SPRTAN: SCALABLE PARAFAC2 FOR LARGE & SPARSE DATA

PERROS I, PAPALEXAKIS E E, WANG F, ET AL.

WHY TENSOR AND WHICH FACTORIZATION

- Tensor is useful when analyze multi-modal data
- Clinical research usually focuses on three way tensor and analysis extend from 2D to 3D (patient, diagnosis, medication)
- CANDECOMP/PARAFAC (CP) is among the most popular tensor factorization types
  - Intuitive output structure and uniqueness property makes it easy to interpret
  - Restricted version of Tucker – the core tensor restricted to be diagonal
  - There exists efficient and scalable algorithms and software for the factorization

CANDECOMP/PARAFAC TENSOR FACTORIZATION

DEFINITION 5. *The CANDECOMP-PARAFAC (CP) approach approximates the original tensor $\mathcal{X}$ as a sum of rank-one tensors and is expressed as*

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{A}_r^{(1)} \circ \cdots \circ \mathbf{A}_r^{(N)} = [\![\mathbf{A}^{(1)}; \cdots ; \mathbf{A}^{(N)}]\!]$$

where $\mathbf{A}_r^{(n)}$ corresponds to the r^{th} column of $\mathbf{A}^{(n)}$. We call $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}$ the factor matrices and use $[\![\cdot]\!]$ for a shorthand notation of the sum of rank-one tensors.

PROBLEM

- Irregular time points in EHR require a powerful way to align time
  - Any preprocessing to aggregate across time may lose temporal patterns and usually these method would require continuous and sufficiently long temporal measures
- The problem causes missing data in the tensor and further hinders us to directly apply existing CP algorithms

CP-WOPT

- Tensor factorization algorithm could be extended to used for regression/classification tasks in the case of missing data

$$
\begin{aligned}
(4.5)\quad & f_{\mathcal{W}}(\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}) \\
& = \left\| \mathcal{W} * \left(\mathcal{X} - [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!] \right) \right\|^2 \\
& = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} w_{i_1 i_2 \cdots i_N}^2 \left\{ x_{i_1 i_2 \cdots i_N}^2 \right. \\
& \left. - 2 x_{i_1 i_2 \cdots i_N} \sum_{r=1}^{R} \prod_{n=1}^{N} a_{i_n r}^{(n)} + \left(\sum_{r=1}^{R} \prod_{n=1}^{N} a_{i_n r}^{(n)} \right)^2 \right\}.
\end{aligned}
$$

PARAFAC2

- Flexible version of CP
  - Irregular time points make slices of the tensor along the subject (patient) dimension to have different sizes
  - Only requires the same factor along one mode and allows the other factor matrices to vary

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$$

$$\mathbf{X}_k \approx \mathbf{U}\,\mathbf{S}_k\,\mathbf{V}^T$$

Irregular time points

$$\mathbf{X}_k \approx \mathbf{U}_k\,\mathbf{S}_k\,\mathbf{V}^T$$

X_k is the kth slice of tensor along subject mode and U_k is of size I_k * R

ALGORITHM

Classic algorithm for PARAFAC2 follows an Alternating Least Square approach

$$\min_{\{\mathbf{U}_k\},\{\mathbf{S}_k\},\mathbf{V}} \sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{U}_k\mathbf{S}_k\mathbf{V}^T||_F^2$$

$$\text{subject to: } \mathbf{U}_k = \mathbf{Q}_k\mathbf{H},\ \mathbf{Q}_k^T\mathbf{Q}_k = \mathbf{I} \text{ and } \mathbf{S}_k \text{ to be diagonal}$$

Problem: ALS on sparse matrices dominated by MTTKRP (Matricized-Tensor-Times-Khatri-Rao-Product), which requires huge storage and computational cost

Further leads to slow and complicated implementation, which hinders the extensive use of PARAFAC2

SPARTAN

- Construct a MTTKRP kernel for each one of the tensor modes
- The design makes the computation to be easily parallelized over K independent sub-problems
- Details listed in the paper, Section 4.2

RESULT

- The proposed algorithm is tested on two real datasets
  - CHOA and MovieLens
  - Implementation in MATLABR2015b
- Speed
  - SPARTan is reported to be 13 times faster than classic PARAFAC2 algorithm with low rank factorization and 22 times faster with high rank factorization
  - Run time per iteration increases dramatically for classic algorithm when rank increases; for SPARTan only increases slightly
  - SPARTan more scalable
- Phenotyping – captures the temporal trends regarding the evolution of the phenotypes for patients over time

RUBIK

- Example: some phenotypes should be related to hypertension
- Guidance could be encoded as a vector where positive entries indicate relevant feature dimensions
  - binary vector as one of the columns in the corresponding factor matrix

The value of the element in the diagnosis mode of the bias tensor corresponding to hypertension represents the possibility of any given patient having hypertension

$$\min_{\mathcal{X},\mathcal{T},\mathcal{C}} \left\{ \mathbf{\Phi}(\mathcal{X},\mathcal{T},\mathcal{C}) \right\}, \quad \text{s.t. } \underbrace{\mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{O})}_{\text{Completion}}$$

where,

$$\mathbf{\Phi} = \underbrace{\|\mathcal{X} - \mathcal{C} - \mathcal{T}\|_F^2}_{\text{Factorization error}} + \frac{\lambda_a}{2} \underbrace{\|(\mathbf{A}^{(p)} - \hat{\mathbf{A}}^{(p)})\mathbf{W}\|_F^2}_{\text{Guidance information}} \qquad (1)$$

$$+ \frac{\lambda_q}{2} \underbrace{\|\mathbf{Q} - \mathbf{A}^{(k)T}\mathbf{A}^{(k)}\|_F^2}_{\text{Pairwise constraint}}$$

$$\underbrace{\mathcal{T} = [\![\mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!] \in \Omega_{\mathcal{T}}}_{\text{Interaction tensor}}, \quad \underbrace{\mathcal{C} = [\![\mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!] \in \Omega_{\mathcal{C}}}_{\text{Bias tensor}}$$

Biased Tensor

$$\Omega_{\mathcal{T}} = \Omega_{A_1} \times \cdots \times \Omega_{A_N}, \ \Omega_{A_n} = \{\underbrace{\mathbf{A} \in \{0\} \cup [\gamma_n, +\infty)^{I_n \times R}}_{\text{Sparse representation}}\}$$

$$\Omega_{\mathcal{C}} = \Omega_{u_1} \times \cdots \times \Omega_{u_N}, \ \Omega_{u_n} = \{\mathbf{u} \in [0, +\infty)^{I_n \times 1}\}$$