Illustration by Chris Brigman

Practical Leverage-Based Sampling for Low-Rank Tensor Decomposition

Tamara G. Kolda

Sandia National Labs, Livermore, CA

www.kolda.net

Joint work with

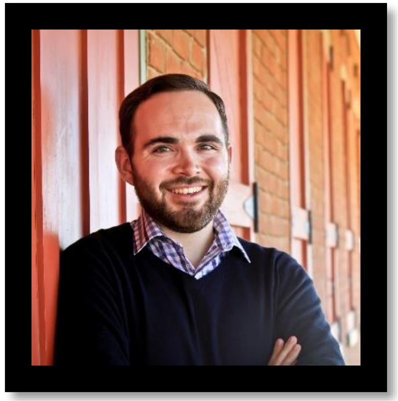
Brett Larsen

Stanford University

Supported by the DOE Office of Science Advanced Scientific Computing Research (ASCR) Applied Mathematics. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

Funding & Reference

- Tammy & Brett were funded by
  Department of Energy (DOE) Office of Science
  Advanced Scientific Computing Research (ASCR)
  Applied Mathematics Program



- Brett was also funded by
  DOE Computational Science Graduate Fellowship
  (CSGF), administered by the Krell Institute

- B. W. Larsen, T. G. Kolda. **Practical Leverage-Based Sampling for Low-Rank Tensor Decomposition**. arXiv:2006.16438,2020. http://arxiv.org/abs/2006.16438

A Tensor is an Multi-Way Array

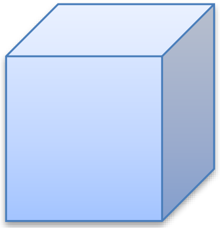
1st-order Tensor (vector)

$\mathbf{x}$

2nd-order Tensor (matrix)

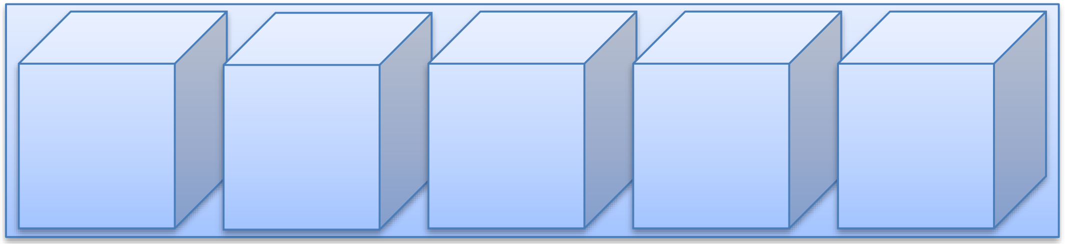
$\mathbf{X}$

3rd-Order Tensor



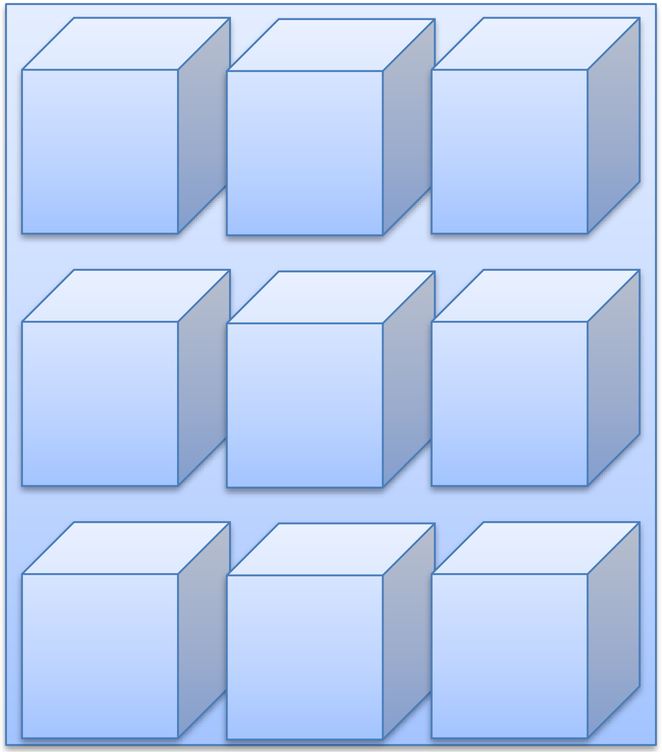
$\mathcal{X}$

4th-Order Tensor



$\mathcal{X}$

5th-Order Tensor

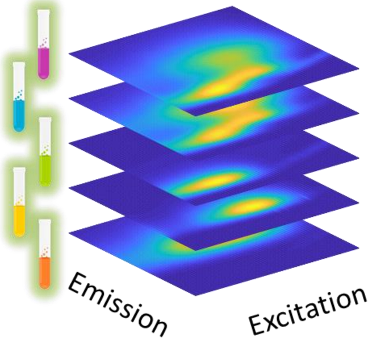


$\mathcal{X}$

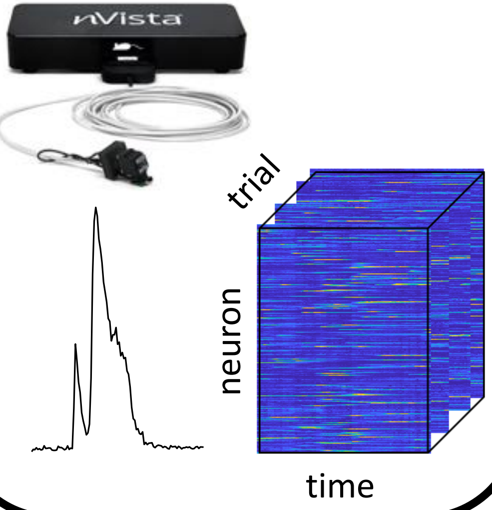
Tensors Come From Many Applications

- **Chemometrics:** Emission x Excitation x Samples (Fluorescence Spectroscopy)
- **Neuroscience:** Neuron x Time x Trial
- **Criminology:** Day x Hour x Location x Crime (Chicago Crime Reports)
- **Machine Learning:** Multivariate Gaussian Mixture Models Higher-Order Moments
- **Transportation:** Pickup x Dropoff x Time (Taxis)
- **Sports:** Player x Statistic x Season (Basketball)
- **Cyber-Traffic:** IP x IP x Port x Time
- **Social Network:** Person x Person x Time x Interaction-Type
- **Signal Processing:** Sensor x Frequency x Time
- **Trending Co-occurrence:** Term A x Term B x Time

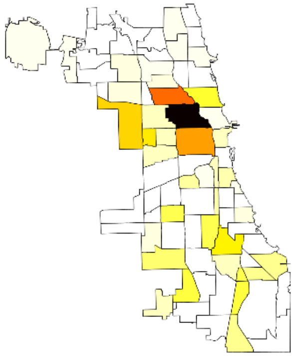
Chemometrics



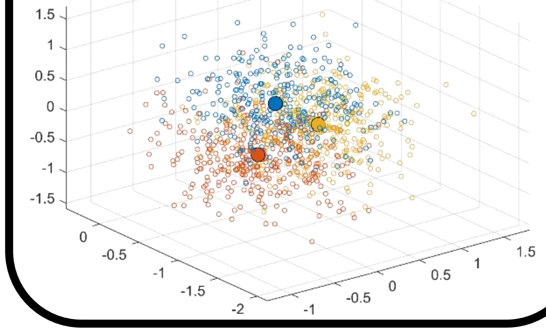
Neuroscience



Criminology



Machine Learning

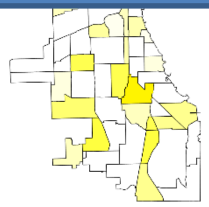


Tensors Come From Many Applications

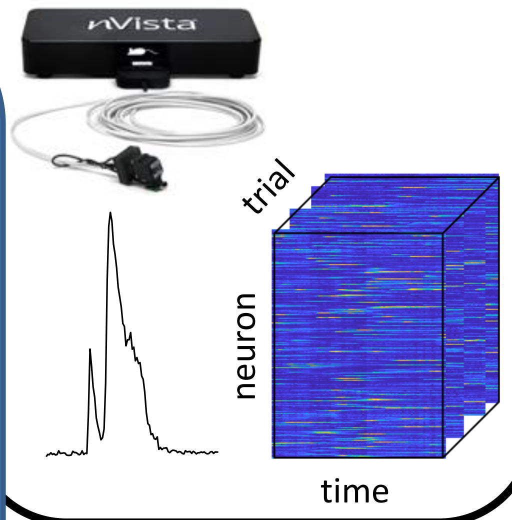
- **Chemometrics:** Emission x Excitation x Samples (Fluorescence Spectroscopy)
- **Neuroscience:** N
- **Criminology:** D (Chicago Crime
- **Machine Learn** Mixture Model
- **Transportation**
- **Sports:** Player
- **Cyber-Traffic:** I
- **Social Network** Interaction-Type
- **Signal Processing:** Sensor x Frequency x Time
- **Trending Co-occurrence:** Term A x Term B x Time

Chemometrics

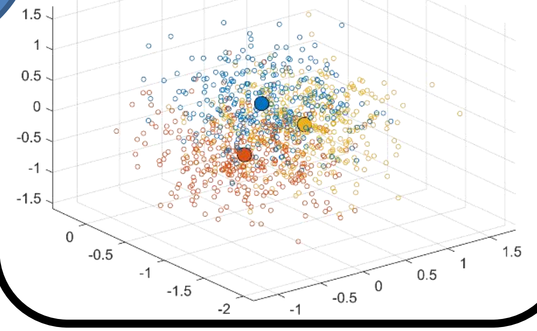
Tensor Decomposition Finds Patterns in Massive Data (Unsupervised Learning)



