CP-ORTHO: An Orthogonal Tensor Factorization Framework for Spatio-Temporal Data

Ardavan Afshar  
Georgia Institute of Technology  
afshar8@gatech.edu

Joyce C. Ho  
Emory University  
joyce.c.ho@emory.edu

Bistra Dilkina  
Georgia Institute of Technology  
bdilkina@cc.gatech.edu

Ioakeim Perros  
Georgia Institute of Technology  
perros@gatech.edu

Elias B. Khalil  
Georgia Institute of Technology  
elias.khalil@cc.gatech.edu

Li Xiong  
Emory University  
lxiong@emory.edu

Vaidy Sunderam  
Emory University  
vss@emory.edu

ABSTRACT
Extracting patterns and deriving insights from spatio-temporal data finds many target applications in various domains, such as in urban planning and computational sustainability. Due to their inherent capability of simultaneously modeling the spatial and temporal aspects of multiple instances, tensors have been successfully used to analyze such spatio-temporal data. However, standard tensor factorization approaches often result in components that are highly overlapping, which hinders the practitioner’s ability to interpret them without advanced domain knowledge. In this work, we tackle this challenge by proposing a tensor factorization framework, called CP-ORTHO, to discover distinct and easily-interpretable patterns from multi-modal, spatio-temporal data. We evaluate our approach on real data reflecting taxi drop-off activity. CP-ORTHO provides more distinct and interpretable patterns than prior art, as measured via relevant quantitative metrics, without compromising the solution’s accuracy. We observe that CP-ORTHO is fast, in that it achieves this result in 5x less time than the most accurate competing approach.

KEYWORDS
Tensor Factorization; Unsupervised Learning

1 INTRODUCTION
Spatio-temporal data analysis refers to the extraction of insights out of data containing spatial and temporal properties. A fundamental task is to automatically identify the underlying temporal trends and location patterns for sub-groups of the data instances. In urban planning, for instance, the identification of underlying temporal trends and location patterns is essential for city planners and traffic managers in order to improve a city’s road infrastructure.

A crucial characteristic of spatio-temporal datasets is that they are inherently multi-modal, due to the simultaneous presence of both location and time modes which describe each of the available instances. Thus, one of the most natural ways to model such datasets is via tensors (i.e., multi-way arrays) [8]. One of the most popular tensor analysis methods is the canonical polyadic (CP) decomposition (also known as PARAFAC or CANDECOMP) [4, 6]. CP decomposes a tensor into a sum of (rank-one) outer products which effectively represents the underlying data concepts. Its popularity owes to its intuitive output structure and uniqueness property that make the model reliable to interpret [8, 9]. However, the resulting factors of the CP model are usually highly overlapping, which hampers their interpretability without advanced domain knowledge. In order to obtain more concise and easily-interpretable results, we would prefer the solution’s components to be as distinct as possible. While prior art [11] has attempted to tackle this challenge, it promotes non-overlapping results for a specific, fixed tensor mode; such an objective is still limited by the need for prior domain knowledge or even an arbitrary choice of a tensor mode.

To tackle the challenges introduced above, we propose CP-ORTHO, a non-negative tensor factorization framework to discover distinct and easily-interpretable patterns from multi-modal, non-negative data. We propose a fast projected gradient optimization scheme to fit our objective. Our experimental evaluation on a real, publicly-available dataset showcases that CP-ORTHO achieves the best of both worlds in terms of solution distinctiveness and speed, without sacrificing the model’s accuracy. To promote reproducibility, our code is open-sourced and publicly available at https://github.com/aafshar/CP-ORTHO.

2 PRELIMINARIES AND NOTATION
The order indicates the number of tensor modes (N), R denotes the number of pursued components (tensor rank). Matricization is the process of converting a tensor into a matrix without changing its values. The mode-n matricization of $X \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is indicated as $X^{(n)} \in \mathbb{R}^{I_n \times I_1 \cdot I_{n-1} \cdot I_{n+1} \cdot \cdots \cdot I_N}$. The multiplication of tensor $X$ with N vectors in N modes is $X \times_1 A^{(1)} \times_2 A^{(2)} \cdots \times_N A^{(N)}$, where and $A^{(n)}$ is the r-th factor (column) of mode n and $X_{ik_{n}}$ indicates the multiplication in mode n. The tensor inner product of $X$ and $Y$ is defined as $X \cdot Y = \sum_{k_1=1}^{I_1} \sum_{k_2=1}^{I_2} \cdots \sum_{k_N=1}^{I_N} X_{i_1 k_1 \cdot i_2 k_2 \cdot \cdots \cdot i_N k_N} M_{i_1 k_1 \cdot i_2 k_2 \cdot \cdots \cdot i_N k_N}$

