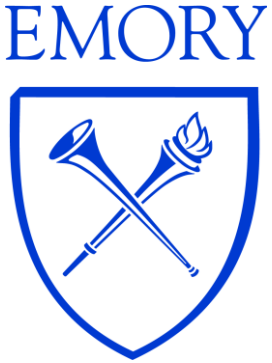


COPA: Constrained PARAFAC2 for Sparse & Large Datasets

Ardavan (Ari) Afshar[1], Ioakeim Perros[1], Evangelos E. Papalexakis[2], Elizabeth Searles[3], Joyce Ho[4], Jimeng Sun[1]

[1]Georgia Tech, [2]UC Riverside

[3]Children's Healthcare of Atlanta, [4]Emory University

Computational Phenotyping from Electronic Health Records (EHRs)

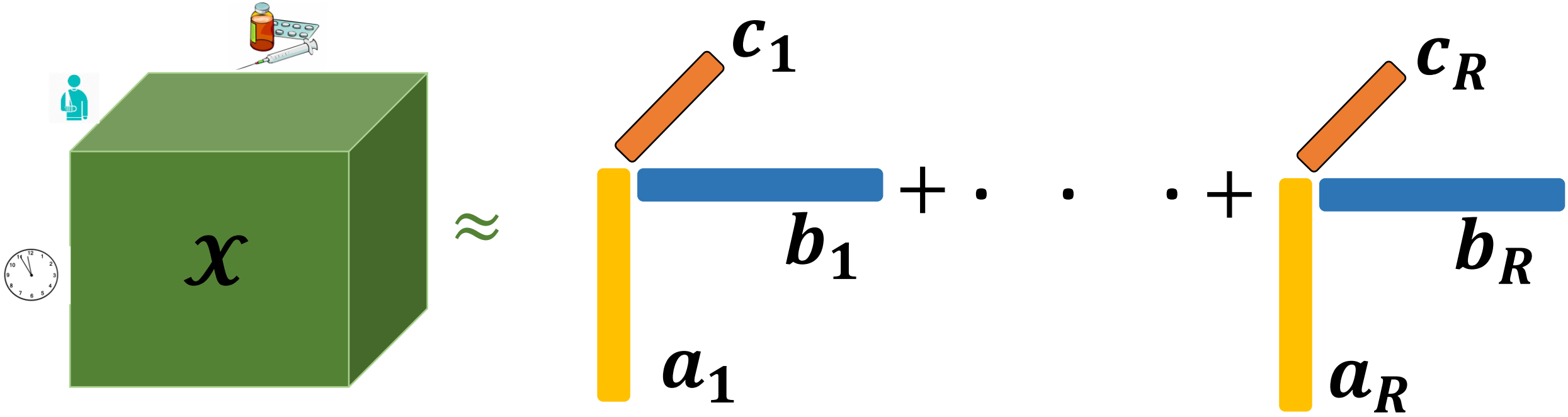
Phenotyping Algorithm

Medical Concepts (Phenotypes)

- Leukemia
- Neurological Disorders
- Congenial Anomalies
- Sickle Cell Anemia

(Phenotyping with Tensor Factorization)

- Regular Tensor Factorization Approaches (CP Decomposition)

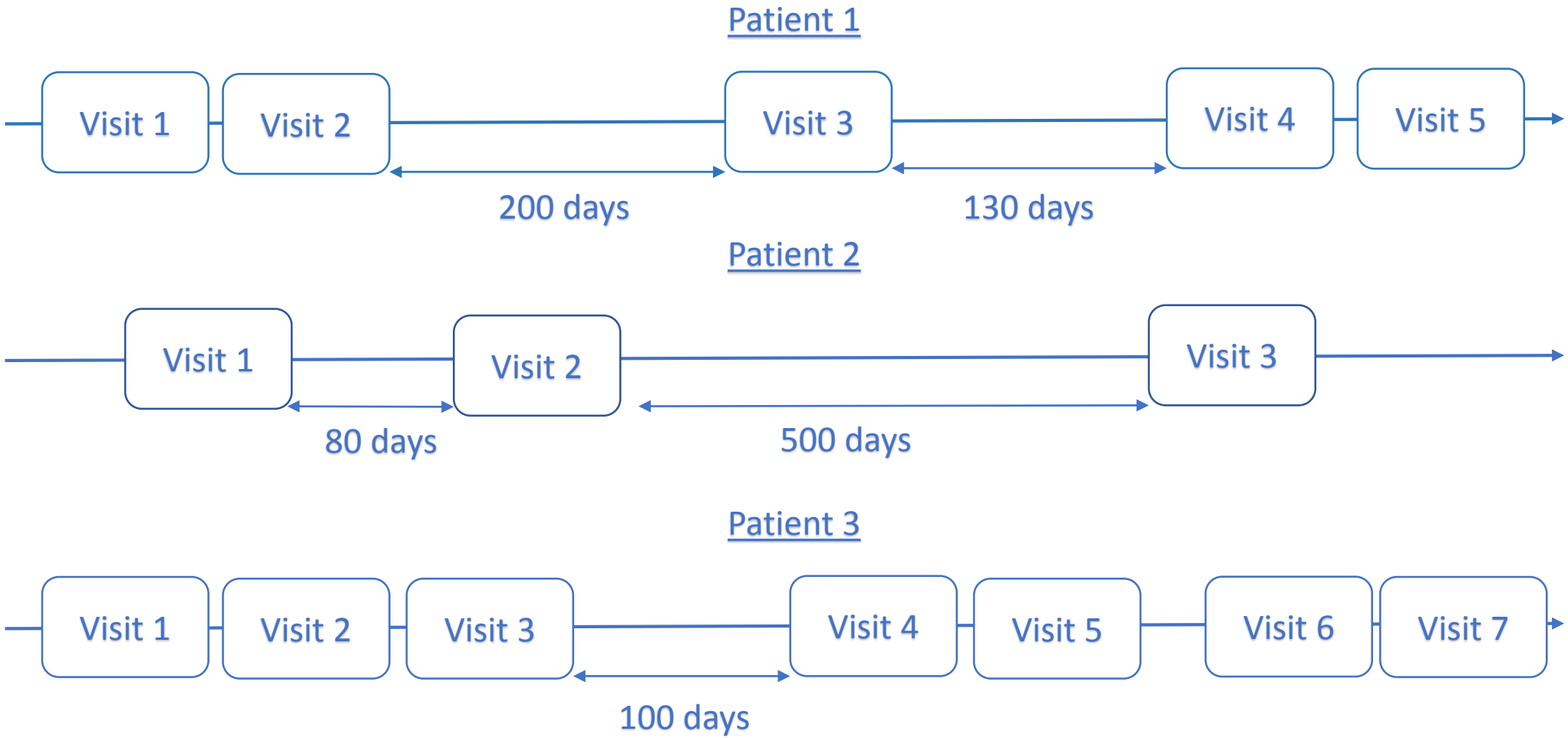


$$\mathcal{X} \approx [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$

Carroll, J. Douglas, and Jih-Jie Chang. "Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition." Psychometrika 35.3 (1970): 283-319.