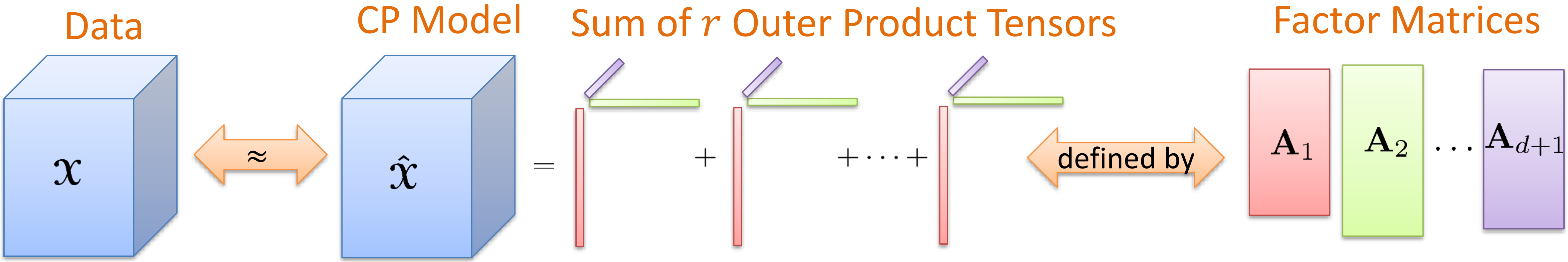
Neuroscience



Machine Learning



Tensor Decomposition Identifies Factors



$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

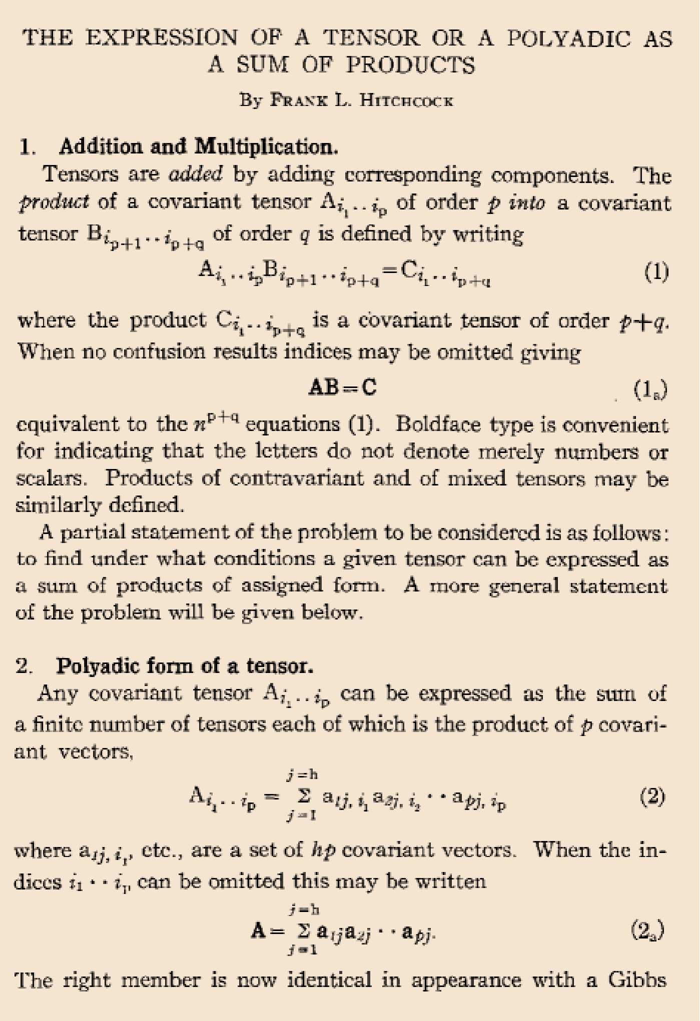
$[\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{d+1}]\!] \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

$\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$

Model Rank

CP First Invented in 1927

Frank Lauren Hitchcock
MIT Professor
(1875–1957)



THE EXPRESSION OF A TENSOR OR A POLYADIC AS A SUM OF PRODUCTS

By Frank L. Hitchcock

1. Addition and Multiplication.

Tensors are *added* by adding corresponding components. The *product* of a covariant tensor $A_{i_1 \cdots i_p}$ of order p *into* a covariant tensor $B_{i_{p+1} \cdots i_{p+q}}$ of order q is defined by writing

$$A_{i_1 \cdots i_p} B_{i_{p+1} \cdots i_{p+q}} = C_{i_1 \cdots i_{p+q}} \tag{1}$$

where the product $C_{i_1 \cdots i_{p+q}}$ is a covariant tensor of order $p+q$. When no confusion results indices may be omitted giving

$$\mathbf{AB} = \mathbf{C} \tag{1$_a$}$$

equivalent to the n^{p+q} equations (1). Boldface type is convenient for indicating that the letters do not denote merely numbers or scalars. Products of contravariant and of mixed tensors may be similarly defined.

A partial statement of the problem to be considered is as follows: to find under what conditions a given tensor can be expressed as a sum of products of assigned form. A more general statement of the problem will be given below.

2. Polyadic form of a tensor.

Any covariant tensor $A_{i_1 \cdots i_p}$ can be expressed as the sum of a finite number of tensors each of which is the product of p covariant vectors,

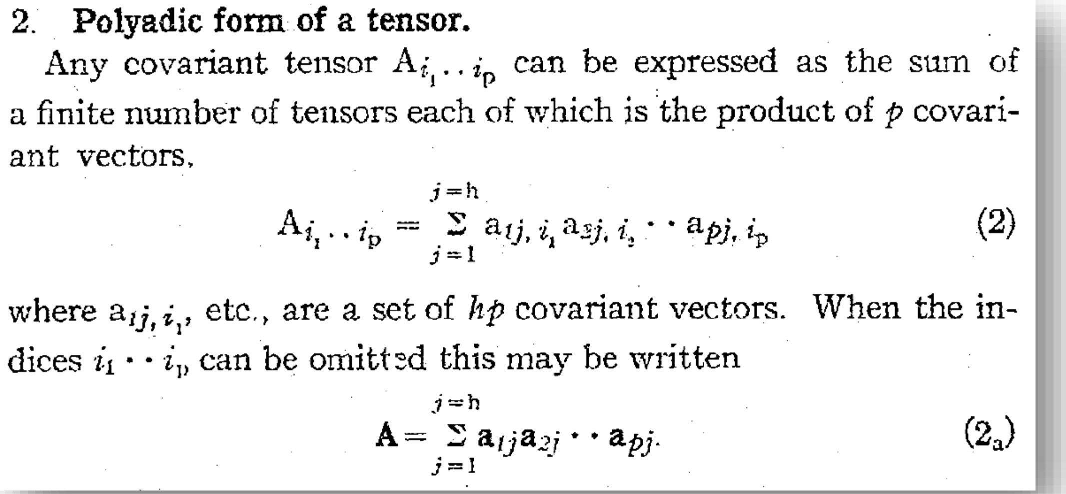
$$A_{i_1 \cdots i_p} = \sum_{j=1}^{j=h} a_{1j,\, i_1} a_{2j,\, i_2} \cdots a_{pj,\, i_p} \tag{2}$$

where $a_{1j,\, i_1}$, etc., are a set of hp covariant vectors. When the indices $i_1 \cdots i_p$ can be omitted this may be written

$$\mathbf{A} = \sum_{j=1}^{j=h} \mathbf{a}_{1j} \mathbf{a}_{2j} \cdots \mathbf{a}_{pj}. \tag{2$_a$}$$

The right member is now identical in appearance with a Gibbs

F. L. Hitchcock, ***The Expression of a Tensor or a Polyadic as a Sum of Products***, Journal of Mathematics and Physics, 1927



2. Polyadic form of a tensor.

Any covariant tensor $A_{i_1 \cdots i_p}$ can be expressed as the sum of a finite number of tensors each of which is the product of p covariant vectors,

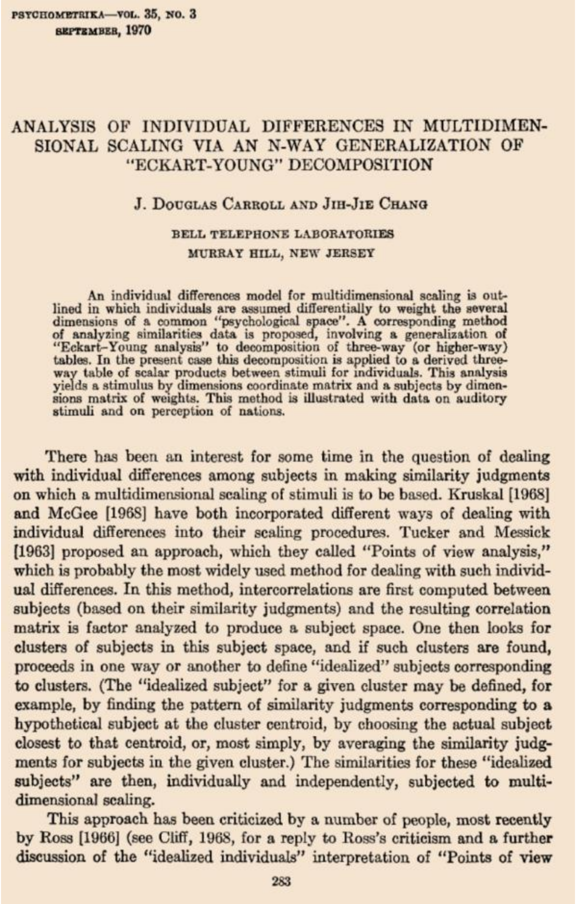
$$A_{i_1 \cdots i_p} = \sum_{j=1}^{j=h} a_{1j,\, i_1} a_{2j,\, i_2} \cdots a_{pj,\, i_p} \tag{2}$$

where $a_{1j,\, i_1}$, etc., are a set of hp covariant vectors. When the indices $i_1 \cdots i_p$ can be omitted this may be written

$$\mathbf{A} = \sum_{j=1}^{j=h} \mathbf{a}_{1j} \mathbf{a}_{2j} \cdots \mathbf{a}_{pj}. \tag{2$_a$}$$

CP Independently Reinvented (twice) in 1970

CANDECOMP: Canonical Decomposition



PSYCHOMETRIKA—VOL. 35, NO. 3
SEPTEMBER, 1970

ANALYSIS OF INDIVIDUAL DIFFERENCES IN MULTIDIMENSIONAL SCALING VIA AN N-WAY GENERALIZATION OF "ECKART-YOUNG" DECOMPOSITION

J. DOUGLAS CARROLL AND JIH-JIE CHANG

BELL TELEPHONE LABORATORIES
MURRAY HILL, NEW JERSEY

An individual differences model for multidimensional scaling is outlined in which individuals are assumed differentially to weight the several dimensions of a common "psychological space". A corresponding method of analyzing similarities data is proposed, involving a generalization of "Eckart–Young analysis" to decomposition of three-way (or higher-way) tables. In the present case this decomposition is applied to a derived three-way table of scalar products between stimuli for individuals. This analysis yields a stimulus by dimensions coordinate matrix and a subjects by dimensions matrix of weights. This method is illustrated with data on auditory stimuli and on perception of nations.

There has been an interest for some time in the question of dealing with individual differences among subjects in making similarity judgments on which a multidimensional scaling of stimuli is to be based. Kruskal [1968] and McGee [1968] have both incorporated different ways of dealing with individual differences into their scaling procedures. Tucker and Messick [1963] proposed an approach, which they called "Points of view analysis," which is probably the most widely used method for dealing with such individual differences. In this method, intercorrelations are first computed between subjects (based on their similarity judgments) and the resulting correlation matrix is factor analyzed to produce a subject space. One then looks for clusters of subjects in this subject space, and if such clusters are found, proceeds in one way or another to define "idealized" subjects corresponding to clusters. (The "idealized subject" for a given cluster may be defined, for example, by finding the pattern of similarity judgments corresponding to a hypothetical subject at the cluster centroid, by choosing the actual subject closest to that centroid, or, most simply, by averaging the similarity judgments for subjects in the given cluster.) The similarities for these "idealized subjects" are then, individually and independently, subjected to multidimensional scaling.

