Tensors and Probability: An Intriguing Union

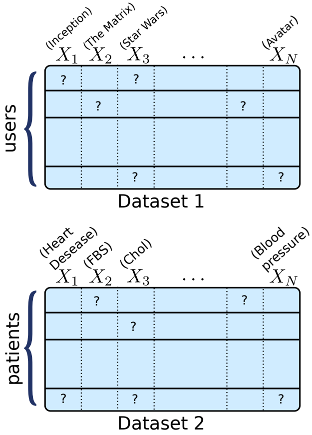
Nikos Sidiropoulos (joint work with Nikos Kargas, Xiao Fu)

November 29, 2018

University of Virginia

Motivation

- Infer missing values?
- ML: Matrix completion
low rank
- SSP: Gold standard:
statistical inference
but ... joint distribution?



Joint PMF estimation

- Without structural assumptions, joint PMF estimation is mission impossible (10 variables, 10 values each $\rightarrow 10^{10}$ parameters).

- Generic way to control joint PMF complexity?

- Is it possible to discover the underlying structure?

- Joint PMF recovery by observing subsets of variables? Possible?

Sneak preview

We will see that:

• Full joint PMF can be *provably* recovered from third-order marginal PMFs ...

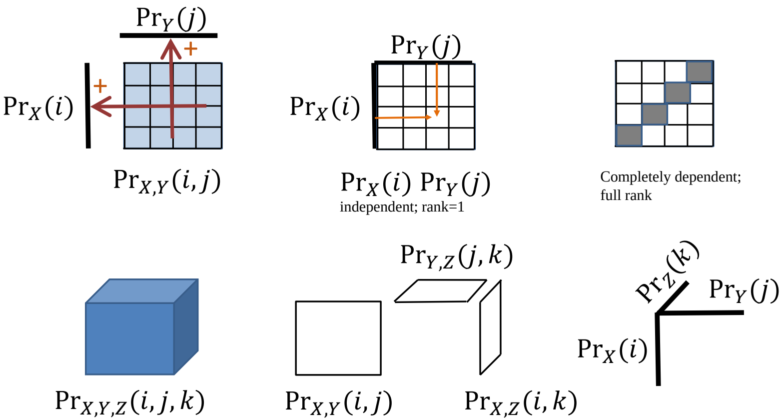
• ... provided joint PMF rank is not too large (RVs are reasonably (in)dependent).

Kolmogorov extension:

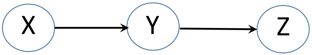
• Consistent specification of finite-dimensional distributions implies unique ∞-dim measure;

• Specification of third-order distributions implies unique higher-order, under rank condition (our result)

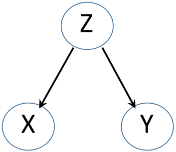
Joint PMF from marginals ('projections')?



Graphical models? — Structure?

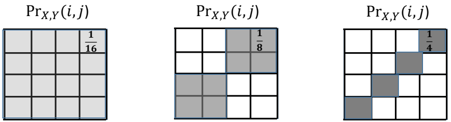


$$\Pr_{X,Y,Z}(i,j,k) = \Pr_{Z|X,Y}(k|i,j)\Pr_{X,Y}(i,j)$$
$$= \Pr_{Z|Y}(k|j)\Pr_{X,Y}(i,j) = \frac{\Pr_{Z,Y}(k,j)}{\Pr_Y(j)}\Pr_{X,Y}(i,j)$$
$$= \frac{\Pr_{Z,Y}(k,j)\Pr_{X,Y}(i,j)}{\sum_z \Pr_{Z,Y}(k,j)}$$



$$\Pr_{X,Y,Z}(i,j,k) = \Pr_{X,Y|Z}(i,j|k)\Pr_Z(k)$$
$$= \Pr_{X|Z}(i|k)\Pr_{Y|Z}(j|k)\Pr_Z(k)$$
$$= \frac{\Pr_{X,Z}(i,k)\Pr_{Y,Z}(j,k)}{\Pr_Z(k)}$$
$$= \frac{\Pr_{X,Z}(i,k)\Pr_{Y,Z}(j,k)}{\sum_X \Pr_{X,Z}(i,k)}$$