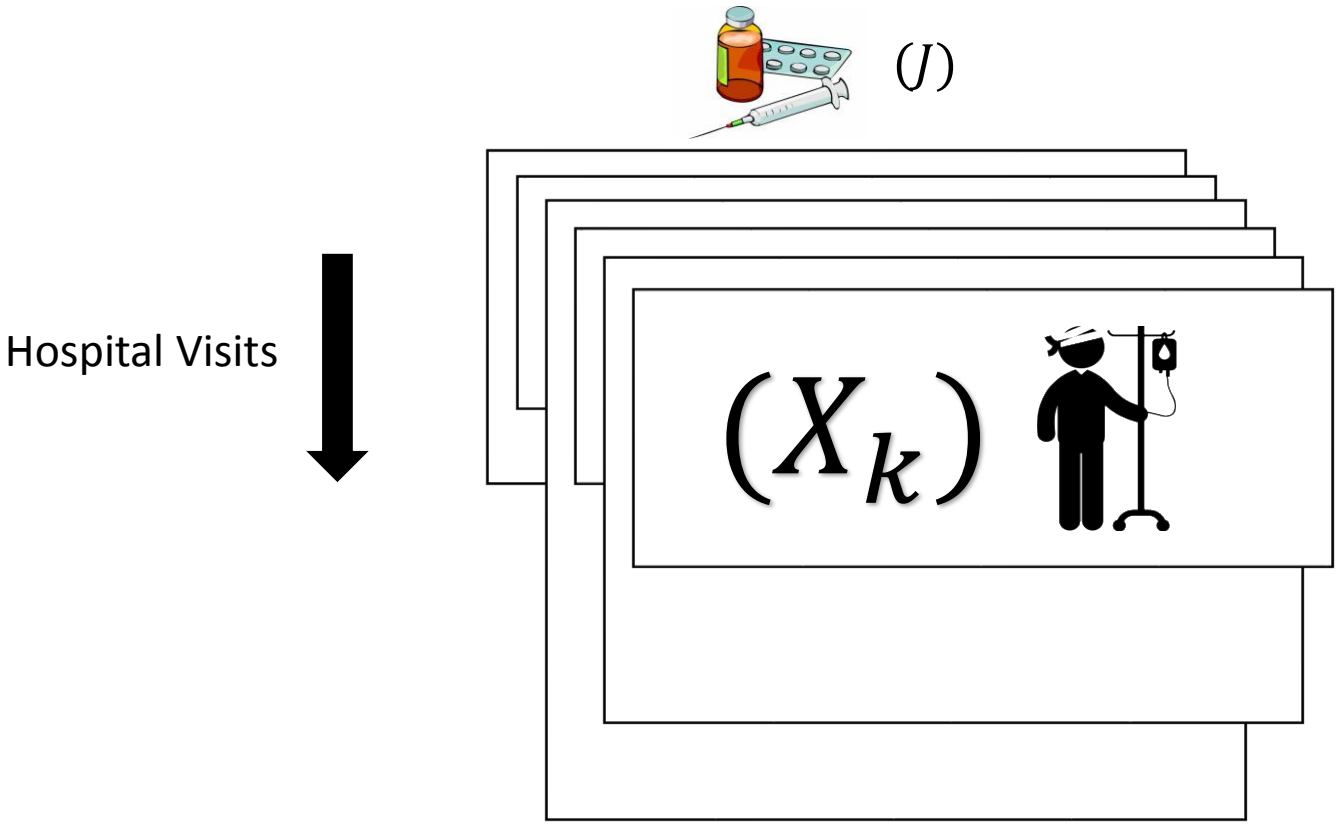
Challenges: Temporal Mismatch in Phenotyping

- Variable # hospital visits.



Background: PARAFAC2

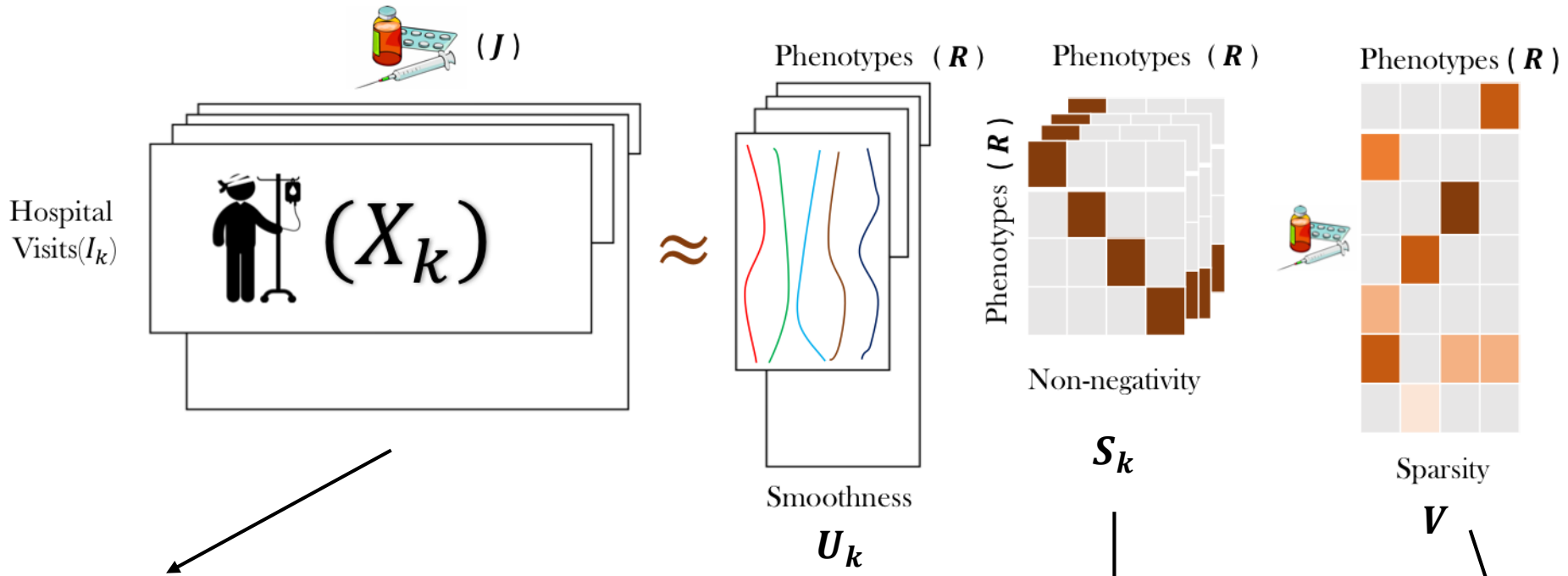
Input data: K subjects (patients), J medical feature and I_k hospital visits per patient.



$$X_k \in R^{I_k \times J}$$

Kiers, Henk AL, Jos MF Ten Berge, and Rasmus Bro. "PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model." *Journal of Chemometrics: A Journal of the Chemometrics Society* 13.3-4 (1999): 275-294.