A rank-one tensor $X$ is equal to the outer product of N vectors: $X = A^{(1)} \circ A^{(2)} \circ \cdots \circ A^{(N)}$. The CANDECOMP-PARAFAC (CP)
We express the spatio-temporal dataset as an observed non-negative tensor:

\[ X = \sum_{r=1}^{R} \lambda_r A_r^{(1)} \circ A_r^{(2)} \circ \cdots \circ A_r^{(N)} \]

where factors \( A_r^{(n)} \) are orthogonal if at least one pair of their decomposed factors has a zero inner product. When the input data and the decomposed factors are non-negative, the above property implies that there is no overlap between them, thus improving interpretability. Table 1 summarizes the notations used in this paper.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \circ )</td>
<td>Element-wise Multiplication</td>
</tr>
<tr>
<td>( \otimes )</td>
<td>Outer product</td>
</tr>
<tr>
<td>( \langle A, B \rangle )</td>
<td>Khatri Rao Product</td>
</tr>
<tr>
<td>( \bullet )</td>
<td>Tensor Inner Product</td>
</tr>
<tr>
<td>( \times )</td>
<td>Tensor-Vector Multiplication</td>
</tr>
</tbody>
</table>

**Table 1: Symbols and notations used.**

Spatio-temporal data can usually be represented as event occurrences. Thus, our input tensor consists of non-negative values. In that case, interpretability is improved if the CP decomposition is constrained to contain non-negative factors as well (where both the weights and the factor matrices are non-negative). To summarize the above constraints, we provide the following constrained optimization problem:

\[
\mathcal{F} = \min_{\lambda, A^{(1)}, \ldots, A^{(N)}} \frac{1}{2} \left\| X - \left[ \lambda; A^{(1)}, A^{(2)}, \ldots, A^{(N)} \right] \right\|_F^2 \\
\text{s.t. } (A^{(1)} \circ A^{(1)}) \ast \cdots \ast (A^{(N)} \circ A^{(N)}) = I \\
\left\| A_r^{(n)} \right\|_2 = 1, \forall n,r \\
A_r^{(n)} \in [0,1]^{I_n \times R}, \lambda \in [0, +\infty)^R
\]

The first constraint (5) is the orthogonality constraint for the \( R \) rank-one tensors, and constraint (6) indicates that each column of the factor matrices must sum to one. The last two constraints are the nonnegative constraints on the weights and the factor matrices.

### 3.2 Algorithm

Since restricting the tensor decomposition to only orthogonal rank-one tensors can yield undesirable results (failing noise or poor approximation of the observed tensor), we relax the orthogonality and unit-norm constraints [1]. This also allows us to use a first-order optimization algorithm to find the solution. We use the quadratic penalty method and convert the orthogonality and unit-norm constraints to penalty terms. Note that while our problem still has nonnegative constraints, we can utilize projected gradient descent to discover the weights and the factor matrices simultaneously. For notational convenience, we reformulate the objective function (4) and the orthogonality and unit vector penalty terms using the mode-n matricization of our tensor, \( X \). The new objective function is:

\[
\mathcal{F} = \frac{1}{2} \left\| X^{(n)} - A^{(n)} \phi^{(n)} \right\|_F^2 + \frac{\psi_l}{2} \sum_{n=1}^{C_n} \left( \left\| A_r^{(n)} \right\|_2 - 1 \right)^2 \\
+ \frac{\psi_u}{2} \left\| (A^{(1)} \circ A^{(1)}) \ast C_{n} - I \right\|_F
\]
where \( \psi_a \) and \( \psi_b \) are the regularization parameters, \( \Phi(n) = (A^T \odot A_1 \odot A_2 \odot \ldots \odot A_n) \), and \( C_p = (A_1^T A_1^T \odot \ldots \odot (A^T n-1 \odot A^{T(n-1)}) \). 

**Projected Gradient Descent.** The gradient can be computed by taking the partial derivative with respect to all the factors simultaneously, vectorizing the partials, and concatenating them together (i.e., \( x = vec(A(n)) \)). After computing the gradient, we can use any first-order optimization method. Since our optimization problem has non-negativity constraints, we implement the projected gradient descent method [10] to solve the problem.

To ensure a sufficient decrease in the function at each iteration, we use backtracking line search [12] to find a sufficient value. For Rubik, \( \Delta \) membership (lower is better) and cluster membership \( \Delta C \) values over all tensor elements (higher is better) for the best initialization.

### 4 EXPERIMENTS

We compare the performance of CP-ORTHO with existing methods for constrained CP decomposition: CP-NMU [3], CP-APR [5], and Rubik [11]. A grid search for the best parameters was conducted. The final parameter values for the NYC Taxi dataset for \( R = 5 \) are: \( \psi_a = 15000, \psi_b = 15000, \alpha_A = 4e - 5, \) and \( \alpha_j = 4e - 9 \). For Rubik, we also performed a grid search with respect to its orthogonality parameter and used the value 100 which had the best result.