This approach has been criticized by a number of people, most recently by Ross [1966] (see Cliff, 1968, for a reply to Ross's criticism and a further discussion of the "idealized individuals" interpretation of "Points of view

283

J. Douglas Carroll
Bell Labs
(1939-2011)

Jih-Jie Chang
Bell Labs
(1927-2007)

Richard A. Harshman
Univ. Ontario
(1943-2008)

PARAFAC: Parallel Factors

NOTE: This manuscript was originally published in 1970 and is reproduced here to make it more accessible to interested scholars. The original reference is
Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics, 16*, 1-84. (University Microfilms, Ann Arbor, Michigan, No. 10,085).

FOUNDATIONS OF THE PARAFAC PROCEDURE: MODELS AND CONDITIONS FOR AN "EXPLANATORY" MULTIMODAL FACTOR ANALYSIS

by

Richard A. Harshman

UCLA

Working Papers in Phonetics

16

December, 1970

CP: CANDECOMP/PARAFAC — In 2000, Henk Kiers proposed this *compromise* name

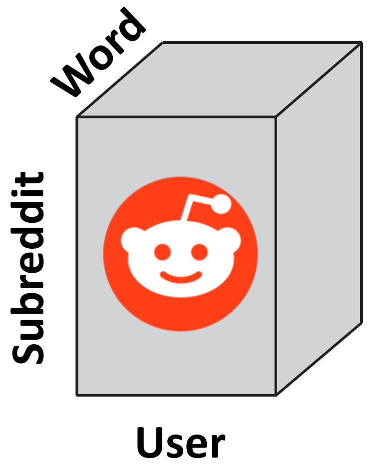
CP: Canonical Polyadic — 2010: Pierre Comon, Lieven DeLathauwer, and others reverse-engineered CP, revising some of Hitchcock's terminology

Many thanks to the following persons for helping me learn about Jih-Jie Chang: Fan Chung, Ron Graham, Shen Lin (husband), May Chang (niece), Lili Bruer (daughter).

Example Sparse Multiway Data: Reddit

- Reddit is an American social news aggregator, web content rating, and discussion website
  - A "subreddit" is a discussion forum on a particular topic
- Tensor obtained from frost.io (http://frostt.io/tensors/reddit-2015/)
  - Built from reddit comments posted in the year 2015
  - Users and words with less than 5 entries have been removed

For perspective, chance of being struck by lightening in your life $\approx$ 1 in 10^6



Reddit Tensor

8 million users

200 thousand subreddits

8 million words

4.7 billion non-zeros (>1 in 10^9)

106 gigabytes

$x(i,j,k) = \log$ (1 + the number of times user i used word j in subreddit k)

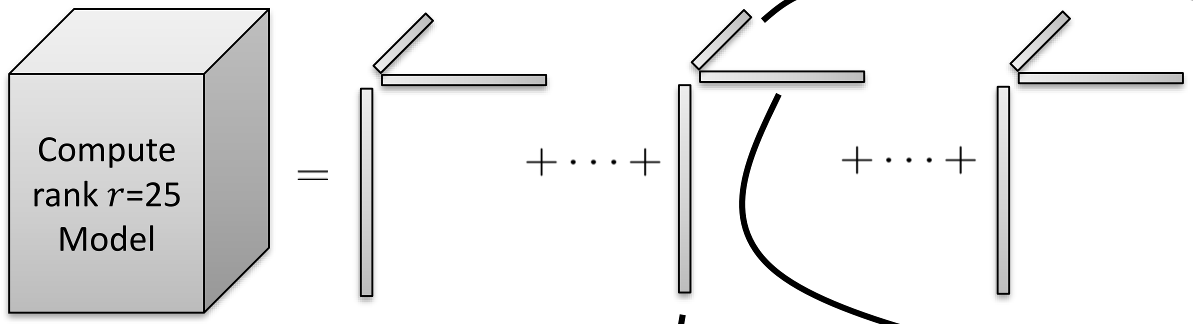
Used a rank $\boldsymbol{r = 25}$ decompsition

Smith et al (2017). "FROSTT: The Formidable Open Repository of Sparse Tensors and Tools"