Most commonly used measure of Dependence: $D := \sum_{i,j} \Pr_{X,Y}(i,j) \ln\left(\frac{\Pr_{X,Y}(i,j)}{\Pr_X(i)\Pr_Y(j)}\right)$



R=1
D=0
Statistically independent

R=2
D=ln(2)
partial statistical dependence

R=4
D=ln(4)
Complete statistical dependence

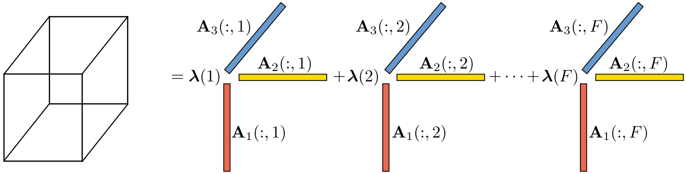
R=1 statistically independent
R=2 can model strong statistical dependence, yields 50% of D of fully dependent case
R=4 maximal statistical dependence

Canonical Polyadic Decomposition (CPD)

N-way tensor (multi-way array) $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ admits a CPD of rank F if it can be decomposed as a sum of F rank-1 tensors.

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

F is the smallest number for which such a decomposition exists.



Canonical Polyadic Decomposition (CPD)

Different ways of writing a CPD model $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$

- Element-wise

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

- Matrix (unfolding)

$$\mathbf{X}^{(n)} = (\mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \odot \cdots \odot \mathbf{A}_1)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_{\mathrm{n}}^{\mathrm{T}}$$

- Vector

$$\mathrm{vec}(\underline{\mathbf{X}}) = (\mathbf{A}_N \odot \cdots \odot \mathbf{A}_1)\, \boldsymbol{\lambda}$$

Link between naive Bayes model and CPD

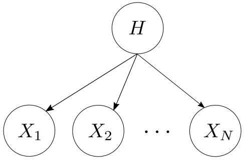
Assume that $\{X_n\}_{n=1}^N$ are conditionally independent given a variable H that takes F distinct values.

$$\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N) = \sum_{f=1}^{F} \mathsf{Pr}(H = f) \prod_{n=1}^{N} \mathsf{Pr}(X_n = i_n | H = f).$$

A special non-negative polyadic decomposition $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$ with

$$\boldsymbol{\lambda}(f) = \mathsf{Pr}(H = f),$$
$$\mathbf{A}_n(i_n, f) = \mathsf{Pr}(X_n = i_n | H = f),$$

where $\mathbf{1}^T \boldsymbol{\lambda} = 1$, $\mathbf{1}^T \mathbf{A}_n = \mathbf{1}^T$.



Naive Bayes Model.

Link between naive Bayes model and CPD

Proposition 1 (Kargas & Sidiropoulos, 2017)

Every *joint PMF can be written as*

$$\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N) = \sum_{f=1}^{F} \mathsf{Pr}(H = f) \prod_{n=1}^{N} \mathsf{Pr}(X_n = i_n | H = f)$$

with $F \leq \min_k (\prod_{\substack{n=1 \\ n \neq k}}^{N} I_n)$

$\rightarrow$ *Every* joint PMF can be represented by a naive Bayes model with a *bounded* number of latent states.

$\rightarrow$ Even when there is no physically meaningful H.

We naturally prefer $F \ll \min_k (\prod_{\substack{n=1 \\ n \neq k}}^{N} I_n)$

Reasonable in practice: random variables are not fully dependent.

Uniqueness of CPD

Definition 1 (Essential uniqueness)

For a tensor $\underline{\mathbf{X}}$ of rank F, we say that a decomposition $\underline{\mathbf{X}} = [\![\mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$ is essentially unique if the factors are unique up to a common permutation and scaling / counter-scaling of columns.

This means that if there exists another decomposition $\underline{\mathbf{X}} = [\![\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_N]\!]$, then, there exists a permutation matrix $\mathbf{\Pi}$ and and diagonal scaling matrices $\mathbf{\Lambda}_n$ such that

$$\widehat{\mathbf{A}}_n = \mathbf{A}_n \mathbf{\Pi} \mathbf{\Lambda}_n \text{ and } \prod_{n=1}^{N} \mathbf{\Lambda}_n = \mathbf{I}.$$

There is no scaling ambiguity for the nonnegative column-normalized representation $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$.

