete random variables from marginal distributions.

Conclusion

- We proposed a **new framework for recovering joint PMF** of any number of discrete random variables from marginal distributions.

- Our approach only uses **two-dimensional marginals**, which naturally has **reduced-sample complexity** and **computational burden**.

Conclusion

- We proposed a **new framework for recovering joint PMF** of any number of discrete random variables from marginal distributions.
- Our approach only uses **two-dimensional marginals**, which naturally has **reduced-sample complexity** and **computational burden**.
- We showed that under certain conditions, the recoverability of joint PMF from pairwise marginals can be **provably guaranteed**.

Conclusion

- We proposed a **new framework for recovering joint PMF** of any number of discrete random variables from marginal distributions.
- Our approach only uses **two-dimensional marginals**, which naturally has **reduced-sample complexity** and **computational burden**.
- We showed that under certain conditions, the recoverability of joint PMF from pairwise marginals can be **provably guaranteed**.
- We proposed a **coupled NMF formulation** as the optimization surrogate for this task, and employed a **Gram-Schmitt-like scalable algorithm as its initialization**.

Conclusion

- We proposed a **new framework for recovering joint PMF** of any number of discrete random variables from marginal distributions.

- Our approach only uses **two-dimensional marginals**, which naturally has **reduced-sample complexity** and **computational burden**.

- We showed that under certain conditions, the recoverability of joint PMF from pairwise marginals can be **provably guaranteed**.

- We proposed a **coupled NMF formulation** as the optimization surrogate for this task, and employed a **Gram-Schmitt-like scalable algorithm as its initialization**.

- We showed that the initialization method is **effective even under the finite-sample** case and can **empirically enhance performance of an EM algorithm**.

Thank You

Back up Slides

Coupled Tensor Decomposition

- Kargas et al. showed if $F \leq \frac{(\lfloor \frac{N}{3} \rfloor I+1)^2}{16}$, where $I = I_1 = \ldots = I_N$, recoverability of the joint PMF can be guaranteed almost surely, if $\boldsymbol{A}_n$'s follow any joint absolutely continuous distribution [Kargas et al., 2018].

- To estimate the $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$, the following estimator was constructed:

$$\begin{aligned} \underset{\{\boldsymbol{A}_k\}_{k=1}^K, \boldsymbol{\lambda}}{\text{minimize}} \; & \sum_{\ell=1}^{K} \sum_{m=\ell+1}^{K} \sum_{n=m+1}^{K} \left\| \underline{\boldsymbol{X}}_{\ell,m,n} - [\![\boldsymbol{\lambda}, \boldsymbol{A}_\ell, \boldsymbol{A}_m, \boldsymbol{A}_n]\!] \right\|_F^2 \\ \text{subject to} \; & \mathbf{1}^\top \boldsymbol{A}_k = \mathbf{1}^\top, \; \boldsymbol{A}_k \geq \mathbf{0}, \; \forall k \\ & \mathbf{1}^\top \boldsymbol{\lambda} = 1, \; \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

- An *alternating least squares* (ALS) based algorithm was proposed to handle the above.

- Note that the constraints are added because the columns of $\boldsymbol{A}_n$ are conditional PMFs and $\boldsymbol{\lambda}$ is the PMF of the latent variable

Pairwise Approach - Main Hurdles

- **Identifiability**
  - A natural thought to handle the identifiability problem of $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top$ would be to employ **NMF (nonnegative matrix factorization)** tools, since the latent factors are all nonnegative.

- **High rank**
  - The uniqueness of NMF models holds only if $F \leq \min\{I_j, I_k\}$ for $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top \in \mathbb{R}^{I_j \times I_k}$.
  - Note that F is the inner dimension of $\boldsymbol{A}_j \in \mathbb{R}^{I_j \times F}, \boldsymbol{A}_k \in \mathbb{R}^{I_k \times F}$ and the dimension of $\boldsymbol{D}(\boldsymbol{\lambda}) \in \mathbb{R}^{F \times F}$.
  - F could be much larger than the I_j's. i.e., $F \gg \min\{I_j, I_k\}$.