Interpreting Reddit Components

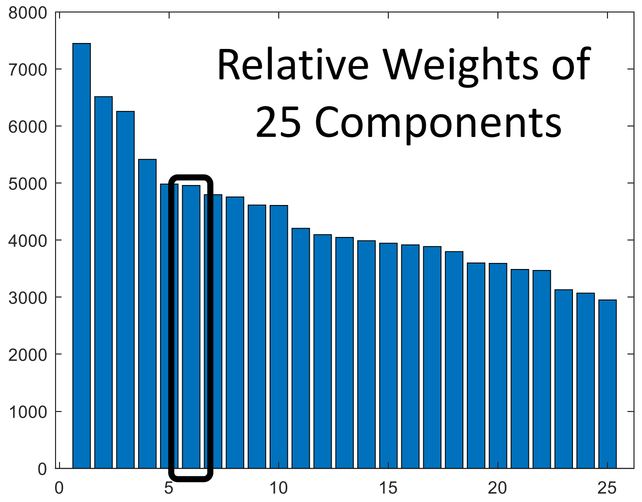
Reddit Tensor

Component 6

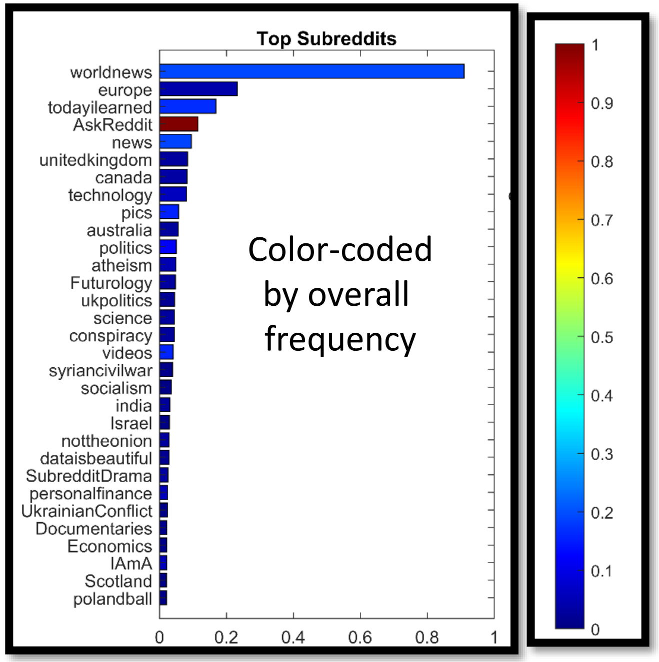


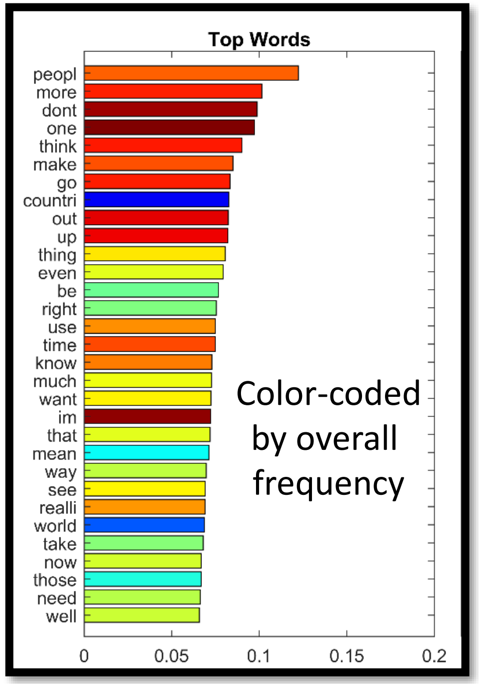
Largest entries, in absolute value

variance explained ≈ 6%

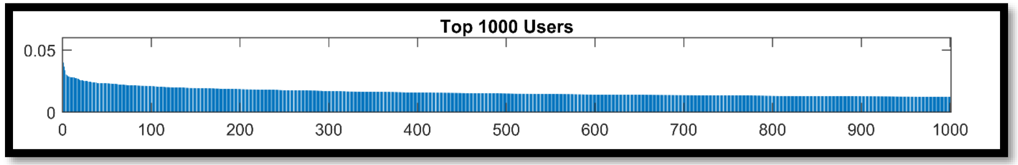


Largest of 200K entries, in absolute value



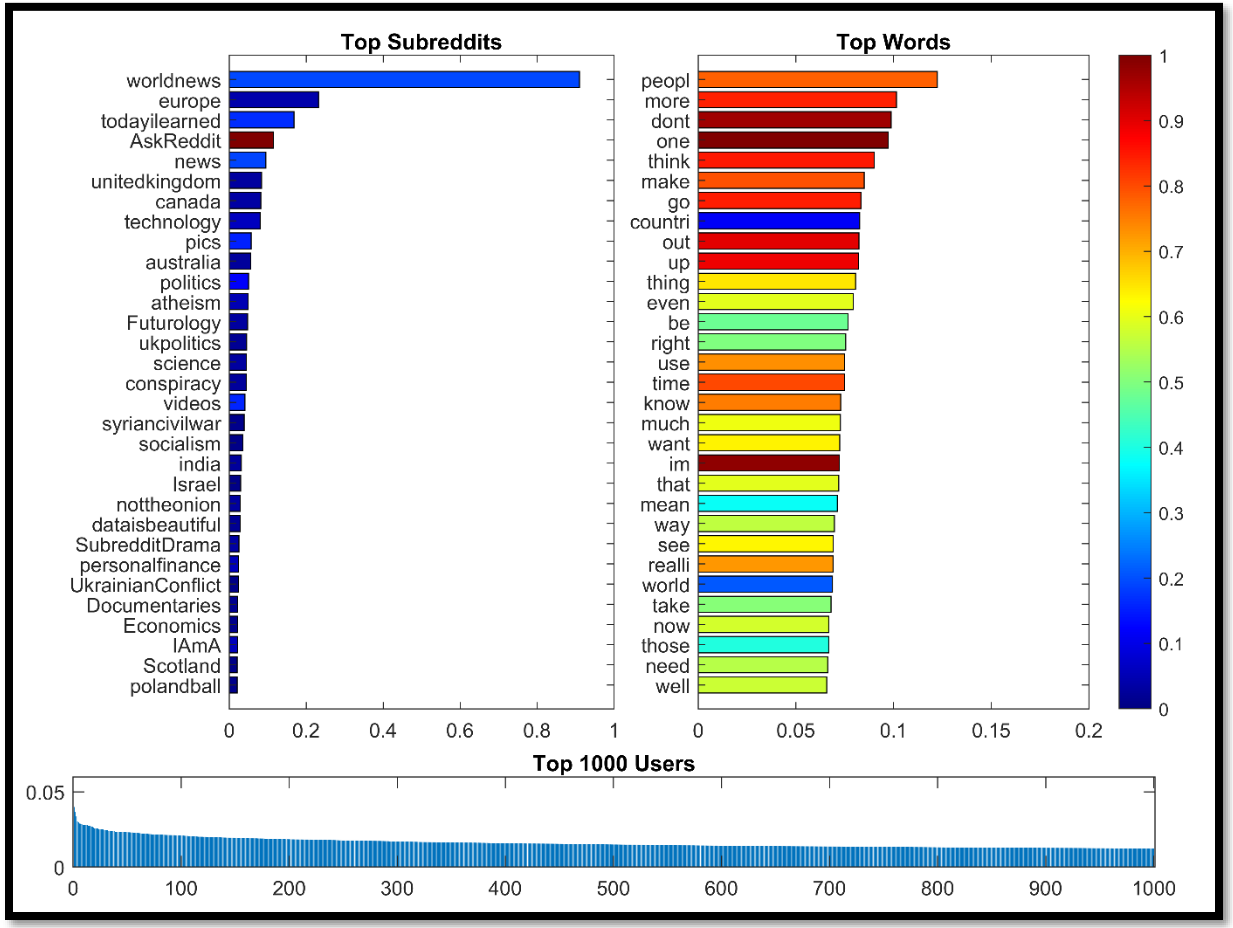


Largest of 8M entries, in absolute value