Model Interpretation for phenotyping via EHRs



Input

Each matrix X_k represents patient k with J medical features and I_k hospital visits.

The r^{th} column of U_k represents the **evolution of phenotype** r.

The diagonal S_k provides the **importance membership** of R phenotype for patient k.

Each **column** of V represents a **phenotype**.

Contributions of COPA

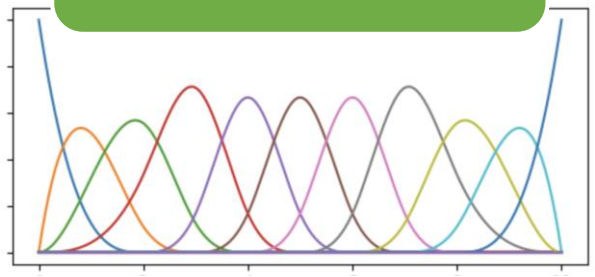
	Marble	Rubik	PARAFAC2	SPARTan	Helwig	COPA
Smoothness					✓	✓
Sparsity	✓	✓				✓
Scalability				✓		✓
Handle irregular tensors			✓	✓	✓	✓

- Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.
- Wang, Yichen, et al. "Rubik: Knowledge guided tensor factorization and completion for health data analytics." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.
- Kiers, Henk AL, Jos MF Ten Berge, and Rasmus Bro. "PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model." *Journal of Chemometrics: A Journal of the Chemometrics Society* 13.3-4 (1999): 275-294.
- Kiers, Henk AL, Jos MF Ten Berge, and Rasmus Bro. "PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model." Journal of Chemometrics: A Journal of the Chemometrics Society 13.3-4 (1999): 275-294.
- Helwig, Nathaniel E. "Estimating latent trends in multivariate longitudinal data via Parafac2 with functional and structural constraints." Biometrical Journal 59.4 (2017): 783-803.

Smoothness on U_k

$X_k \in R^{I_k \times J}$

$M_k \in R^{I_k \times l}$



SVD

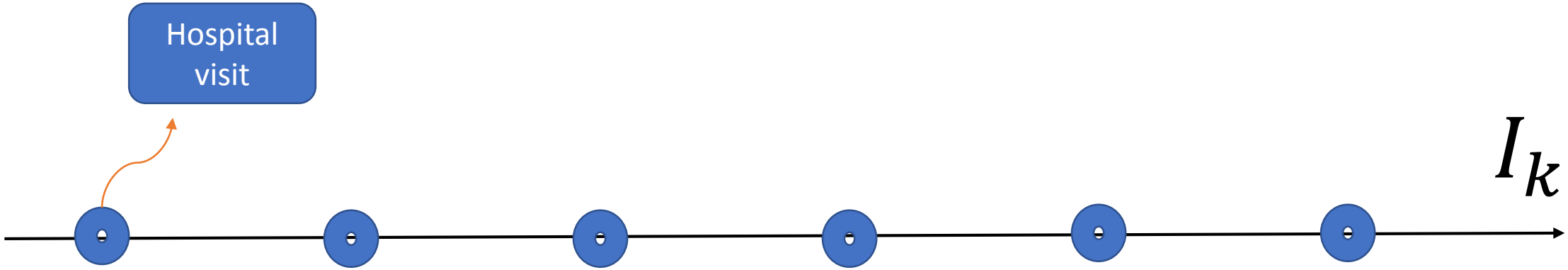
$[C_k, O_k, P_k^T]$

$X_k' = C_k^T X_k$

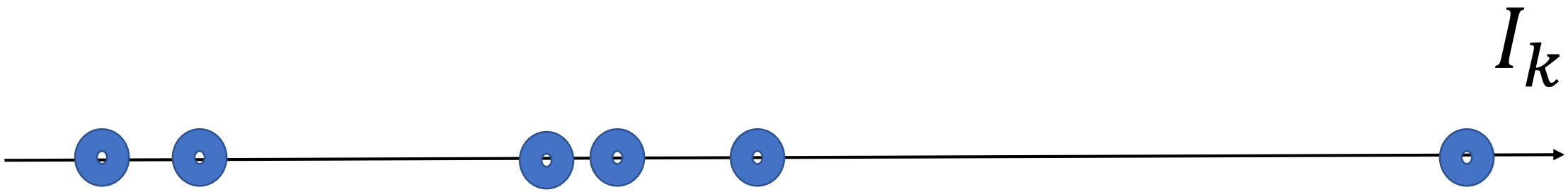
Apply COPA

U_k is smooth

Creating Basis Functions

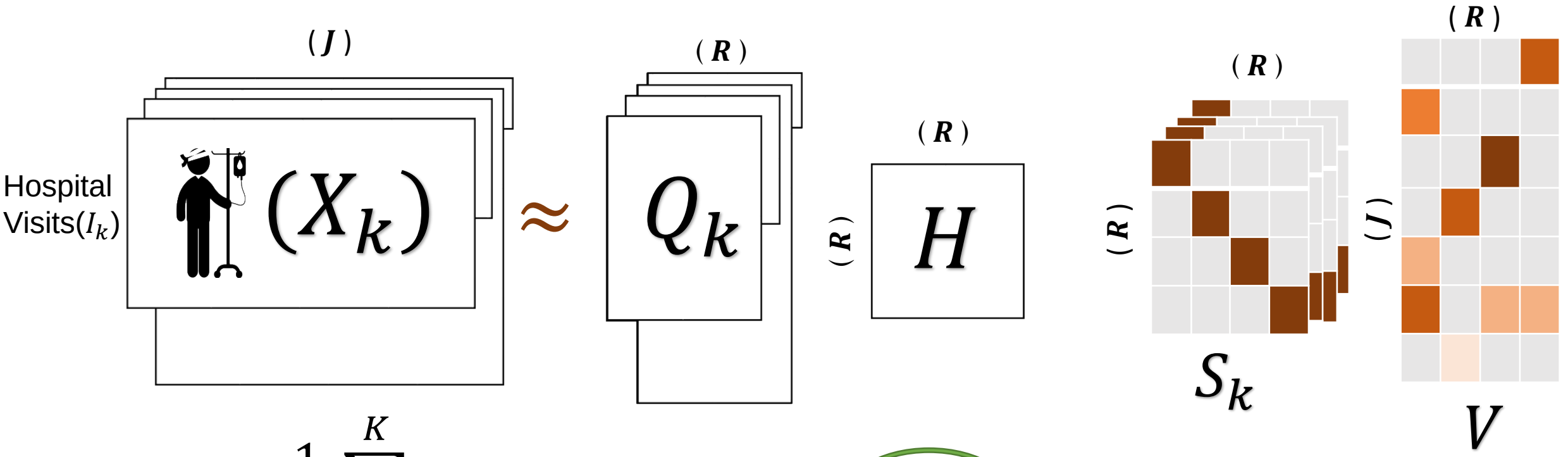


Ignoring the gap between two hospital visits



considers the gap between two hospital visits

COPA Framework



$$\underset{\{U_k\},\{S_k\},V}{\text{Minimize}} \ \frac{1}{2}\sum_{k=1}^{K}\left|\left|X_k - U_k S_k V^T\right|\right|_F^2 + \lambda\left|\left|V\right|\right|_0$$