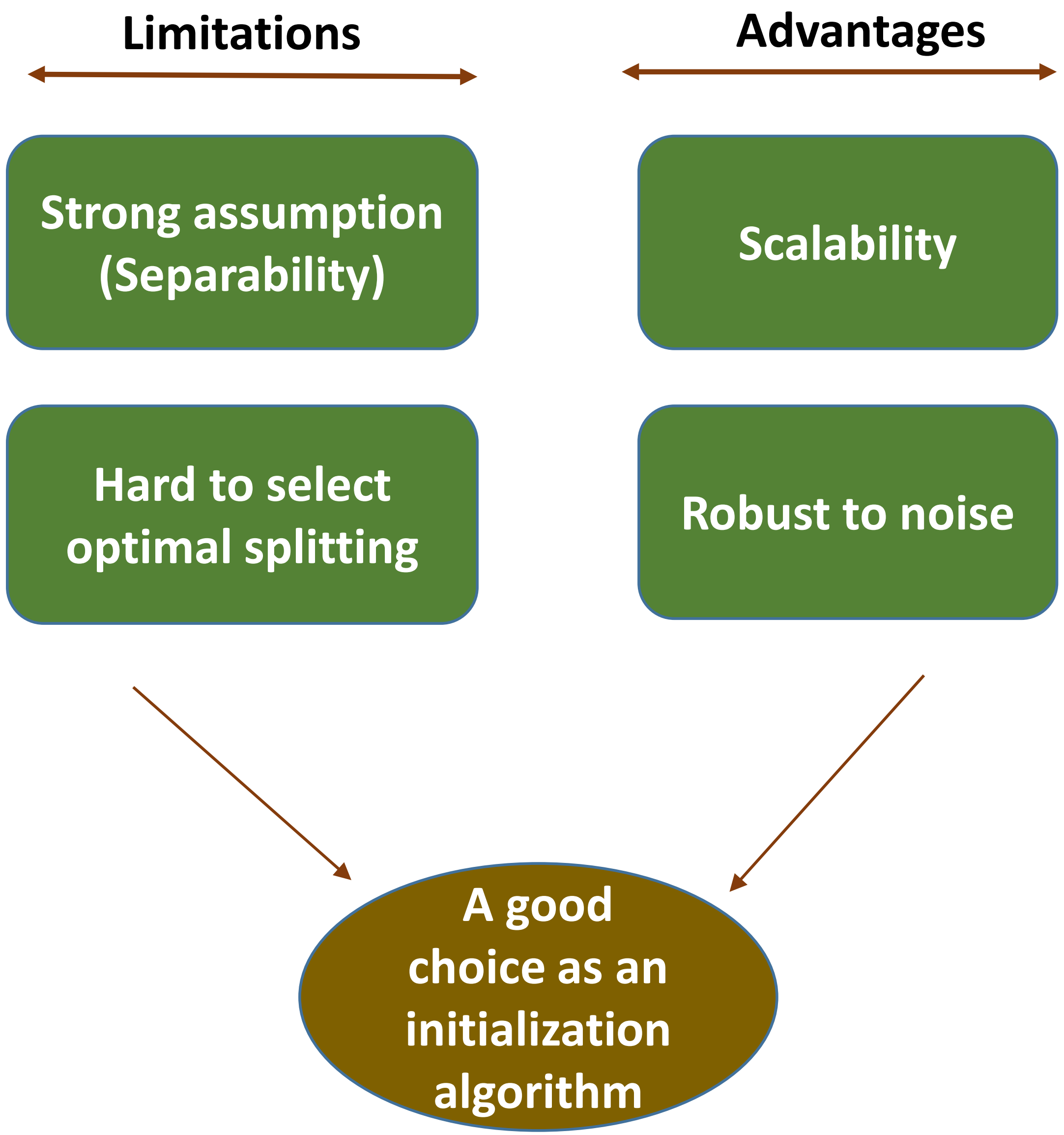
Pairwise Approach - Main Hurdles

- **Identifiability**
  - A natural thought to handle the identifiability problem of $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top$ would be to employ **NMF (nonnegative matrix factorization)** tools, since the latent factors are all nonnegative.

- **High rank**
  - The uniqueness of NMF models holds only if $F \leq \min\{I_j, I_k\}$ for $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top \in \mathbb{R}^{I_j \times I_k}$.
  - Note that F is the inner dimension of $\boldsymbol{A}_j \in \mathbb{R}^{I_j \times F}, \boldsymbol{A}_k \in \mathbb{R}^{I_k \times F}$ and the dimension of $\boldsymbol{D}(\boldsymbol{\lambda}) \in \mathbb{R}^{F \times F}$.
  - F could be much larger than the I_j's. i.e., $F \gg \min\{I_j, I_k\}$.

This means that we have to judiciously use the available NMF results to argue for joint PMF recoverability.