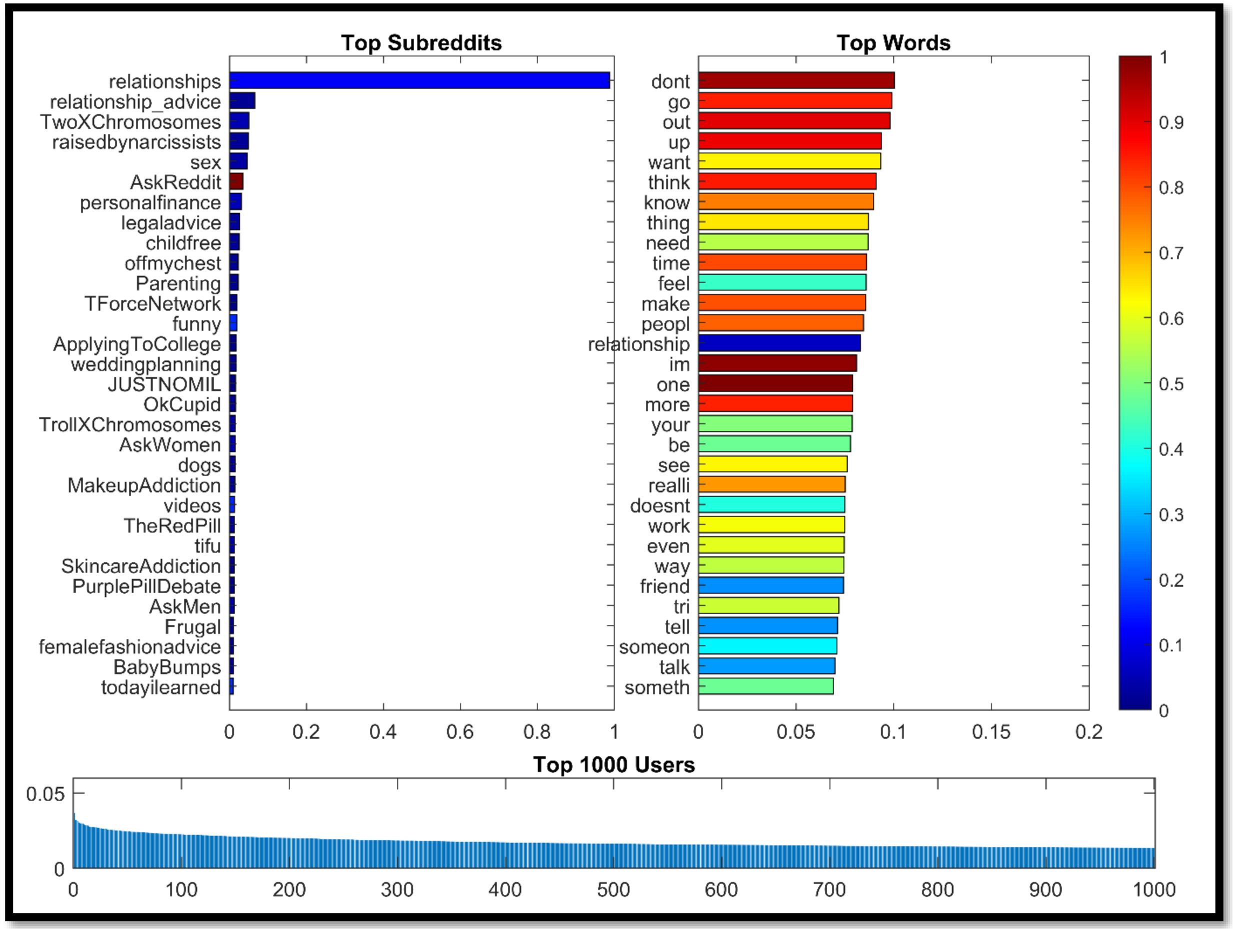


Example Reddit Components Include Rare Words Apropos to High-Scoring Reddits

Component #6: International News

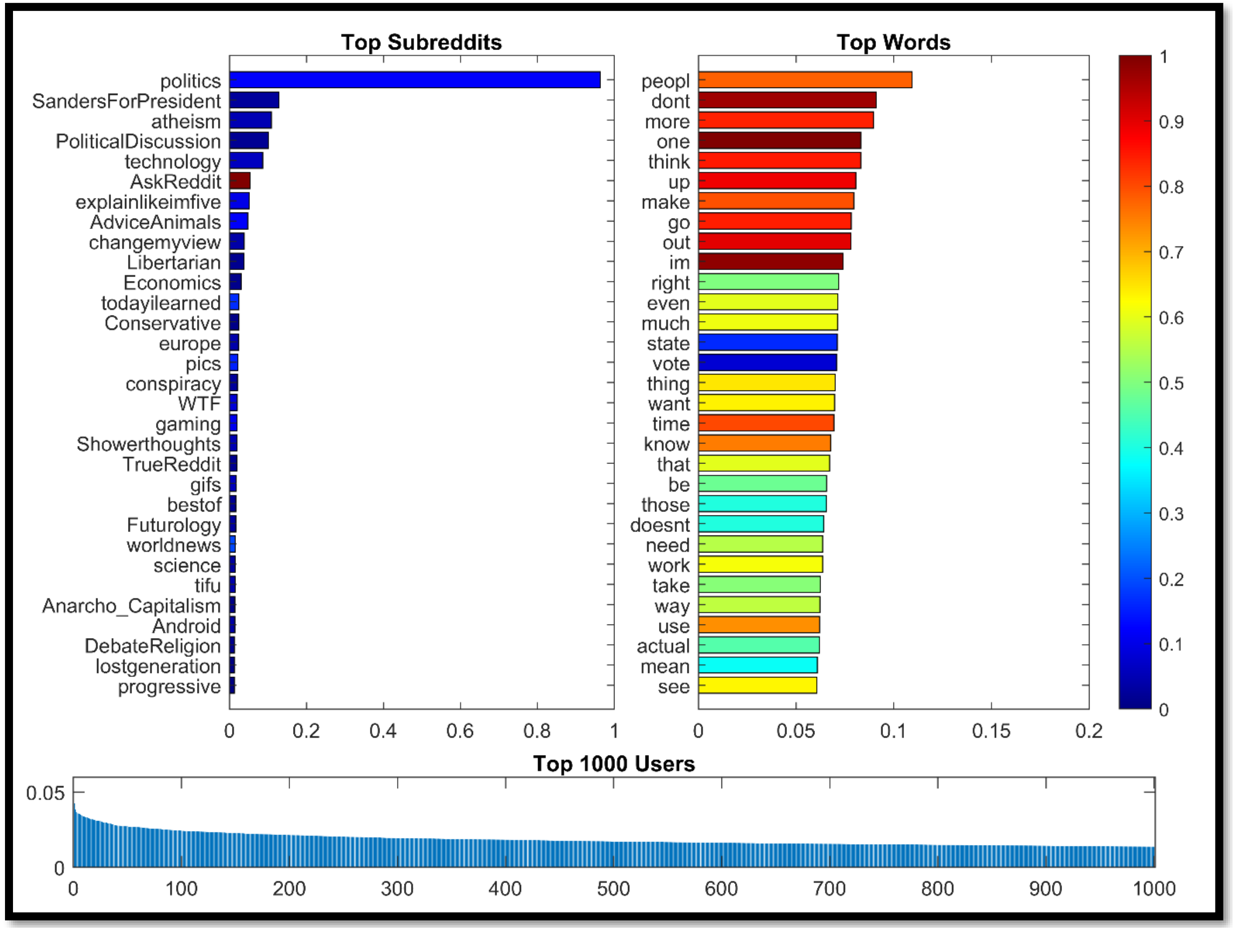


Component #8: Relationships

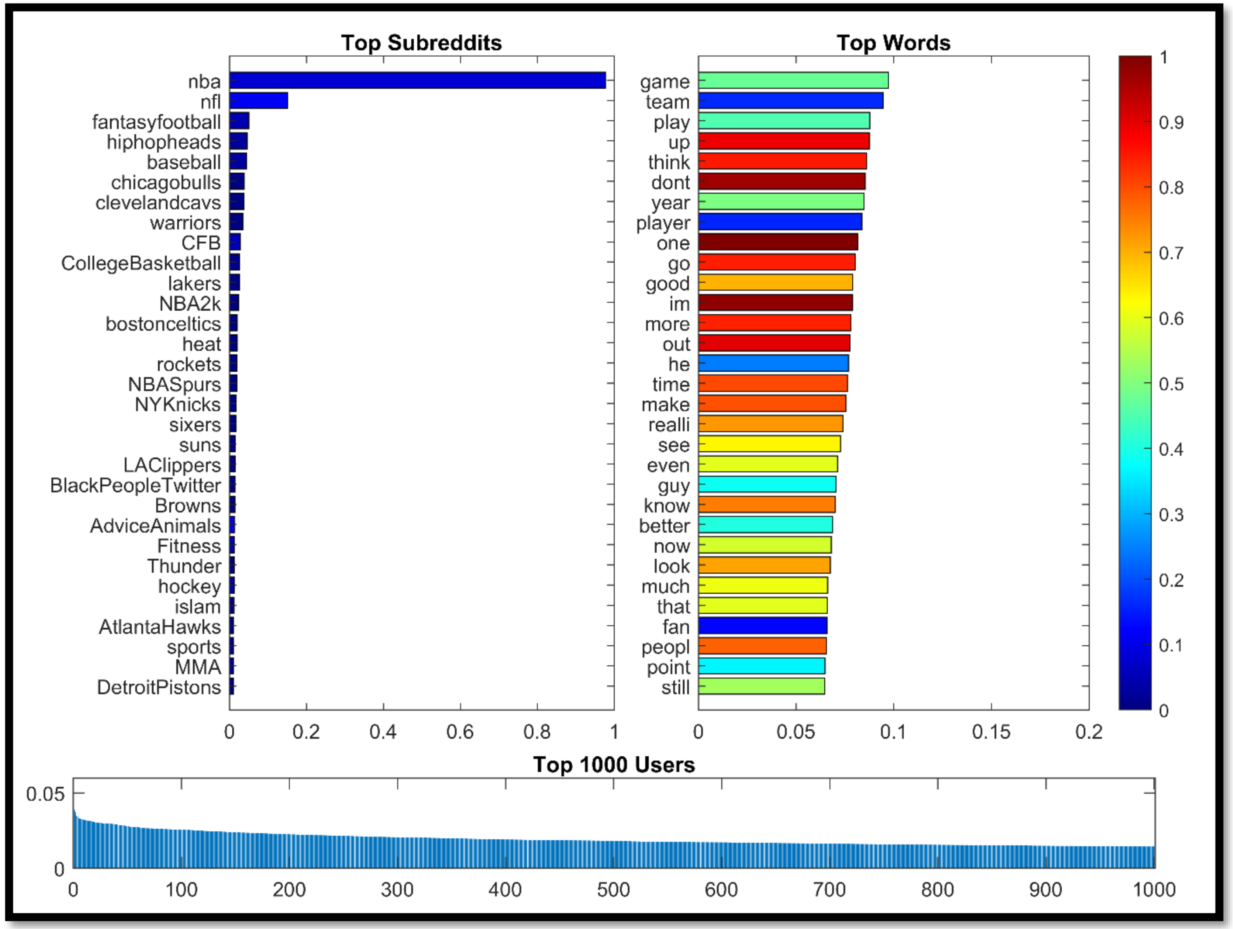


Example Reddit Components Include Rare Words Apropos to High-Scoring Reddits

Component #9: U.S. Politics (2015)