Uniqueness of CPD

Let $\underline{\mathbf{X}} = [\![\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$, where $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times F}$, $\mathbf{A}_2 \in \mathbb{R}^{I_2 \times F}$, $\mathbf{A}_3 \in \mathbb{R}^{I_3 \times F}$ with $I_1 \leq I_2 \leq I_3$.

Theorem 1 (Chiantini & Ottaviani 2012)

If $\min(I_1, I_2) \geq 3$ *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* $F \leq (I_1 - 1)(I_2 - 1)$.

Theorem 2 (Chiantini & Ottaviani 2012)

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.*

Joint PMF indentifiability from marginals?

Is a PMF identifiable from lower-order marginals? Let

$$\underline{\mathbf{X}}(i_1, \ldots, i_N) = \mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N)$$

For brevity, let's focus on triples of random variables.

Assume that third-order marginal distributions are available i.e.,

$$\underline{\mathbf{X}}_{jkl}(i_j, i_k, i_l) = \mathsf{Pr}(X_j = i_j, X_k = i_k, X_l = i_l)$$

A key observation

We saw that every PMF can be decomposed as

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

• The PMF of any subset of rvs is also a non-negative CPD model.
e.g., every marginal PMF of 3 variables X_j, X_k, X_l can be decomposed as

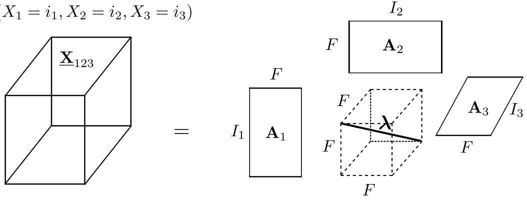
$$\mathsf{Pr}(i_j, i_k, i_l) = \sum_{f=1}^{F} \mathsf{Pr}(f)\mathsf{Pr}(i_j|f)\mathsf{Pr}(i_k|f)\mathsf{Pr}(i_l|f),$$

since $\sum_{i_n=1}^{I_n} \mathsf{Pr}(i_n|f) = 1$.
• A non-negative CPD model that depends only on 3 factors and the same hidden variable.

A key observation

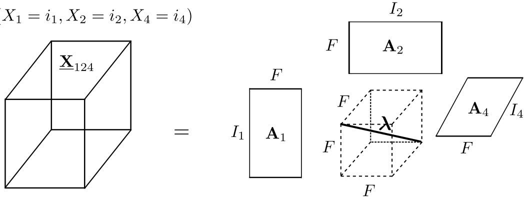
$$\boldsymbol{\lambda}(f) = \mathsf{Pr}(H = f)$$
$$\mathbf{A}_n(i_n, f) = \mathsf{Pr}(X_n = i_n | H = f)$$

$\mathsf{Pr}(X_1 = i_1, X_2 = i_2, X_3 = i_3)$



$$\underline{\mathbf{X}}_{123} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$$
$$\mathbf{X}_{123}^{(1)} = (\mathbf{A}_3 \odot \mathbf{A}_2)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^T$$

$\mathsf{Pr}(X_1 = i_1, X_2 = i_2, X_4 = i_4)$



$$\underline{\mathbf{X}}_{124} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_4]\!]$$
$$\mathbf{X}_{124}^{(1)} = (\mathbf{A}_4 \odot \mathbf{A}_2)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^T$$

Joint PMF identifiability

- Sufficient conditions for coupled CPD with one common factor: [Sørensen & De Lathauwer, 2015]

- Lower-order marginal distributions (tensors) share multiple factors.

$\rightarrow$ Better approach: Consider third-order marginals for random variables X_1, X_2, and a third random variable.

$$\begin{bmatrix} \mathbf{X}_{123}^{(1)} \\ \mathbf{X}_{124}^{(1)} \\ \vdots \\ \mathbf{X}_{12N}^{(1)} \end{bmatrix} = \begin{bmatrix} (\mathbf{A}_3 \odot \mathbf{A}_2)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^{\mathrm{T}} \\ (\mathbf{A}_4 \odot \mathbf{A}_2)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^{\mathrm{T}} \\ \vdots \\ (\mathbf{A}_N \odot \mathbf{A}_2)\mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^{\mathrm{T}} \end{bmatrix} = \left(\begin{bmatrix} \mathbf{A}_3 \\ \mathbf{A}_4 \\ \vdots \\ \mathbf{A}_N \end{bmatrix} \odot \mathbf{A}_2 \right) \mathrm{diag}(\boldsymbol{\lambda})\mathbf{A}_1^{\mathrm{T}}$$

Aggregate single-CPD model!