Sparsity on V

$$U_k = Q_k H, Q_k^T Q_k = I \qquad \text{For all k=1,...K}$$

$$S_k \geq 0\,, \qquad \text{For all k=1,...K}$$
$$H \geq 0, V \geq 0$$

Non-negativity constraint

Solution for factor matrix $\{Q_k\}$

$$\underset{Q_k}{\text{Minimize}} \quad \frac{1}{2}\left|\left|X_k - Q_k H S_k V^T\right|\right|_F^2$$

Trace Properties

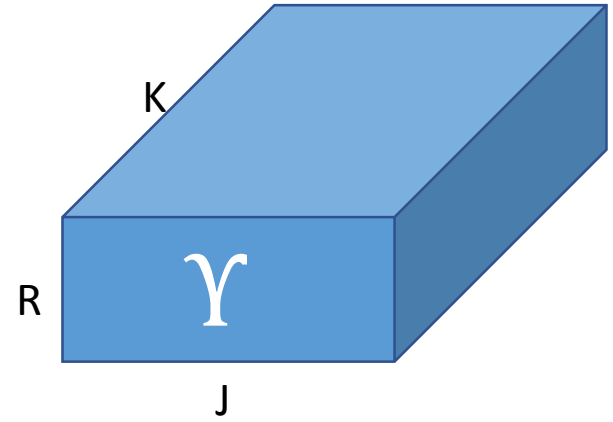
$$\underset{Q_k}{\text{Minimize}} \quad \frac{1}{2}\left|\left|X_k V S_k H^T - Q_k\right|\right|_F^2$$

Apply SVD

$$[B_k, \Sigma_k, C_k] = SVD(X_k V S_k H^T)$$
$$Q_k = B_k C_k^T$$

Solutions for factor matrices $\mathbf{H}, \{S_k\}, V$

$$\|X_k - Q_k H\, S_k V^T\|_F^2 \quad \xrightarrow{\times\ Q_k^T} \quad \left\|Q_K^T X_k - H\, S_k V^T\right\|_F^2$$



$$\underset{H,W,V}{\text{Minimize}} \quad \frac{1}{2}\left|\left|\Upsilon - [H,V,W]\right|\right|_F^2 + \lambda\left|\left|V\right|\right|_0$$

$$H \geq 0, V \geq 0, \mathrm{W} \geq 0,$$

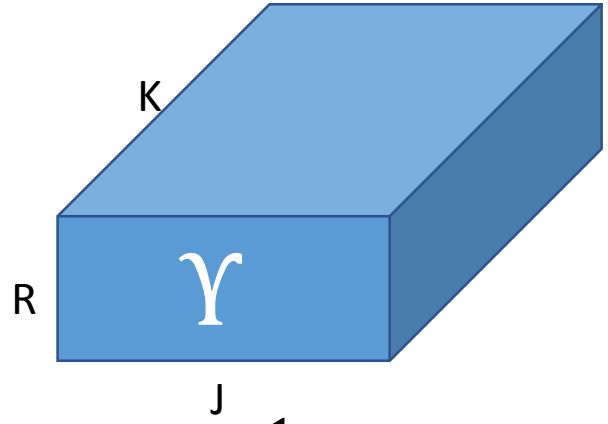
Where $S_k = diag\big(W(k,:)\big)$ and $W \in R^{K\times R}$

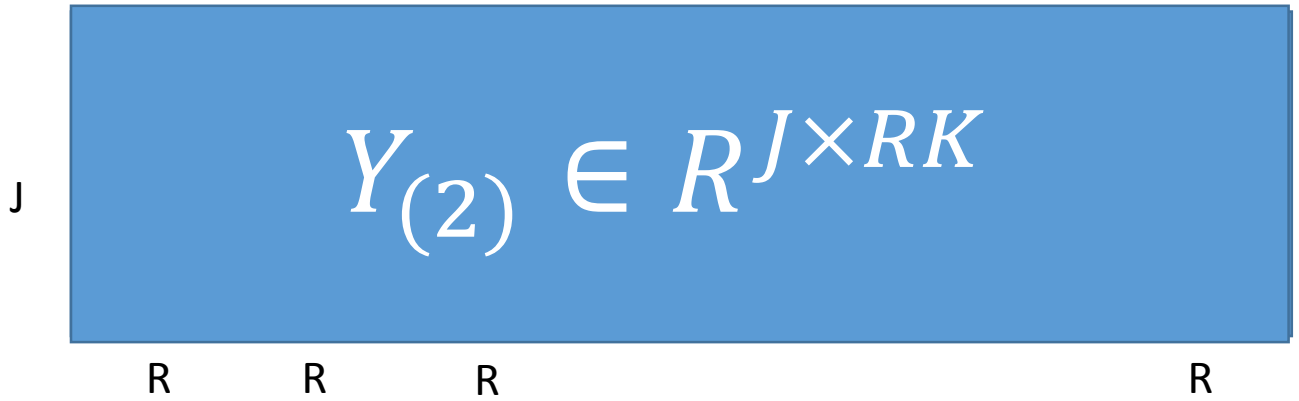
Sparsity on V

$$\underset{V}{\text{Minimize}} \quad \frac{1}{2}\left|\left|\Upsilon - [H, V, W]\right|\right|_F^2 + \lambda \left|\left|\overline{V}\right|\right|_0$$

$$V = \overline{V}$$

Auxiliary Variable handles sparsity





$$\underset{V}{\text{Minimize}} \quad \frac{1}{2}\left|\left|Y_{(2)}^T - (W \odot H)V^T\right|\right|_F^2 + \lambda \left|\left|\overline{V}\right|\right|_0$$

$$V^T = \overline{V}$$

Real Datasets Description

- Children Healthcare of Atlanta (CHOA):
  - ✓Medical features contain diagnosis and medications.
  - ✓Hospital visits are based on days of visits.
- Centers for Medicare and Medicaid (CMS):
  - ✓CMS is a set of realistic data set publicly available, however, protecting the privacy of patients.
  - ✓Medical features just contains diagnosis.