### 4.1 Data Description

We evaluate CP-ORTHO on NYC Taxi Data\(^1\) of taxi trips, as provided by the New York City (NYC) Taxi and Limousine Commission. We use all the data from January 2015 to June 2016 which has 215,519,509 trips (average of 393,283 trips per day). We investigate a rectangular area of 40 km x 31 km that covers Manhattan and some of the surrounding boroughs. The area is divided into grids of 250 meter x 250 meter. A third-order binary tensor is constructed from the drop-off data that represents 19,200 grid cell locations for 24 hours by 7 days.

### 4.2 Quantitative Metrics

Since all four models are only guaranteed to converge to local minima, we evaluate all competing approaches by running them for 10 random initializations. Table 2 reports the average and best fit over these 10 runs, as well as the mean running time. Regarding the accuracy of the solution, the results suggest that the mean and best fit of CP-ORTHO are comparable or better than the other models. At the same time, CP-ORTHO is 5 times faster than CP-APR, which is the most accurate competing method.

We propose two quantitative measures to assess a model’s ability to produce interpretable results. The measures correspond to the ability to produce distinct patterns and “hard” cluster results. Distinct patterns make the assignment of meaning and interpretations to the factors easier. Thus, the pattern distinctiveness metric, captures to what extent the discovered patterns are distinct form each other via the pairwise orthogonality of the rank-one tensors of components, \( \prod \delta_{n} A_i, A_j, A_k \), \( i, j \in [1, R], i < j \). For this metric, we desire lower values as it corresponds to pairs of components that are (close to) orthogonal. Table 3 summarizes the average pairwise pattern distinctiveness of the four models. CP-ORTHO has the lowest orthogonality between the pairs of corresponding components.

### Table 2: Comparison of fit and running time (10 seeds).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fit</th>
<th>Running Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± std</td>
<td>Best</td>
</tr>
<tr>
<td>CP-ORTHO</td>
<td>0.7330±0.0065</td>
<td><strong>0.7384</strong></td>
</tr>
<tr>
<td>CP-APR [5]</td>
<td>0.7371±0.0022</td>
<td>0.7381</td>
</tr>
<tr>
<td>CP-NMU [3]</td>
<td>0.7266±0.0054</td>
<td>0.7306</td>
</tr>
<tr>
<td>Rubik [11]</td>
<td>0.6852±0.0042</td>
<td>0.6894</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Pattern Distinctiveness</th>
<th>Cluster Membership(( \Delta C ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± std</td>
<td>Mean ± std</td>
</tr>
<tr>
<td>CP-ORTHO</td>
<td>0.1386±0.0646</td>
<td>0.5774±0.2995</td>
</tr>
<tr>
<td>CP-APR</td>
<td>0.1676±0.1064</td>
<td>0.4978±0.2912</td>
</tr>
<tr>
<td>CP-NMU</td>
<td>0.4923±0.1314</td>
<td>0.2418±0.2056</td>
</tr>
<tr>
<td>Rubik</td>
<td>0.2164±0.1064</td>
<td>0.3800±0.2300</td>
</tr>
</tbody>
</table>

The second measure captures the ability of each method to produce a factorization that discriminatively assigns samples to a corresponding pattern, i.e. the “hardness” of the clustering. Each factor element of the rank-one can be interpreted as the corresponding score of how much it belongs to each pattern. Thus, we analyze cluster membership score, the difference between the highest and second-highest pattern membership score, to capture how clearly the factorization assigns each element to a distinct pattern.

### 4.3 Case Study on NYC Taxi Data

Below, we study the patterns extracted from the methods under comparison. In Figure 2, we assign each combination of (day, hour) to a certain rank-one component (out of 5). To do so, we compute the outer product \( A_i \odot A_i \) for each i-th component. Then, we assign each (day, hour) combination to the component with the maximum value in the corresponding entry of the outer product.

While the drop-off locations for component 5 (Figure 3b) are more spread out over all the boroughs including the Bronx, Brooklyn, and Queens. We will provide more extensive experimental results in the paper’s extended version.

5 CONCLUSION

In this work, we propose CP-ORTHO, a tensor factorization framework enabling to find more meaningful and distinct patterns in spatio-temporal data. As measured via relevant quantitative metrics, CP-ORTHO provides more distinct and interpretable patterns than prior art, without compromising the solution’s accuracy. It is also fast, in that it achieves this result in 5x less time than the most accurate competing approach. Future directions include: 1) extending the model to include count data using KL-divergence; 2) incorporating alternative optimization approaches to reduce the number of hyper-parameters; 3) proposing an explicit metric for hard clustering in tensor factorization.

6 ACKNOWLEDGMENT

This research was supported in part by AFOSR grant FA9550-17-1-0006 and NSF grant CNS-1618932. The work of Ioakeim Perros was partially supported by NSF, award number CCF-1153568.