Joint PMF identifiability

More generally, consider a partition of the variables into 3 disjoint subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ such that the third-order marginals $\mathsf{Pr}(i_j, i_k, i_l),\ \forall j \in \mathcal{S}_1, \forall k \in \mathcal{S}_2, \forall l \in \mathcal{S}_3$ are available.
Define the following factors

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

with $u_t \in \mathcal{S}_1$, $v_t \in \mathcal{S}_2$, $w_t \in \mathcal{S}_3$.
We obtain a single non-negative CPD model

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

Assuming that $I_1 = \ldots = I_N = I$, $\underline{\widehat{\mathbf{X}}} \in \mathbb{R}^{I|\mathcal{S}_1| \times I|\mathcal{S}_2| \times I|\mathcal{S}_3|}$.

Joint PMF identifiability

Application of the uniqueness results for 3-way tensors gives

Theorem 3

- $I \leq N$ *The joint PMF is almost surely identifiable from the third-order marginals if* $F \leq I(N-2)$.
- $N \leq I$ *The joint PMF is almost surely identifiable from the third-order marginals if* $F \leq \left(\lfloor \frac{\sqrt{NI-1}}{I} \rfloor I - 1\right)^2$.

Theorem 4

The joint PMF is almost surely identifiable from the third-order marginals if $F \leq \frac{(\lfloor \frac{N}{3} \rfloor I+1)^2}{16}$.

Note: F can be of order $O(N^2 I^2)$.

Joint PMF identifiability

What about higher order marginals?

Assume that fourth-order marginals are available.
Similar to the 3-way case

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

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

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

where $\bar{\mathbf{A}}_3 = \widehat{\mathbf{A}}_4 \odot \widehat{\mathbf{A}}_3$.
In this case, identifiability can be guaranteed for much higher rank.

Algorithmic approach

Assume that we are given incomplete vector realizations (missing entries OK).

Estimate third-order marginal distributions from sample averages.

$$\underline{\mathbf{X}}_{jkl}(i_j, i_k, i_l) = \widehat{\mathsf{Pr}}(X_j = i_j, X_k = i_k, X_l = i_l)$$

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

Does the low-rank assumption hold in practice?

The empirical joint PMF of 3 randomly selected variables from different datasets was factored using a non-negative CPD model with various ranks.

Relative 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}

Problem formulation

[S2] We propose solving the following optimization problem

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

It is an instance of coupled tensor factorization.

Example

Assume that we want to estimate a joint PMF of 4 variables given third-order marginals. In this case, the cost function will be

$$f(\{\mathbf{A}_n\}_{n=1}^4, \boldsymbol{\lambda}) = \frac{1}{2}\Big(\|\underline{\mathbf{X}}_{123} - [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]\|_F^2 + \|\underline{\mathbf{X}}_{124} - [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_4]\!]\|_F^2$$

$$+ \|\underline{\mathbf{X}}_{134} - [\![\boldsymbol{\lambda}, \mathbf{A}_1, \mathbf{A}_3, \mathbf{A}_4]\!]\|_F^2 + \|\underline{\mathbf{X}}_{234} - [\![\boldsymbol{\lambda}, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4]\!]\|_F^2 \Big)$$

Algorithm

We solve problem (1) using an alternating optimization approach. Cyclically update variables $\mathbf{A}_n$ and $\boldsymbol{\lambda}$.

The optimization problem with respect to $\mathbf{A}_j$ becomes

$$\min_{\mathbf{A}_j} \sum_{k \neq j} \sum_{\substack{l \neq j \\ l > k}} \quad \frac{1}{2} \left\| \mathbf{X}_{jkl}^{(1)} - (\mathbf{A}_l \odot \mathbf{A}_k) \mathrm{diag}(\boldsymbol{\lambda}) \mathbf{A}_j^T \right\|_F^2$$

$$\text{subject to} \qquad \mathbf{A}_j \geq \mathbf{0},\ \mathbf{1}^T \mathbf{A}_j = \mathbf{1}^T.$$

Note that we have dropped the terms that do not depend on $\mathbf{A}_j$.