Dataset	# Patients	# Medical Features	# Hospital visits	#non-zero elements
CHOA	247,885	1388	857	11 Million
CMS	843,162	284	1500	84 Million

Evaluation Metrics

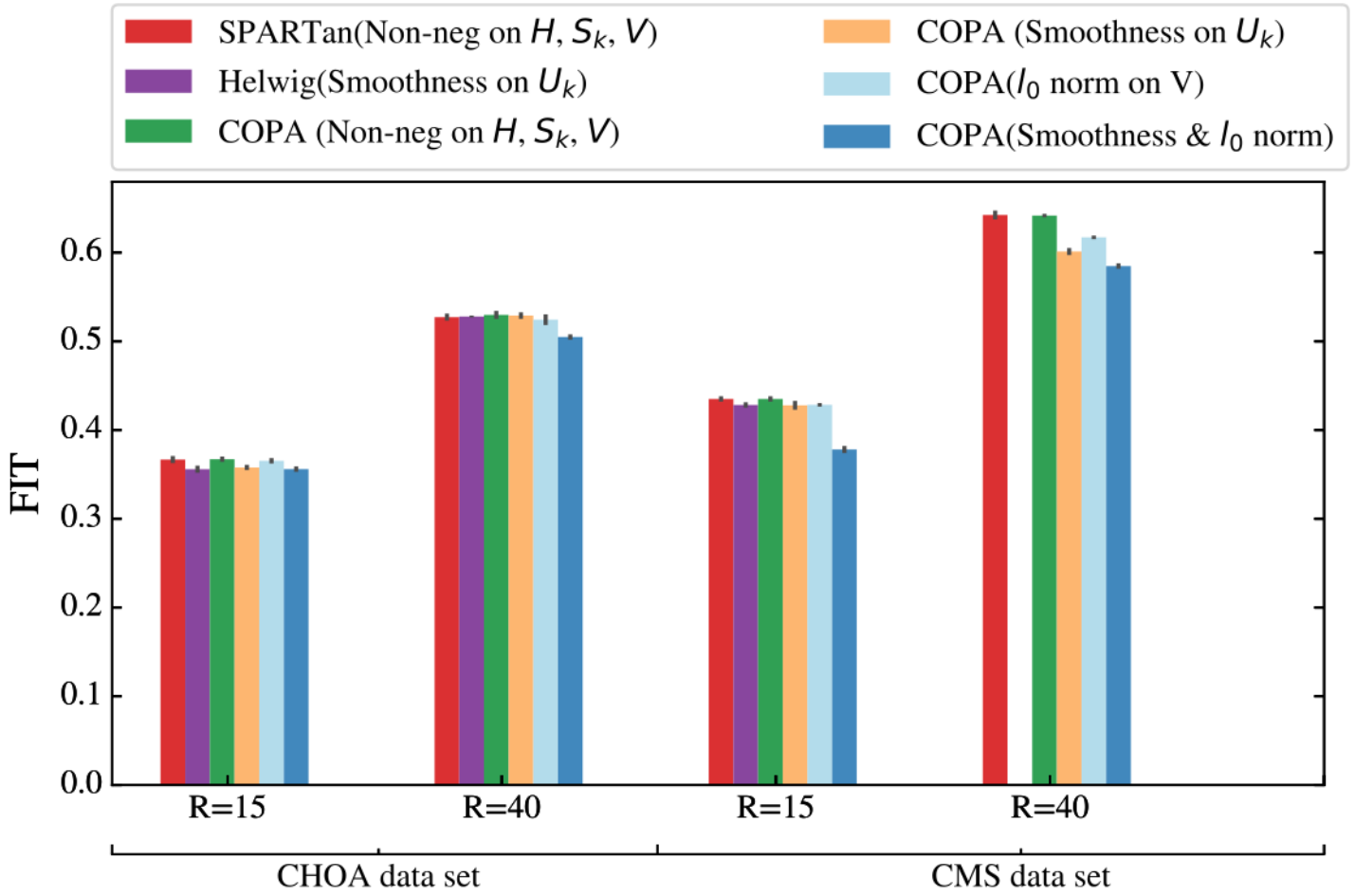
FIT $= 1 - \frac{\sum_{k=1}^{K} \|X_k - U_k S_k V^T\|_F^2}{\sum_{k=1}^{K} \|X_k\|_F^2}$ — Higher better

Sparsity $= \frac{nz(V)}{size(V)}$ — Higher better

Running-Time (In Seconds) — Lower better

Quantitative assessment of Constraints

This plot shows the impact of each constraint on the FIT values across both datasets for two different target ranks (R={15,40})



Higher better

R={15,40}

CHOA, CMS

* The missing purple bar in the forth column is out of memory failure for Helwig method.

Quantitative assessment of Constraints

Sparsity Metric

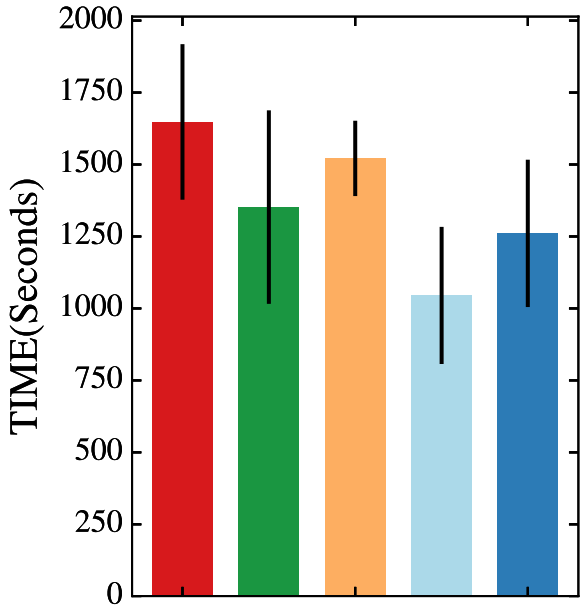
	CHOA		CMS	
Algorithm	R=15	R=40	R=15	R=40
COPA	**0.9886±0.0035**	**0.9897±0.0027**	**0.995±0.0001**	**0.9963±0.0002**
SPARTan	0.7127±0.0161	0.8127±0.0029	0.1028±0.0032	0.2164±0.0236