SPA based Algorithm



Synthetic Data Simulations

- We consider $N = 10$ random variables with n-th variable taking I discrete values.
- The rank F is fixed to be 5.
- The columns of the conditional PMF matrices (factor matrices) $\boldsymbol{A}_n \in \mathbb{R}^{I_n \times F}$ and the prior probability vector $\boldsymbol{\lambda} \in \mathbb{R}^F$ are generated using dirichlet distribution with parameter $\boldsymbol{\alpha} = \mathbf{1} \in \mathbb{R}^F$.
- We assume that the pairwise marginals of the random variables $\boldsymbol{X}_{jk}$'s are available such that $\boldsymbol{X}_{jk} = \boldsymbol{A}_j \boldsymbol{D}(\boldsymbol{\lambda}) \boldsymbol{A}_k^\top$ for all $j, k \in \{1, \ldots, N\}, j \neq k$.
- We run the experiment for different values of I ranging from 5 to 25.
- For each I, we run 10 Monte Carlo simulations by randomly generating the factor matrices $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$.

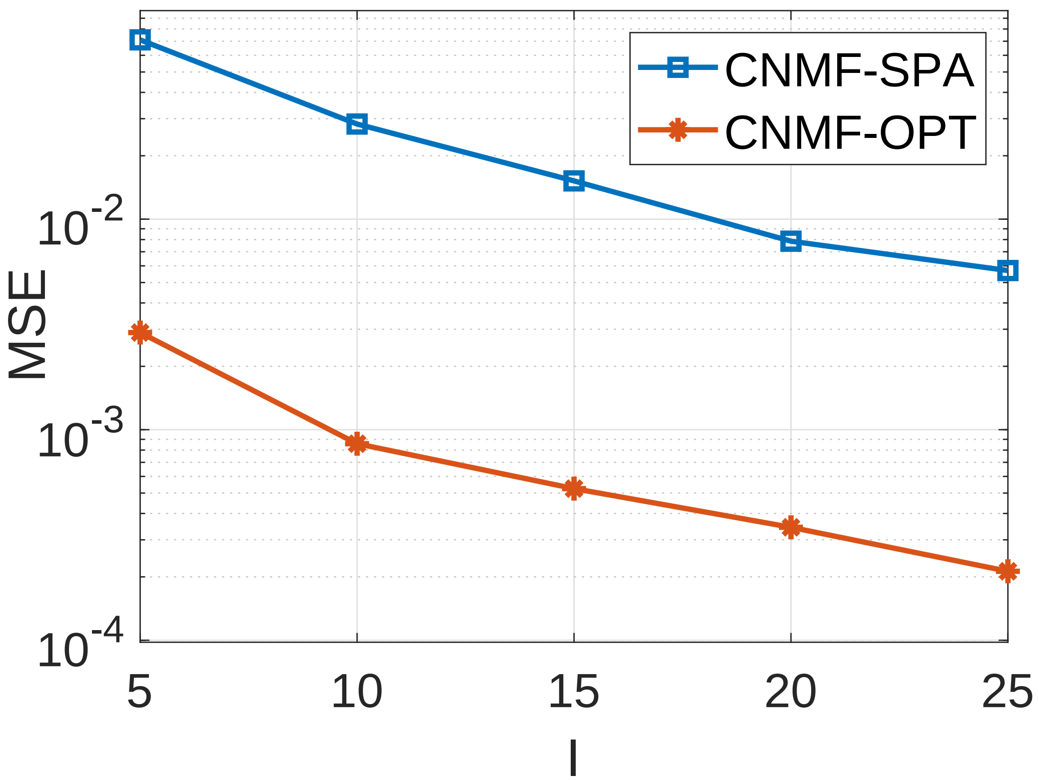


Figure 1: MSE for $N = 10, F = 5$ with different values of I

Joint PMF Learning Using Third Order Marginals

- Direct CPD of $\underline{\boldsymbol{X}}$ is not possible since estimating $\underline{\boldsymbol{X}}$ is difficult. **However, estimating the joint PMF of a subset of random variables can be possible.**

- Suppose third-order marginals are available $\text{Pr}(i_j, i_k, i_\ell)$, which can be expressed as [Kargas et al., 2018]

$$\text{Pr}(i_j, i_k, i_\ell) = \sum_{f=1}^{F} \text{Pr}(f)\text{Pr}(i_j|f)\text{Pr}(i_k|f)\text{Pr}(i_\ell|f).$$

- Let $\underline{\boldsymbol{X}}_{jk\ell}(i_j, i_k, i_\ell) = \text{Pr}(i_j, i_k, i_\ell)$. Then, we have $\underline{\boldsymbol{X}}_{jk\ell} = [\![\boldsymbol{\lambda}, \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!]$,

- If the $\underline{\boldsymbol{X}}_{jk\ell}$'s admit essentially unique CPD, then $\boldsymbol{A}_n$'s and $\boldsymbol{\lambda}$ can be identified from the marginals.