Algorithm

Similarly, the optimization problem with respect to $\boldsymbol{\lambda}$ becomes

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

Both problems are linearly constrained quadratic programs, and can be solved to optimality by standard solvers e.g., ADMM.

Synthetic dataset

$K = 20$ Monte Carlo simulations with randomly generated low-rank tensors

- Number of variables: $N = 5$.
- Alphabet size: $I_n = 10$, $n = 1, \ldots, 5$.
- Rank: $F \in \{5, 10, 15\}$.
- Exact marginals of pairs triples and quadruples of variables are available

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

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

where $\mathbf{\Pi}$ is a permutation matrix to fix the permutation ambiguity.

Synthetic dataset

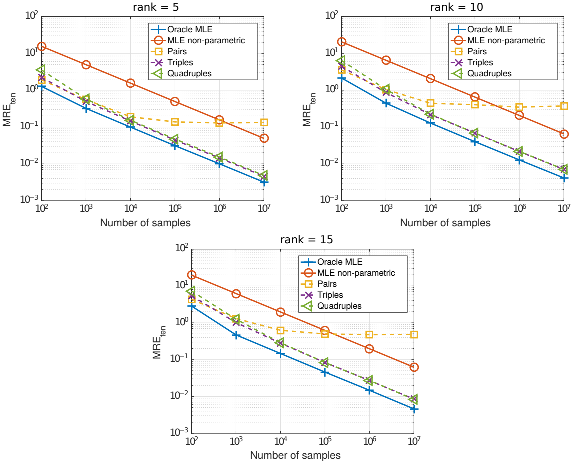
Rank		$\mathrm{MRE}_{\mathrm{fact}}$	$\mathrm{MRE}_{\mathrm{ten}}$
	Pairs	0.277	0.148
$F = 5$	Triples	1.18×10^{-7}	4.58×10^{-8}
	Quadruples	3.39×10^{-8}	1.19×10^{-8}
	Pairs	0.440	0.187
$F = 10$	Triples	3.58×10^{-7}	8.70×10^{-8}
	Quadruples	1.26×10^{-7}	2.58×10^{-8}
	Pairs	0.466	0.184
$F = 15$	Triples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

Synthetic dataset

$K = 20$ Monte Carlo simulations with randomly generated low-rank tensors

- $I_n = 10$, $n = 1, \ldots, 5$
- $F \in \{5, 10, 15\}$
- Generate M 5-dimensional data points by drawing samples from the PMF. For each data point $\mathbf{s}_m$:
  - First draw a sample h_m according to $\boldsymbol{\lambda}$.
  - Then the data point $\mathbf{s}_m$ is generated by drawing its elements independently from $\{\mathbf{A}_n\}(:, h_m)_{n=1}^{N}$.

Synthetic dataset



Mean relative error of the estimated joint PMF.

Classification task

- 7 different datasets from the UCI machine learning repository were selected.
- From each dataset select discrete features.
- Estimate lower-order marginal distributions of pairs, triples and quadruples of variables.
- For each dataset let X_N be the label and $X_1, \ldots, X_{N-1}$ the features.
- 20% used as test set, 10% as validation set and 70% as training set.
- F in the range $[1, 20]$.
- MAP estimator of the label

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

- Return the model that reports highest accuracy in validation set.

Misclassification error on different UCI datasets.