Higher better

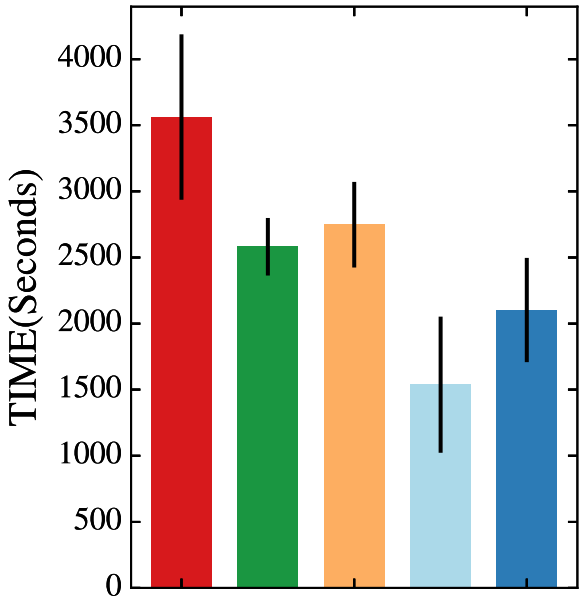
5 different random initialization

Scalability

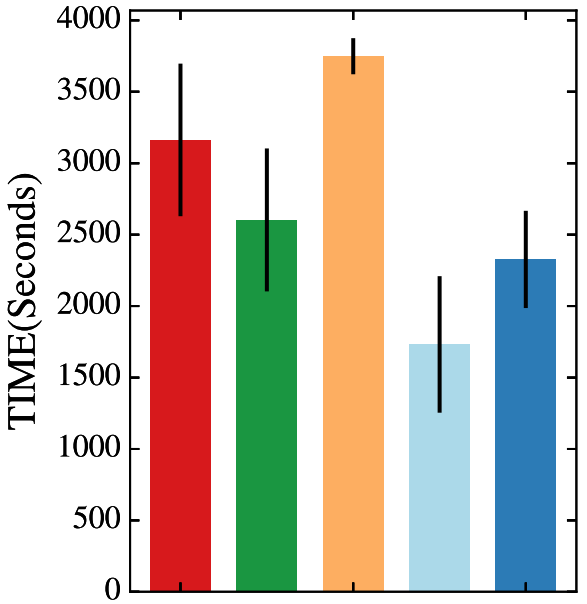
Lower better



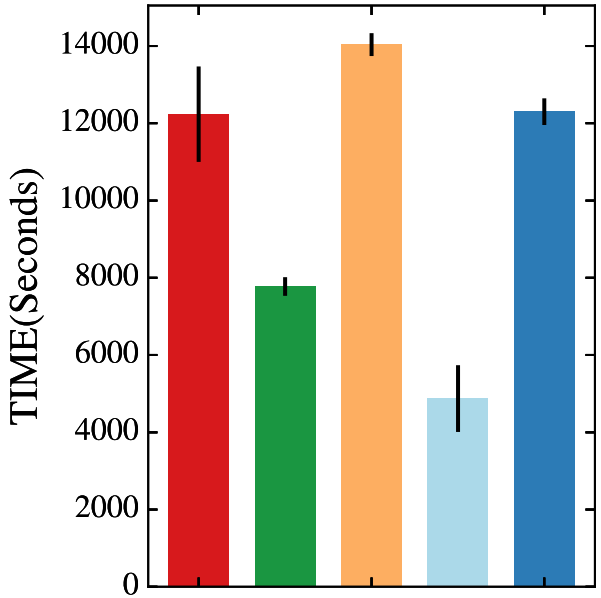
CHOA, R=15



CHOA, R=40



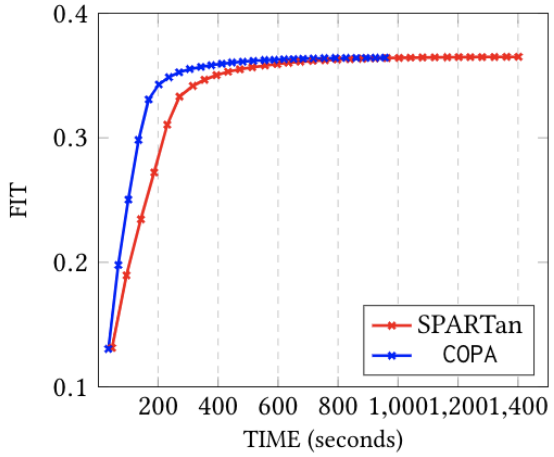
CMS, R=15



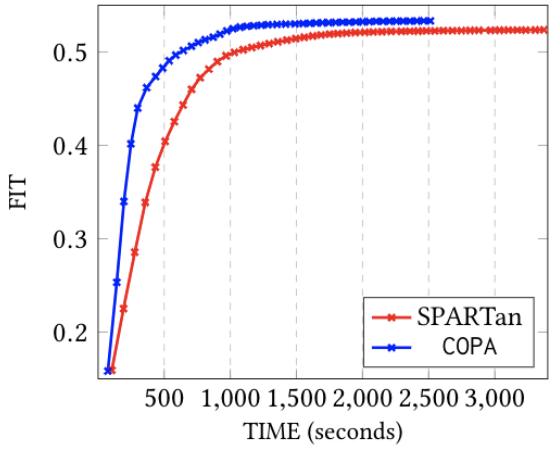
CMS, R=40

The Total Running Time comparison (average and standard deviation) in seconds for different versions of COPA and SPARTan for 5 different random initializations.

FIT-TIME (Convergence)

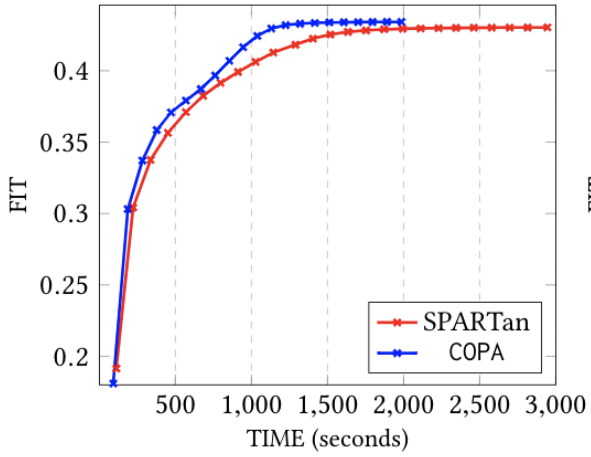


(a) 15 Components

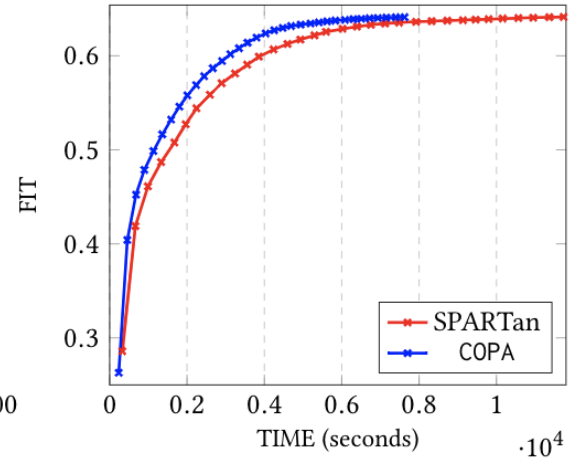


(b) 40 Components

CHOA



(a) 15 Components.



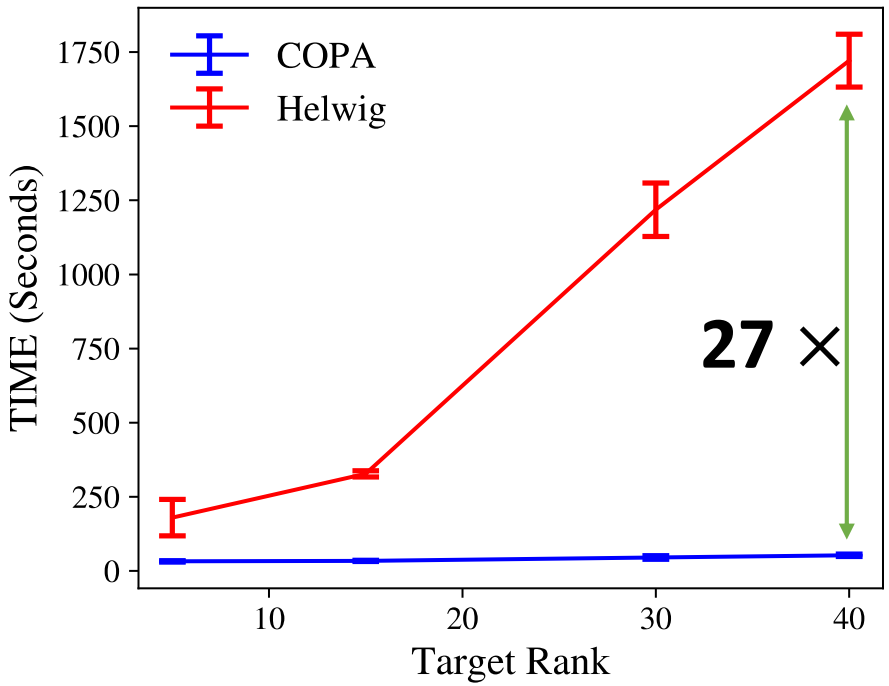
(b) 40 Components.