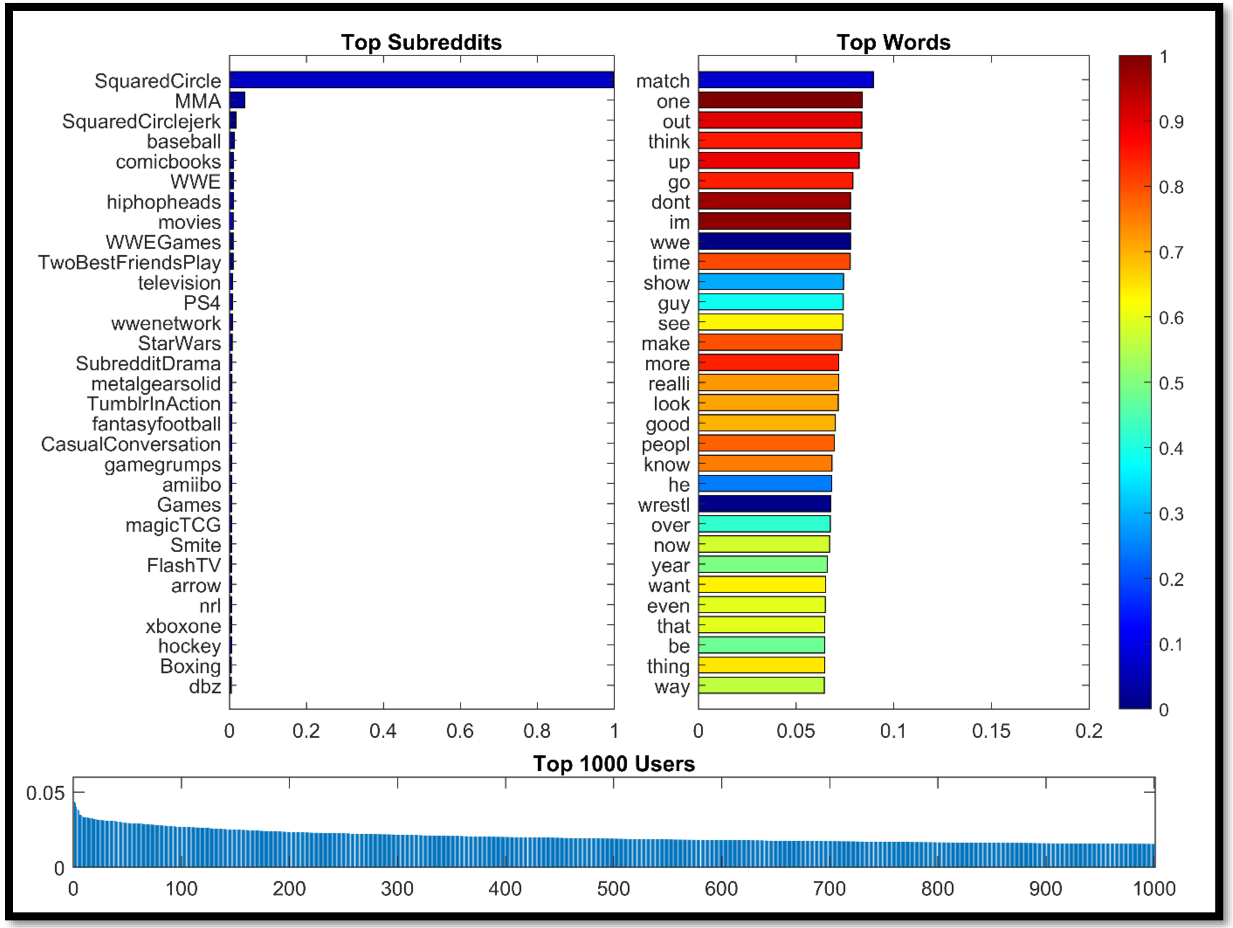


Component #11: Sports

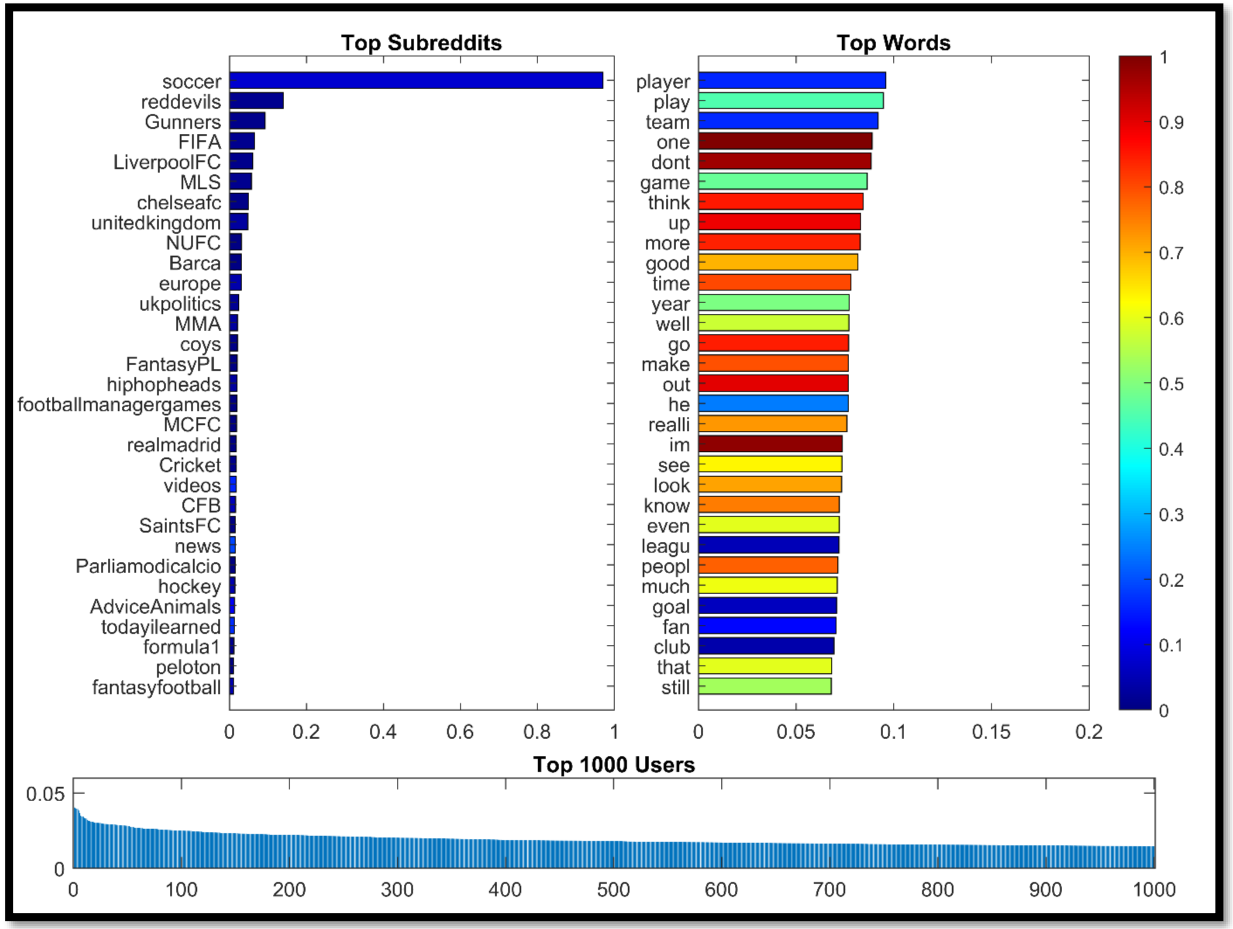


Example Reddit Components Include Rare Words Apropos to High-Scoring Reddits

Component #15: Wrestling

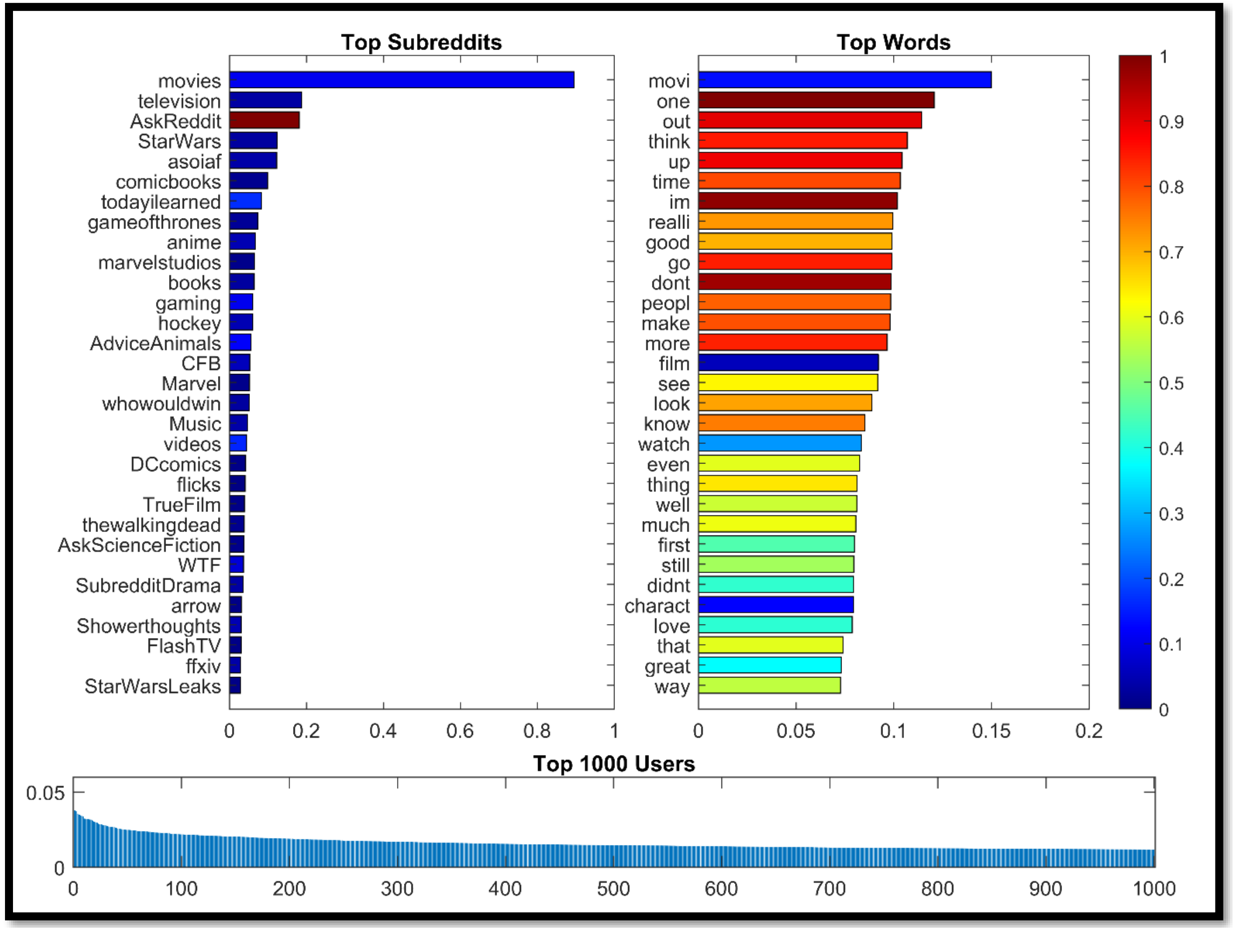


Component #18: Soccer

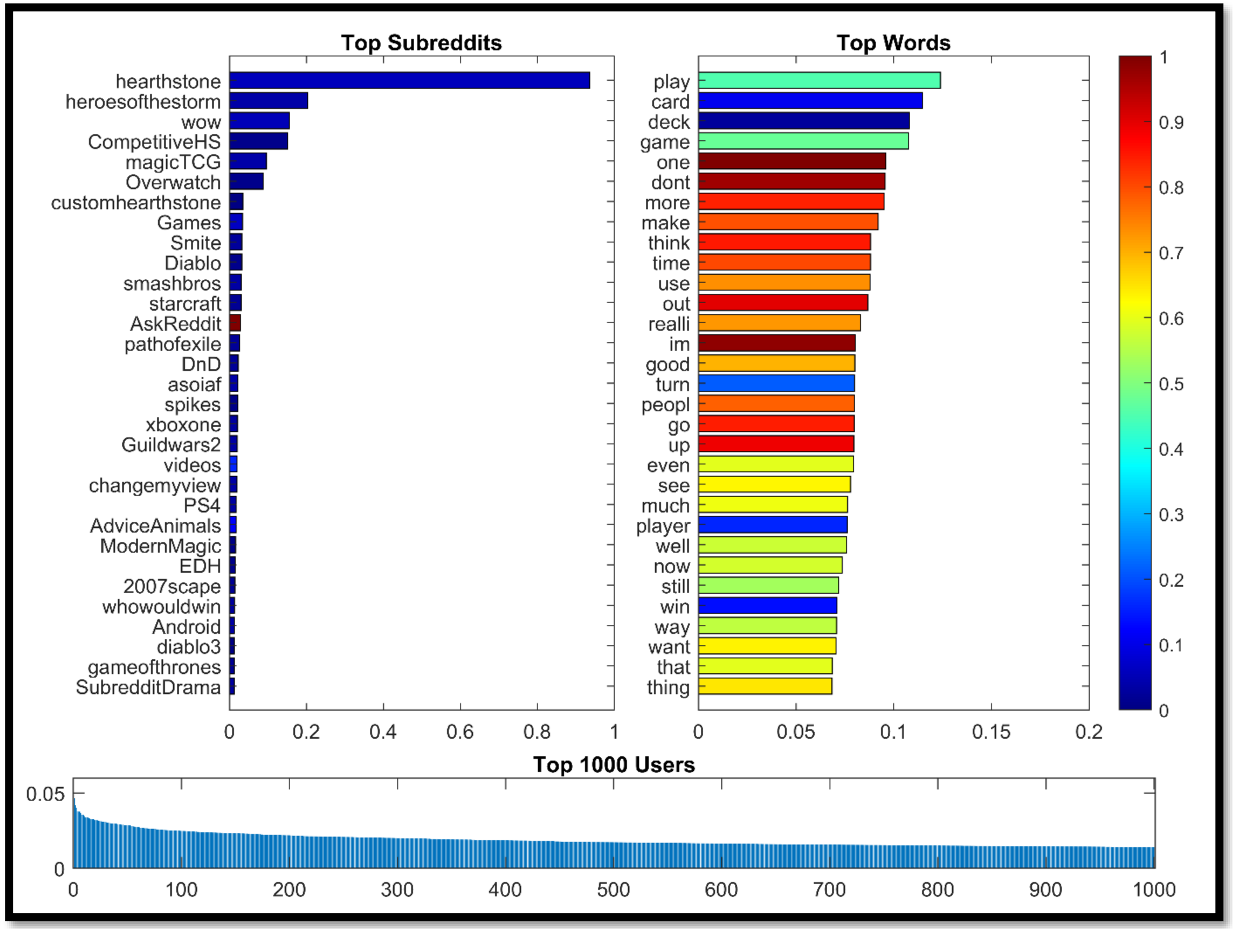


Example Reddit Components Include Rare Words Apropos to High-Scoring Reddits

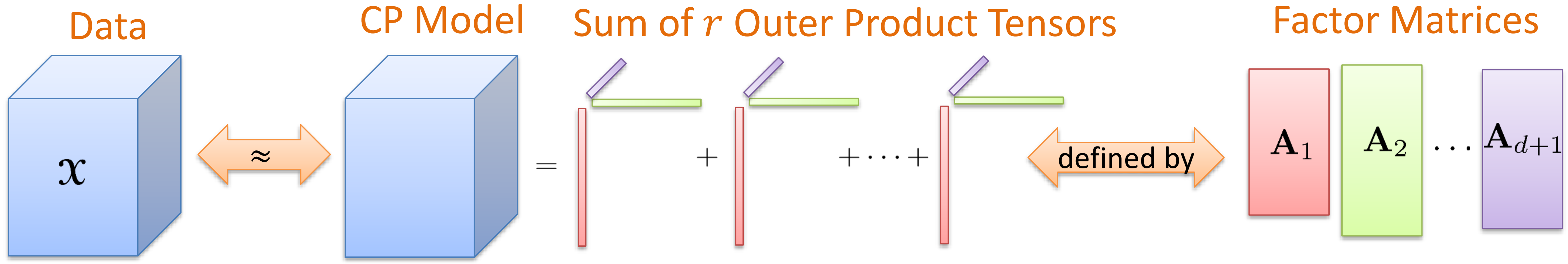
Component #19: Movies & TV



Component #18: Computer Card Game



Interpretation as Sum of Outer Products



$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

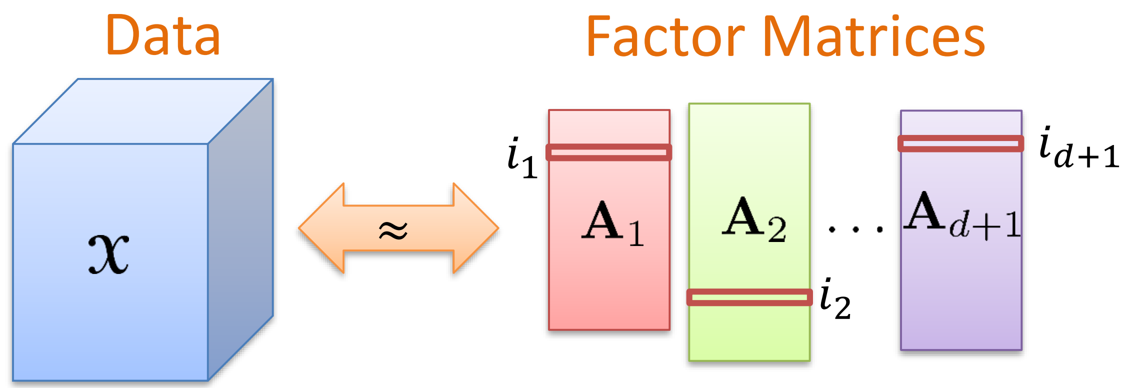
$[\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{d+1}]\!] \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

$\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$

Model Rank

$$x(i_1, i_2, \ldots, i_{d+1}) \approx \sum_{j=1}^{r} \prod_{k=1}^{d+1} a_k(i_k, j)$$

Interpretation as Row-wise Products



$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$ $\quad$ $\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$

Number of tensor elements = $\prod_{k=1}^{d+1} n_k$

$$x(i_1, i_2, \ldots, i_{d+1}) \approx \sum_{j=1}^{r} \prod_{k=1}^{d+1} a_k(i_k, j)$$