	Binary				
Method	INCOME	CREDIT	HEART	MUSHROOM	VOTES
CP (Pairs)	0.177±0.004	0.134±0.019	0.151±0.023	0.010±0.007	0.046±0.024
CP (Triples)	0.175±0.003	0.129±0.018	0.147±0.031	0.006±0.002	0.043±0.024
CP (Quadruples)	**0.171**±0.003	**0.123**±0.018	**0.138**±0.029	0.002±0.001	0.042±0.020
SVM (Linear)	0.179±0.004	0.146±0.027	0.170±0.053	**0**±0	**0.038**±0.025
SVM (RBF)	0.174±0.004	0.136±0.018	0.187±0.055	**0**±0	0.079±0.024
Naive Bayes	0.209±0.005	0.140±0.018	0.166±0.026	0.044±0.005	0.096±0.022

	Multiclass	
Method	CAR	NURSERY
CP (Pairs)	0.128±0.021	0.101±0.009
CP (Triples)	0.089±0.016	0.069±0.011
CP (Quadruples)	0.074±0.015	0.061±0.007
SVM (Linear)	0.065±0.006	0.063±0.004
SVM (RBF)	**0.026**±0.008	**0.006**±0.001
Naive Bayes	0.151±0.016	0.097±0.007

Recommender systems

MovieLens is a collaborative filtering dataset that contains 5-star movie ratings. We extracted 3 small datasets.

- 3 Categories were selected; action, romance and animation.
- Extracted ratings for 20 most rated movies of each smaller dataset.
- 20% used as test set, 10% as validation set and 70% as training set.
- F in the range $[1, 30]$.
- Conditional expectation of a movie's rating is given by

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

- Return the model that reports lowest RMSE in validation set.

RMSE and MAE of different algorithms on MovieLens.

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

Learning Mixtures of Continuous Distributions

Let $\mathcal{X} = \{X_n\}_{n=1}^N$ denote a set of N continuous RVs.
Joint PDF $f_{\mathcal{X}}$ is a mixture of F component distributions if it can be expressed as

$$f_{\mathcal{X}}(x_1, \ldots, x_N) = \sum_{f=1}^{F} w_f f_{\mathcal{X}|H}(x_1, \ldots, x_N | f).$$

Consider the special case of mixture models whose component distributions factor into the product of the associated marginals

$$f_{\mathcal{X}}(x_1, \ldots, x_N) = \sum_{f=1}^{F} w_f \prod_{n=1}^{N} f_{X_n|H}(x_n | f),$$

which can be seen as a continuous extension of the CPD model.
Learning Problem: Find the conditional PDFs as well as the mixing weights given (partially) observed samples.

Learning Mixtures of Smooth Distributions

- Common assumption made in multivariate mixture models is a parametric form of the conditional PDFs (e.g., Gaussian, Laplacian).

- Most popular algorithm for learning a parametric mixture model is Expectation Maximization (EM).

- How do we know whether true mixture components are Gaussian or Laplacian? Convenience ...

- What if we do not not assume a parametric form for the unknown conditional PDFs. Is it possible to recover mixtures of *non-parametric* product distributions from observed samples?

Approach

Consider a discretization of each RV X_n by partitioning its support into I uniform intervals $\{\Delta_n^i = \left(d_n^{i-1}, d_n^i\right)\}_{1\leq i \leq I}$.

Define the probability tensor
$\underline{\mathbf{X}}(i_1,\ldots,i_N) \triangleq \mathsf{Pr}\left(X_1 \in \Delta_n^{i_1},\ldots,X_N \in \Delta_n^{i_N}\right)$

$$
\begin{aligned}
\underline{\mathbf{X}}(i_1,\ldots,i_N) &= \sum_{f=1}^{F} w_f \prod_{n=1}^{N} \int_{\Delta_n^{i_n}} f_{X_n|H}(x_n|f)dx_n \\
&= \sum_{f=1}^{F} w_f \prod_{n=1}^{N} \mathsf{Pr}\left(X_n \in \Delta_n^{i_n} \middle| H = f\right).
\end{aligned}
$$

Let $\mathbf{A}_n(i_n, f) \triangleq \mathsf{Pr}\left(X_n \in \Delta_n^{i_n} \middle| H = f\right)$, $\boldsymbol{\lambda}(f) \triangleq w_f$.

$\underline{\mathbf{X}}$ is an N-way tensor and admits a CPD $\underline{\mathbf{X}} = [\![\boldsymbol{\lambda}, \mathbf{A}_1, \ldots, \mathbf{A}_N]\!]$.

Approach

In practice we do not observe the true $\underline{\mathbf{X}}$ but only (discretized) samples drawn from it.

Often have to deal with missing / limited data; cannot directly estimate $\underline{\mathbf{X}}$ – too many unknowns.