CMS

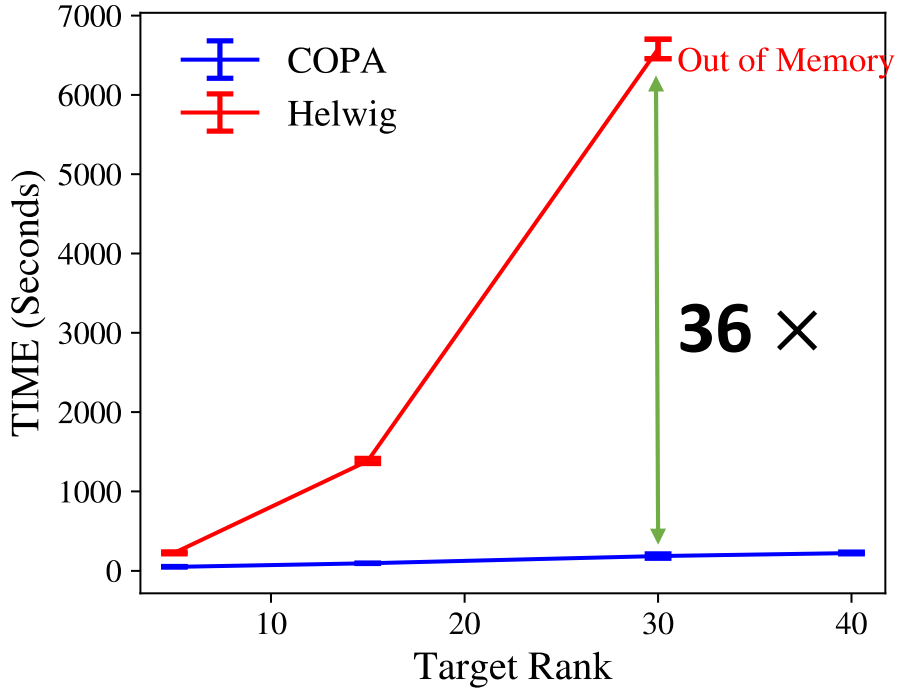
The best Convergence of COPA and SPARTan out of 5 different random initializations with non-negativity constraint on $H, \{S_k\}, V$ on CHOA, CMS data sets for different target ranks (two cases considered: R={15,40})

Scalability

Time in seconds for one iteration (as an average of 5) for different values of R.



CHOA



CMS

Case Study: CHOA Phenotype Discovery

Focus on Medically Complex Patients (MCPs)

Children's SM
Healthcare of Atlanta

A total of **4602** patients are selected with **810** distinct medical features.
We extracted 4 number of phenotypes.

Our goal is:

- Extracting the phenotypes
- Find the temporal evolution of phenotypes for each patient.

Phenotypes Discovered by COPA

Leukemias

Leukemias
Immunity disorders
Deficiency and other anemia
HEPARIN AND RELATED PREPARATIONS
Maintenance chemotherapy; radiotherapy
ANTIEMETIC/ANTIVERTIGO AGENTS
SODIUM/SALINE PREPARATIONS
TOPICAL LOCAL ANESTHETICS
GENERAL ANESTHETICS INJECTABLE
ANTINEOPLASTIC - ANTIMETABOLITES
ANTIHISTAMINES - 1ST GENERATION
ANALGESIC/ANTIPYRETICS NON-SALICYLATE
ANALGESICS NARCOTIC ANESTHETIC ADJUNCT AGENTS
ABSORBABLE SULFONAMIDE ANTIBACTERIAL AGENTS
GLUCOCORTICOIDS

Neurological Disorders

Other nervous system disorders
Epilepsy; convulsions
Paralysis
Other connective tissue disease
Developmental disorders

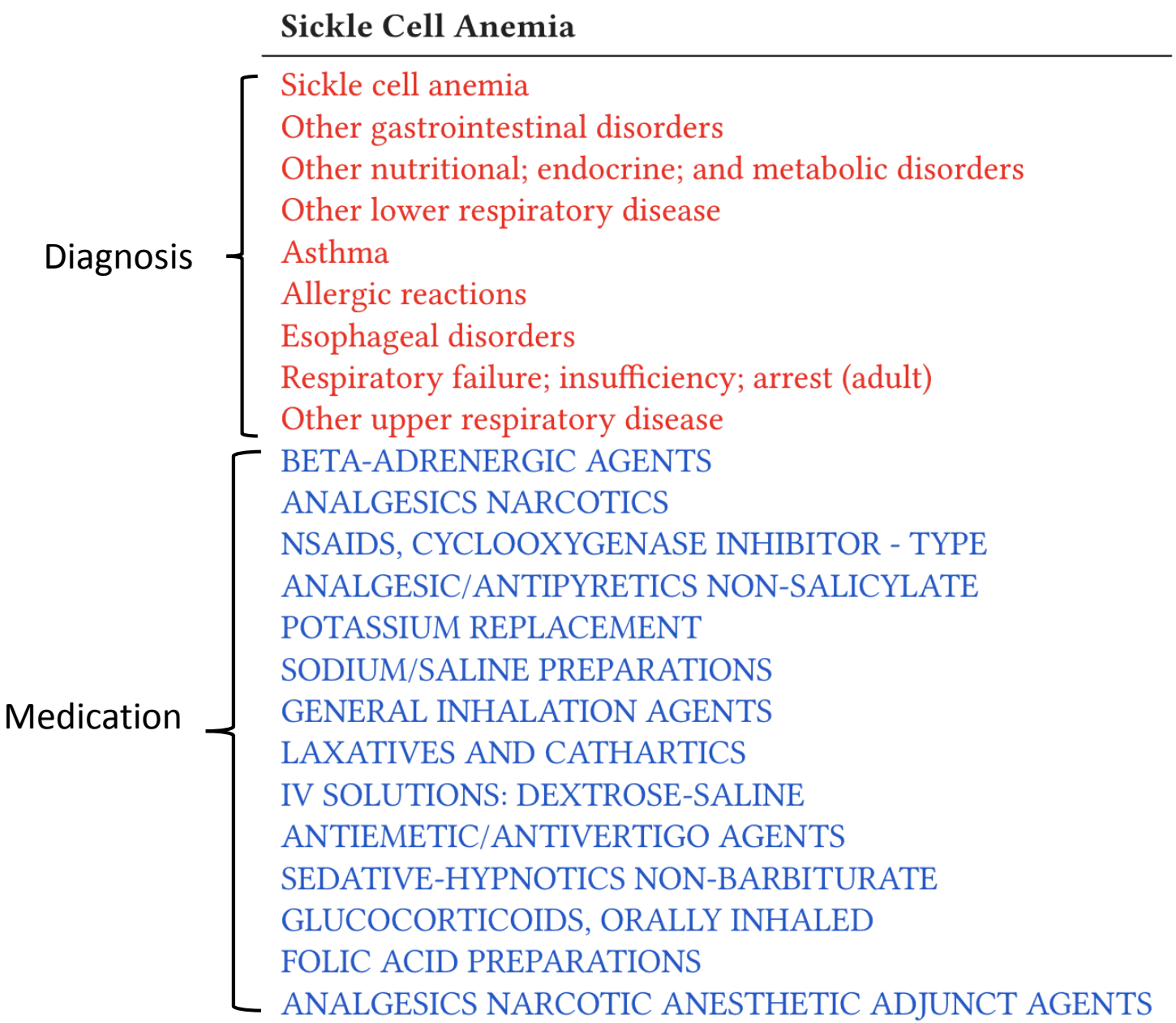
Rehabilitation care; and adjustment of devices
ANTICONVULSANTS