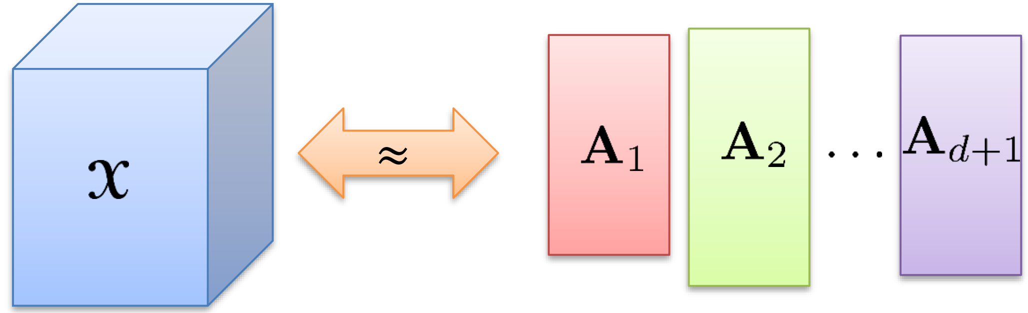
To estimate element $(i_1, i_2, \ldots, i_{d+1})$ of data tensor

- Extract row i_1 of $\mathbf{A}_1$
- Extract row i_2 of $\mathbf{A}_2$
- $\vdots$ $\qquad$ $\vdots$
- Extract row i_{d+1} of $\mathbf{A}_{d+1}$
- Multiply extracted rows elementwise
- Sum result

Doing this for every possible combination of rows yields estimates for every element of the data tensor

Alternating Optimization

Data | Factor Matrices



$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

$\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$

$$\min \sum_{i_1=1}^{n_1} \cdots \sum_{i_{d+1}=1}^{n_{d+1}} \left(x(i_1, \ldots, i_{d+1}) - \sum_{j=1}^{r} \prod_{k=1}^{d+1} a_k(i_k, j) \right)^2$$

- One approach: Alternating Optimization
- Fix *all but one* factor matrix and solve for the remaining one
  - Solve for $\mathbf{A}_1$, fixing $\mathbf{A}_2$ through $\mathbf{A}_{d+1}$
  - Solve for $\mathbf{A}_2$, fixing $\mathbf{A}_1$ and $\mathbf{A}_3$ through $\mathbf{A}_{d+1}$
  - $\vdots$
  - Solve for $\mathbf{A}_{d+1}$, fixing $\mathbf{A}_1$ through $\mathbf{A}_d$
  - Repeat until convergence

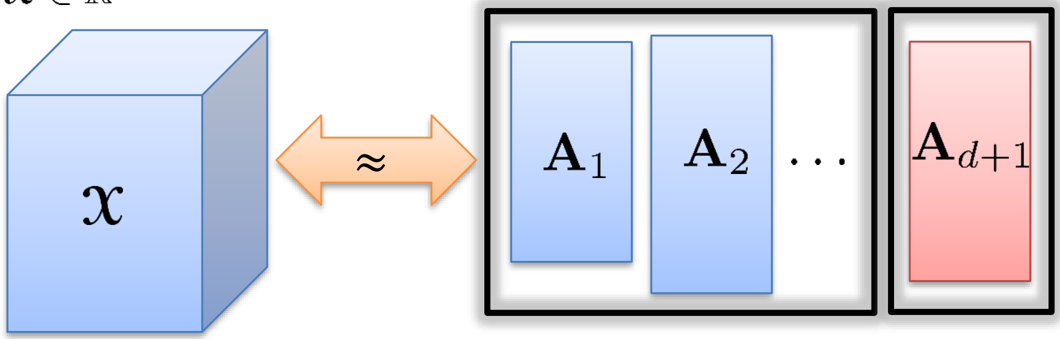
Alternating Optimization Subproblem is Matrix Linear Least Squares Problem

Data

$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_{d+1}}$

Factor Matrices

$\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$



Fix for $(d+1)^{\text{st}}$ subproblem

Unknown

re-index a.k.a. unfold

$$x(i_1, \dots, i_{d+1}) \equiv x(i, \ell)$$

combine and re-index

$$z(i,j) \equiv \prod_{k=1}^{d} a_k(i_k, j)$$

rename

$$b(\ell, j) \equiv a_{d+1}(\ell, j)$$

Rename $n = n_{d+1}$

$$\min_{\mathbf{A}_{d+1}} \sum_{i_1=1}^{n_1} \cdots \sum_{i_{d+1}=1}^{n_{d+1}} \left(x(i_1, \dots, i_{d+1}) - \sum_{j=1}^{r} \prod_{k=1}^{d+1} a_k(i_k, j) \right)^2$$

Re-index to range 1 to $N = \prod_{k=1}^{d} n_k$

$$i \equiv 1 + \sum_{k=1}^{d} (i_k - 1) \prod_{\ell=1}^{k-1} n_\ell$$

$$\min_{\mathbf{B}} \sum_{i=1}^{N} \sum_{\ell=1}^{n} \left(x(i, \ell) - \sum_{j=1}^{r} z(i,j) b(\ell, j) \right)^2$$

Prototypical CP Least Squares Subproblem is "Tall and Skinny"

$$\min_{\mathbf{B}} \sum_{i=1}^{N} \sum_{\ell=1}^{n} \left(x(i,\ell) - \sum_{j=1}^{r} z(i,j) b(\ell,j) \right)^2$$

$$N \gg r, n$$

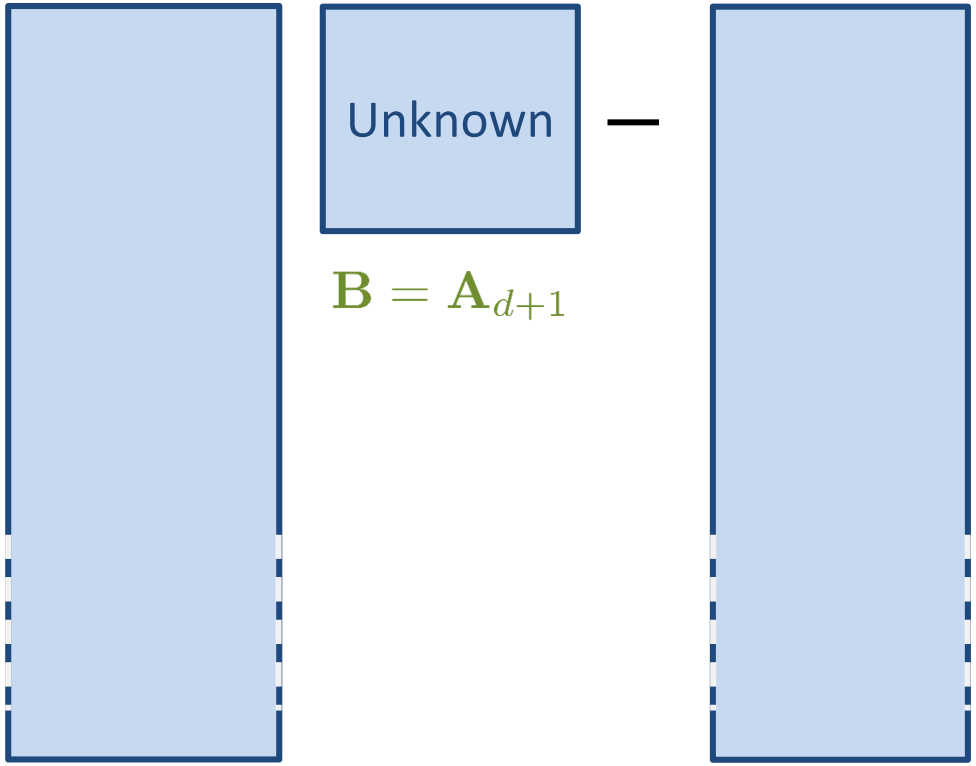
Linking to mode-(d+1) least squares subproblem on prior slide

$$n = n_{d+1}$$

$$N = \prod_{k=1}^{d} n_k$$

$$\min_{\mathbf{B}} \|\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{X}^{\intercal}\|^2$$

$\mathbf{Z} \in \mathbb{R}^{N \times r}$ $\quad \mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\quad \mathbf{X}^{\intercal} \in \mathbb{R}^{N \times n}$



$\mathbf{B} = \mathbf{A}_{d+1}$

Khatri-Rao Product

$\mathbf{Z} = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_1$

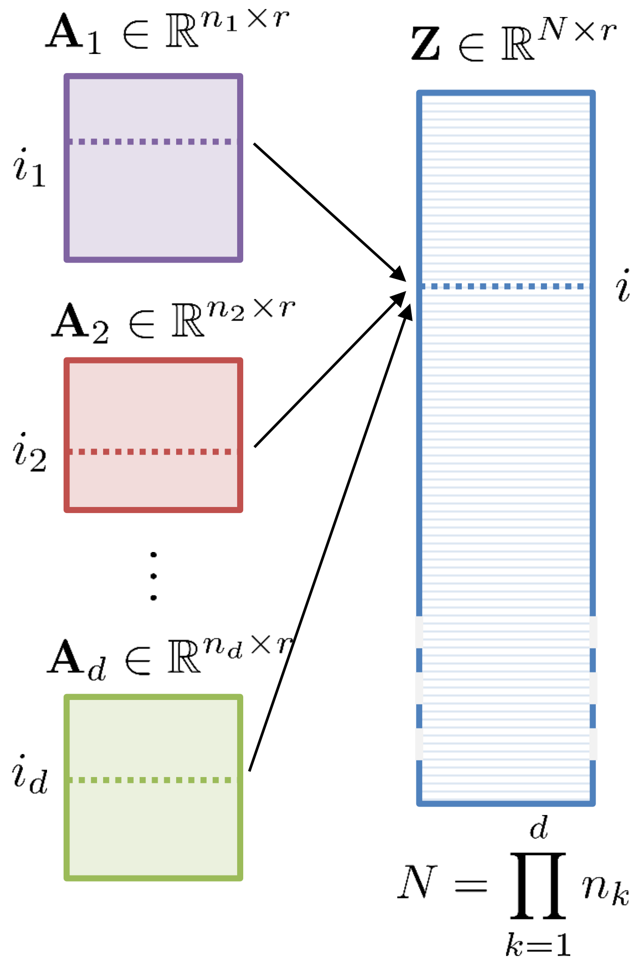
Mode-(d+1) unfolding

$\mathbf{X} = \mathbf{X}_{(d+1)}$

Structure of Khatri-Rao Product (KRP): Hadamard Combinations of Rows of Inputs

KRP of d Matrices: $\mathbf{Z} = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_1$

