Completing a joint PMF from projections: a low-rank coupled tensor factorization approach

Nikos Kargas Nicholas D. Sidiropoulos

University of Minnesota

February 11, 2017

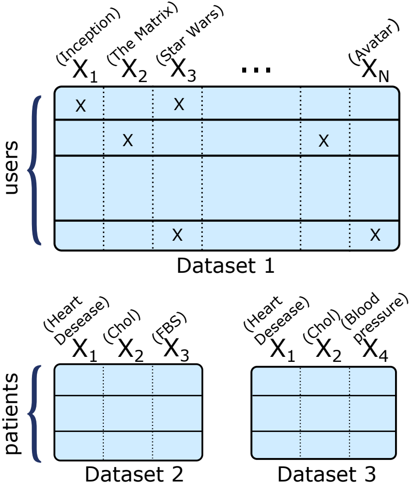
Motivation (1/4)

Dataset 1

- Missing data

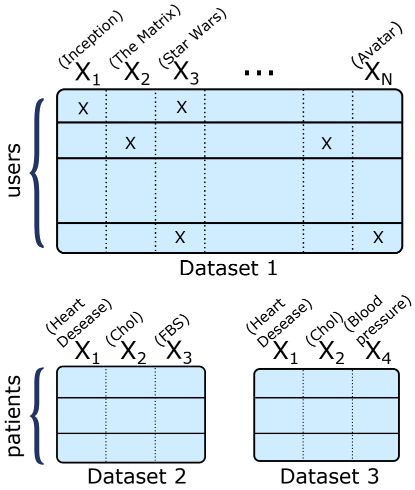
Datasets 2,3

- Common features in different datasets



Motivation (2/4)

Goal: Infer missing values given the observed ones

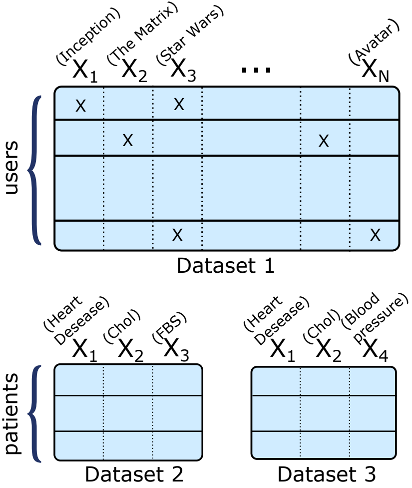


Motivation (3/4)

Why settle for burger when you can have steak? (Paul Newman)

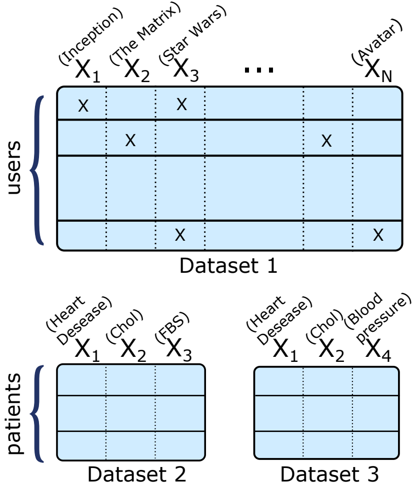
Trained in statistical signal processing?

Can we learn the joint PMF of $X_1, \ldots, X_N$?



Motivation (4/4)

Data completion vs. joint PMF completion



Outline

1. Problem Statement
2. Background
3. Our Approach
4. Results

Problem Statement

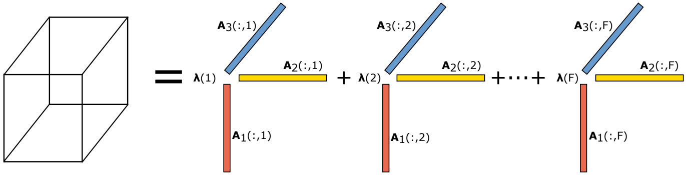
- **Problem Statement**
  - Set of discrete variables $(X_1, \ldots, X_N)$
  - Each one takes I_1, I_2,$\ldots, I_N$ distinct values
  - Partially observed dataset of M discrete samples
  - Our goal is to learn a joint PMF $\mathbb{P}(X_1, \ldots, X_n)$
- **Challenges**
  - Missing values
  - Small number of samples
  - Many parameters (10 variables, 10 values each $\rightarrow 10^{10}$)
- **Proposed Method**
  - Estimate lower-order marginals
  - Tensor factorization approach
  - Fit low-rank tensor [Canonical Polyadic Decomposition (CPD)] model for the joint PMF

Canonical Polyadic Decomposition (CPD) (1/2)

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) (2/2)

Different views of a Tensor

- 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)

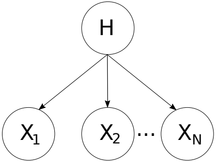
$$\mathbf{X}^{(n)} = (\mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \odot \cdots \odot \mathbf{A}_1)\mathcal{D}(\boldsymbol{\lambda})\mathbf{A}_n^T$$

- Vector

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

CPD and Latent Variable Models (1/3)

A joint PMF of discrete random variables satisfying the naive Bayes hypothesis admits a non-negative CPD [Shashua & Hazan 2005],[Lim & Common, 2009].