- Is it still possible to learn the mixing weights and discretized conditional PDFs?
  - ◇ Yes! Joint factorization of histogram estimates of lower-dimensional PDFs.
- Is it possible to recover non-parametric conditional PDFs from their discretized counterparts?
  - ◇ Yes, if the conditional PDFs are approximately band-limited (smooth).

Toy Example

It is possible to estimate samples of the conditional CDFs from the recovered factor matrices $\mathbf{A}_n$, $n = 1, \ldots, N$.

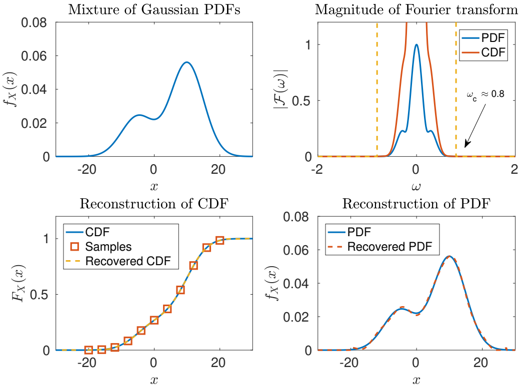


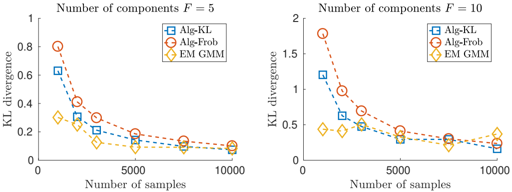
Illustration of the key idea on a univariate Gaussian mixture. The CDF can be recovered from its samples if $T_s \leq \frac{\pi}{0.8}$.

Synthetic dataset

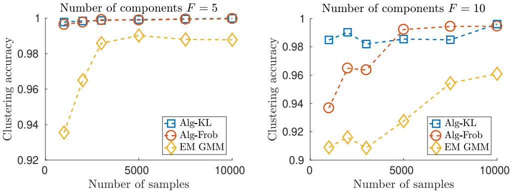
We generate synthetic datasets $\{\mathbf{x}_m\}_{m=1}^{M}$ of varying sample size.

- $I_n = 15$, $n = 1, \dots, 10$
- $F \in \{5, 10\}$
- We explore the following settings for the conditional PDFs: (1) Gaussian (2) Gaussian mixture with two components. Evaluate the performance of the algorithms by computing
  1. Clustering accuracy on $M' = 1000$ test points.
  2. KL divergence between the true and learned model, which is approximated using Monte Carlo integration.

$$\mathrm{D_{KL}}\left(f_{\mathcal{X}}, \widehat{f}_{\mathcal{X}}\right) \approx \frac{1}{M'} \sum_{m'=1}^{M'} \log f_{\mathcal{X}}(\mathbf{x}_{m'}) / \widehat{f}_{\mathcal{X}}(\mathbf{x}_{m'}).$$

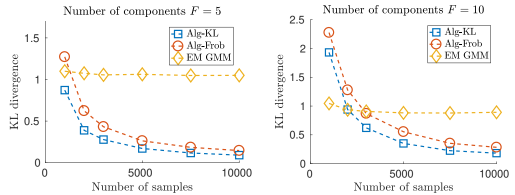


KL divergence (Gaussian).

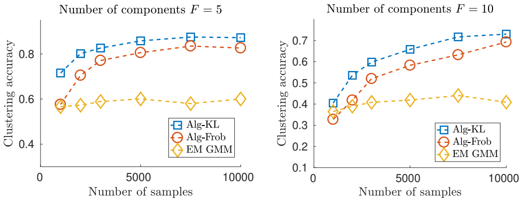


Clustering accuracy (Gaussian).

Synthetic dataset (Gaussian mixture)



KL divergence (GMM).



Clustering Accuracy (GMM).

Take-home points

Concluding remarks

- High dimensional joint PMFs hard to estimate.
- First estimate lower-order marginals.
- Fuse together using coupled CPD to estimate high-order joint.
- Identifiability of full joint PMF when rank is small.
- Analogy to Kolmogorov extension.
- Real-life random variables are never completely dependent.
- Small rank can capture significant statistical dependence.
- Scratched surface – lots of exciting research ahead!

Thank you!
Questions?