Congenital anomalies

Other perinatal conditions
Cardiac and circulatory congenital anomalies
Short gestation; low birth weight
Other congenital anomalies

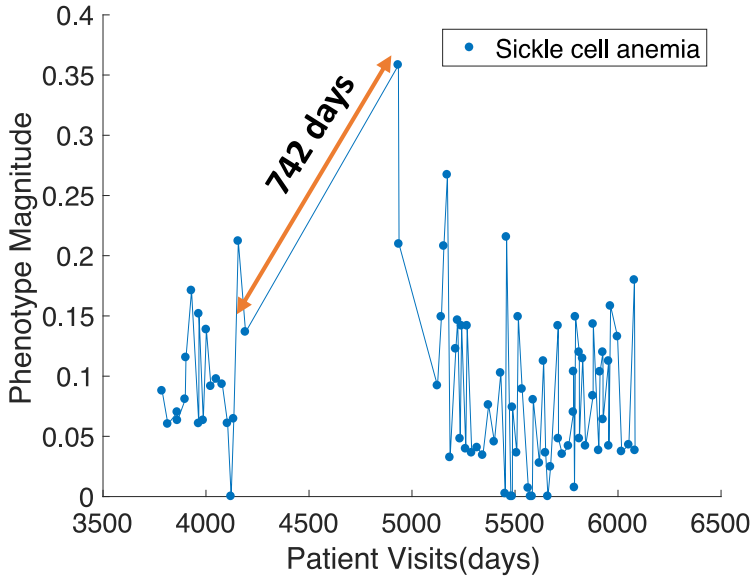
Fluid and electrolyte disorders
LOOP DIURETICS
IV FAT EMULSIONS

No additional post-processing was performed on these results.

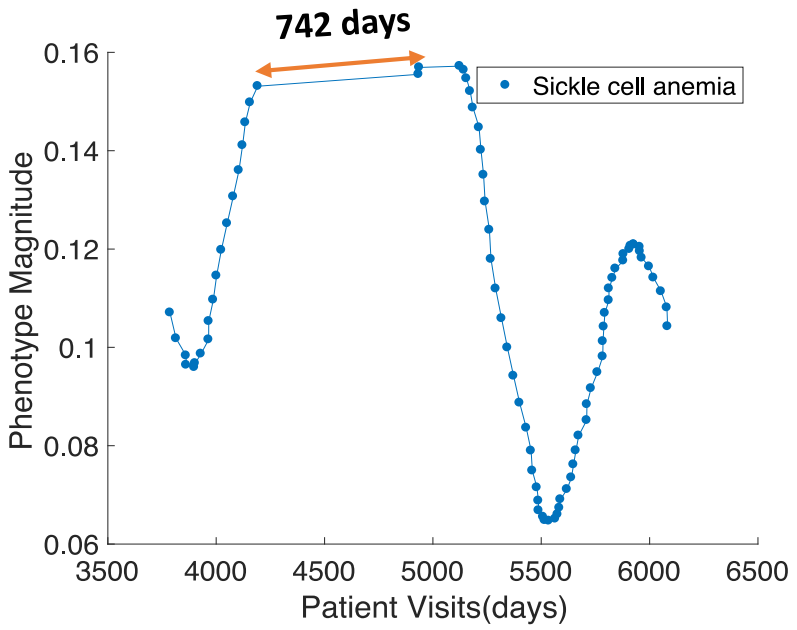


Title annotation is provided by a medical expert.

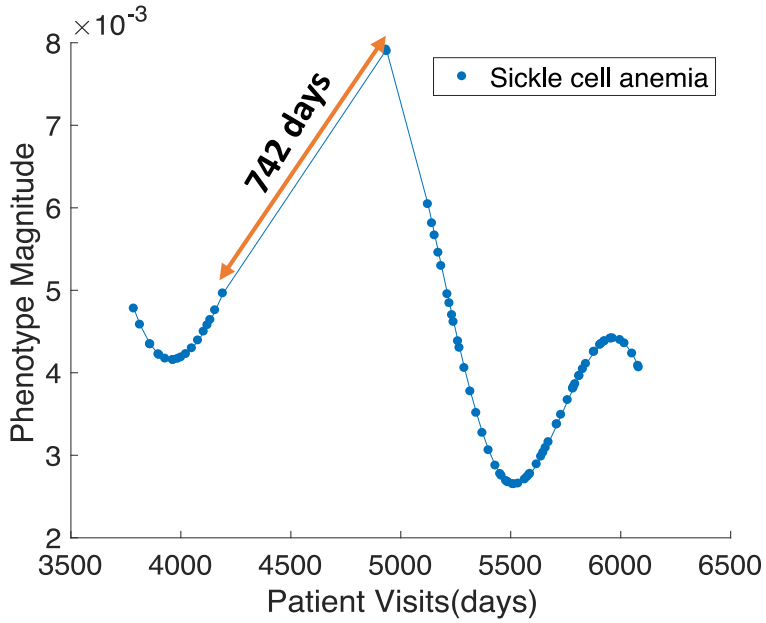
Temporal Phenotyping



SPARTan



Helwig



COPA

aafshar8@gatech.edu

http://www.prism.gatech.edu/~aafshar8/

https://github.com/aafshar/COPA

Ari Afshar

Kimis Perros

Vagelis Papalexakis

Bess Searles

Joyce Ho

Jimeng Sun

SunLab
Enlighten the World

Georgia Tech

UC RIVERSIDE
UNIVERSITY OF CALIFORNIA

EMORY

Children's
Healthcare of Atlanta