Naive Bayes Model.

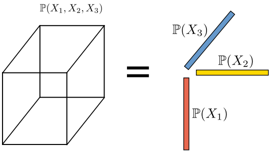
$$\mathbb{P}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \mathbb{P}(f) \prod_{n=1}^{N} \mathbb{P}(i_n|f),$$

where $\mathbb{P}(f) := \mathbb{P}(H = f)$, $\mathbb{P}(i_n|f) := \mathbb{P}(X_n = i_n|H = f)$.

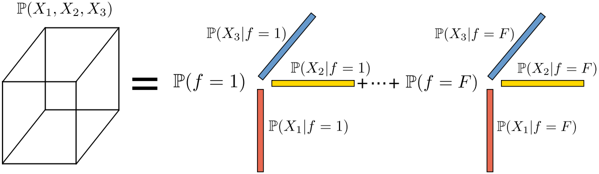
CPD and Latent Variable Models (2/3)

Random variables X_1, X_2, X_3

- Independent
  $\mathbb{P}(i_1, i_2, i_3) = \mathbb{P}(i_1)\mathbb{P}(i_2)\mathbb{P}(i_3)$
- Conditionally independent
  $\mathbb{P}(i_1, i_2, i_3) = \sum_{f=1}^{F} \mathbb{P}(f)\mathbb{P}(i_1|f)\mathbb{P}(i_2|f)\mathbb{P}(i_3|f)$



Rank $= 1$



Rank $=$ F

CPD and Latent Variable Models (3/3)

Interested in cases where the PMF can be approximated by a low-rank CPD model. (Why?)

$$\mathbb{P}(i_1, i_2, \ldots, i_N) \approx \sum_{f=1}^{F} \mathbb{P}(f)\mathbb{P}(i_1|f) \cdots \mathbb{P}(i_N|f)$$

- Is this a reasonable assumption to make?
- What is considered a low-rank model?

Every joint PMF admits a CPD for F large enough.
Upper bound for nonnegative rank $F \leq \min\limits_{k}(\prod_{\substack{n=1\\n\neq k}}^{N} I_n)$.

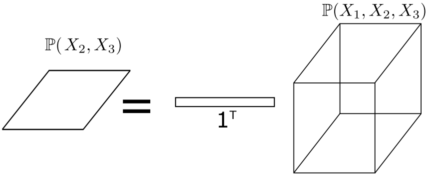
Ideally we would like $F \ll \min\limits_{k}(\prod_{\substack{n=1\\n\neq k}}^{N} I_n)$.

Problem Formulation (1/3)

For brevity, let's focus on triples of random variables.
$\widehat{\mathbb{P}}(X_j, X_k, X_l)$, $j, k, l \in \{1, \ldots, N\}$, $j \neq k, j \neq l, k \neq l$.

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

Observations can be thought as linear combinations of tensor elements.
For example, $\mathbb{P}(X_2, X_3) = \sum_{i_1=1}^{I_1} \mathbb{P}(X_1 = i_1, X_2, X_3)$



Problem Formulation (2/3)

Under the assumption of a low-rank CPD model

$$\mathbb{P}(i_1, i_2, \ldots, i_N) = \sum_{f=1}^{F} \mathbb{P}(f) \prod_{n=1}^{N} \mathbb{P}(i_n | f),$$

it is easy to verify that every marginal PMF can be decomposed as follows

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

a CPD model that depends only on 3 factors, since $\sum_{i_n=1}^{I_n} \mathbb{P}(i_n|f) = 1$.

Problem Formulation (3/3)

Therefore, we propose solving the following optimization problem

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

where $\mathbf{A}_n \in \mathbb{R}_+^{I_n \times F}$, $\boldsymbol{\lambda} \in \mathbb{R}_+^F$. It is an instance of coupled tensor factorization.

Identifiability Considerations (1/2)

Are the model parameters identifiable?
Sufficient conditions for Coupled CPD with one common factor: [Sørensen & De Lathauwer, 2015]

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

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

Identifiability Considerations (2/2)

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

It can be seen as an individual CPD model! Existing results apply.

Theorem (Chiantini & Ottaviani,2012)

If $\sum_{n=3}^{N} I_n \geq F$, $\min(I_1, I_2) \geq 3$, and $(I_1 - 1)(I_2 - 1) \geq F$, then the rank of the tensor is F and the decomposition is essentially unique, almost surely.

Coupling can be further exploited. Many more possibilities.

Alternating Optimization (1/2)

We solve (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

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

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

Alternating Optimization (2/2)

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

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

The two problems are solved via an ADMM algorithm.

Synthetic Dataset (1/3)

$K = 20$ Monte Carlo simulations with randomly generated tensors

- $I_n = 10$, $n = 1, \ldots, 5$
- $F \in \{5, 10, 15\}$
- Marginals of pairs triples and quadruples are given
- Noiseless and noisy data

$$\text{MRE}_{\text{fact}} = \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{\|\mathbf{A}_n^k - \Pi^k \widehat{\mathbf{A}}_n^k\|_F}{\|\mathbf{A}_n\|_F}$$

$$\text{MRE}_{\text{ten}} = \frac{1}{K} \sum_{n=1}^{K} \frac{\|\underline{\mathbf{X}}^k - \widehat{\underline{\mathbf{X}}}^k\|_F}{\|\underline{\mathbf{X}}^k\|_F}$$

Synthetic Dataset (2/3)

Case I: Low Rank

Rank		Rel. Fact. Error	Rel. Ten. Error
$F = 5$	Pairs	0.235	0.124
	Triples	1.24×10^{-6}	2.80×10^{-7}
	Quadruples	8.64×10^{-11}	1.53×10^{-11}
$F = 10$	Pairs	0.412	0.176
	Triples	6.91×10^{-5}	1.36×10^{-5}
	Quadruples	2.17×10^{-9}	3.37×10^{-10}
$F = 15$	Pairs	0.433	0.194
	Triples	8.56×10^{-4}	1.47×10^{-4}
	Quadruples	8.95×10^{-7}	3.63×10^{-8}

Relative factor and tensor error (noiseless data).

Synthetic Dataset (3/3)

Case II: Full Rank

Rank		Rel. Fact. Error	Rel. Ten. Error
$F = 5$	Pairs	0.305	0.17
	Triples	4.5×10^{-3}	4.4×10^{-3}
	Quadruples	4.1×10^{-3}	4×10^{-3}
$F = 10$	Pairs	0.41	0.181
	Triples	10.3×10^{-3}	6.7×10^{-3}
	Quadruples	9.2×10^{-3}	6.1×10^{-3}
$F = 15$	Pairs	0.428	0.19
	Triples	16.2×10^{-3}	8.4×10^{-3}
	Quadruples	14.1×10^{-3}	7.7×10^{-3}

Relative factor and tensor error ($\sigma = 10^{-6}$).

Collaborative Filtering Dataset (1/3)

MovieLens is a collaborative filtering dataset that contains 5-star movie ratings with 0.5 star increments. We extracted 3 small datasets.

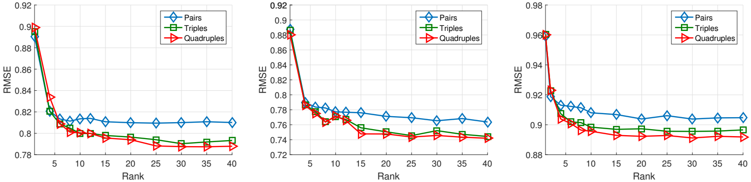
- 3 Categories were selected; action, romance and animation
- Extracted ratings for 10 most rated movies of each smaller dataset
- Performed 20 Monte Carlo simulations
- 20% used as a test set, 10% as a validation set and the remaining as a training set
- F in the range $[1, 40]$
- Run algorithms until convergence (Proposed and Biased Matrix Factorization)
- Return the model that reports best RMSE in validation set

Collaborative Filtering Dataset (2/3)

	MovieLens Dataset 1		**MovieLens Dataset 2**		**MovieLens Dataset 3**	
Method	RMSE	MAE	RMSE	MAE	RMSE	MAE
CP (Pairs)	0.8095	0.6134	0.7637	0.5811	0.9038	0.7028
CP (Triples)	0.7903	0.6003	0.7443	0.5655	0.8955	0.6947
CP (Quadruples)	**0.7874**	**0.5994**	**0.7419**	**0.5624**	**0.8912**	**0.6916**
Global Average	0.9368	0.7157	0.8924	0.7026	1.0102	0.8175
User Average	0.9388	0.6979	0.8008	0.5787	1.0693	0.8106
Item Average	0.8888	0.6863	0.8864	0.6930	0.9549	0.7516
BMF	0.8161	0.6367	0.7443	0.5760	0.9207	0.7293

RMSE and MAE on MovieLens dataset (Ratings are in the range [0.5-5])

Collaborative Filtering Dataset (3/3)



RMSE as a function of rank.

Conclusion

Concluding remarks

- High dimensional joint PMFs hard to estimate
- PMF estimation using lower-order marginals
- Identifiability of parameters when rank is small
- Efficient computation of conditional and marginal distributions

Thank you!
Questions?

- A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 792–799.
- L.-H. Lim and P. Comon, "Nonnegative approximations of nonnegative tensors," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 432–441, July 2009.
- M. Sørensen and L. D. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—Part I: Uniqueness," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 2, pp. 496–522, 2015.
- L. Chiantini and G. Ottaviani, "On generic identifiability of 3-tensors of small rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 3, p. 1018–1037, 2012.