Number of columns is the same in all input matrices, but number of rows varies



Each row of KRP is Hadamard product of specific rows in Factor Matrices:

$$\mathbf{Z}(i,:) = \mathbf{A}_1(i_1,:) * \cdots * \mathbf{A}_d(i_d,:)$$

where

$$i \equiv 1 + \sum_{k=1}^{d} (i_k - 1) \prod_{\ell=1}^{k-1} n_\ell$$

1-1 Correspondence between *linear index and multi index:*

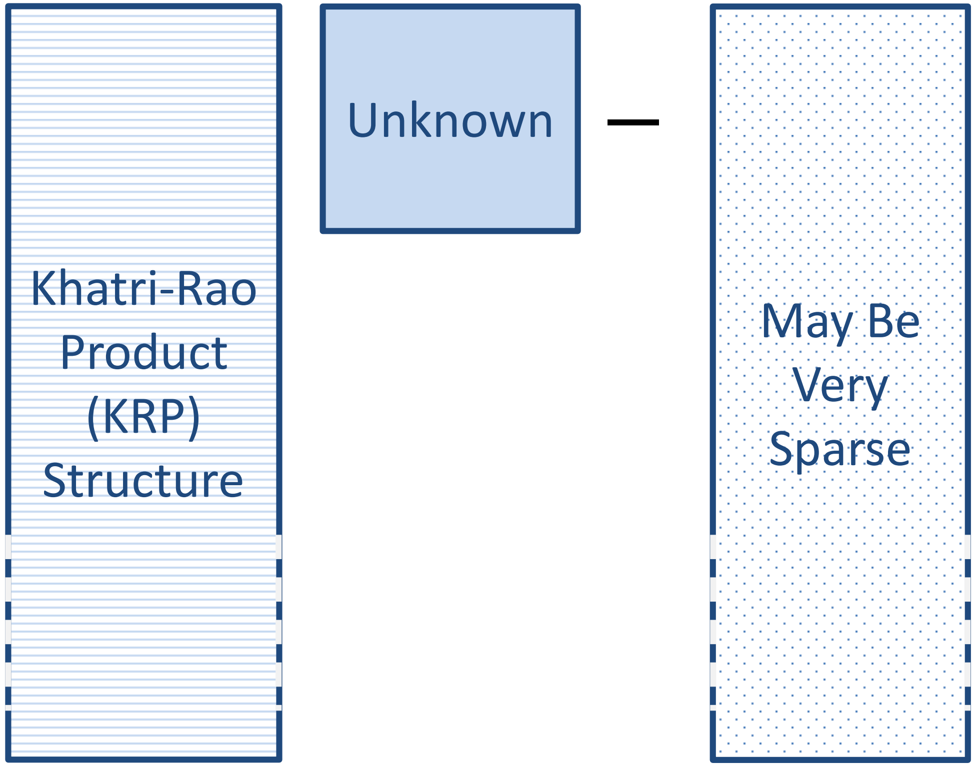
$$i \in [N] \Leftrightarrow (i_1, \ldots, i_d) \in [n_1] \otimes \cdots \otimes [n_d]$$

Prototypical CP Least Squares Problem has Khatri-Rao Product (KRP) Structure

$$\min_{\mathbf{B}} \|\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{X}^{\intercal}\|^2$$

$N \gg r, n$

$\mathbf{Z} \in \mathbb{R}^{N \times r}$ $\mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\mathbf{X}^{\intercal} \in \mathbb{R}^{N \times n}$



- KRP costs $O(Nr)$ to form
- System costs $O(Nnr^2)$ to solve
- KRP structure
  - Cost reduced to $O(Nnr)$
- KRP structure + data sparse
  - Cost reduced to $O\big(r\,\mathrm{nnz}(\mathbf{X})\big)$

Question for today: Suppose this system is extremely large? How can we solve efficiently?

Ingredient #1: Sample Subset of Rows in Overdetermined Least Squares System

Sandia National Laboratories

$$\min_{\mathbf{B}} \|\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{X}^{\intercal}\|^2$$

$\mathbf{Z} \in \mathbb{R}^{N \times r}$ $\quad$ $\mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\quad - \quad$ $\mathbf{X}^{\intercal} \in \mathbb{R}^{N \times n}$

Khatri-Rao Product (KRP) Structure

Unknown

May Be Very Sparse

$$N \gg r, n$$

$$\min_{\mathbf{B}} \|\mathbf{\Omega}\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{\Omega}\mathbf{X}^{\intercal}\|^2$$

$\mathbf{\Omega}\mathbf{Z} \in \mathbb{R}^{s \times r}$ $\quad$ $\mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\quad - \quad$ $\mathbf{\Omega}\mathbf{X}^{\intercal} \in \mathbb{R}^{s \times n}$

Sampled KRP

Unknown

Sampled Data

Complexity reduced from $O(Nnr)$ to $O(snr^2)$

Key surveys:
M. W. Mahoney, ***Randomized Algorithms for Matrices and Data***, 2011;
D. P. Woodruff, ***Sketching as a Tool for Numerical Linear Algebra***, 2014

How to sample so that solution of sampled problem yields something close to the optimal residual of the original problem?

Ingredient #2: Weight Sampled Rows by Probability of Selection to Eliminate Bias

Probability distribution on rows of linear system

$$\sum_{i=1}^{N} p_i = 1$$

Not specifying yet how these probabilities are selected

Pick a single random index ξ with probability p_ξ

Choose

$$\mathbf{\Omega} = \begin{bmatrix} 0 & \cdots & 0 & \frac{1}{\sqrt{p_\xi}} & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{1 \times N}$$

ξth entry

Then (assuming all p_i positive) the sampled the sampled residual equals true residual in expectation:

$$\mathbb{E}\|\mathbf{\Omega}\mathbf{Z}\mathbf{B}^\intercal - \mathbf{\Omega}\mathbf{X}^\intercal\|^2 = \sum_{i=1}^{N} p_i \left(\left\| \frac{1}{\sqrt{p_i}} \mathbf{Z}(i,:)\mathbf{B}^\intercal - \frac{1}{\sqrt{p_i}} \mathbf{X}^\intercal(i,:) \right\|^2 \right)$$
$$= \|\mathbf{Z}\mathbf{B}^\intercal - \mathbf{X}^\intercal\|^2$$

Pick a s random indices ξ_j (with replacement) such that $P(\xi_j = i) = p_i$.

Choose $\mathbf{\Omega} \in \mathbb{R}^{s \times N}$ such that

Not specifying yet how s is determined

$$\omega(j,i) = \begin{cases} \frac{1}{\sqrt{s p_i}} & \text{if } \xi_j = i \\ 0 & \text{otherwise} \end{cases}$$

Each row has a single nonzero!

Then, as before, we have:

$$\mathbb{E}\|\mathbf{\Omega}\mathbf{Z}\mathbf{B}^\intercal - \mathbf{\Omega}\mathbf{X}^\intercal\|^2 = \|\mathbf{Z}\mathbf{B}^\intercal - \mathbf{X}^\intercal\|^2$$

Survey: D. P. Woodruff, ***Sketching as a Tool for Numerical Linear Algebra***, 2014

Optimal Choice for Sampling Probability is Based on Leverage Scores

$$\begin{bmatrix} \boxed{z_{11} & z_{12} & \cdots & z_{1r}} \\ 0 & z_{22} & \cdots & z_{2r} \\ 0 & z_{32} & \cdots & z_{3r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & z_{N2} & \cdots & z_{Nr} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \vdots \\ \nu_N \end{bmatrix}$$

$\ell_1(\mathbf{Z}) = 1$

$\mathbf{Z} \in \mathbb{R}^{N \times r}$

Leverage score:

Let $\mathbf{Q}$ be any orthonormal basis of the column space of $\mathbf{Z}$.

Leverage score of row i:

$$\ell_i(\mathbf{Z}) = \|\mathbf{Q}(i,:)\|_2^2 \in [0,1]$$

Coherence:

$$\mu(\mathbf{Z}) = \max_{i \in [N]} \ell_i(\mathbf{Z})$$

$$r/N \leq \mu(\mathbf{Z}) \leq 1$$

Rough Intuition:

Key rows have high leverage score

$$s = O(\epsilon^{-2} \ln(r)\, r \beta^{-1})$$

where $\beta = \min_{i \in [N]} \frac{r\, p_i}{\ell_i(\mathbf{Z})}$

What if we do uniform sampling? $p_i = \frac{1}{N}$ for all $i \in [N]$,

Case 1: $\mu(\mathbf{Z}) = r/N$ (incoherent)

$\Rightarrow \beta = 1 \Rightarrow s = O(\epsilon^{-2} \ln(r)\, r)$

Case 2: $\mu(\mathbf{Z}) = 1$ (coherent)

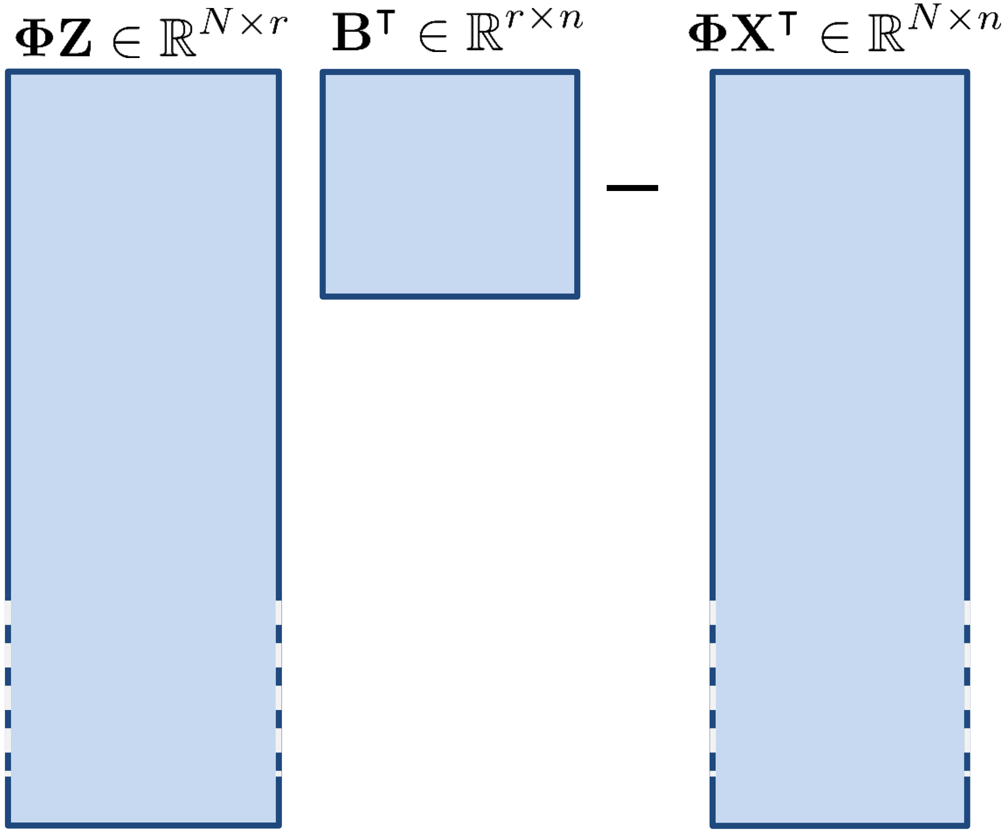
$\Rightarrow \beta = r/N \Rightarrow s = O(\epsilon^{-2} \ln(r)\, N)$

In Case 2, prefer $p_i = \ell_i(\mathbf{Z})/r$, <u>but</u> costs $O(Nr^2)$ to compute leverage scores!

Survey: D. P. Woodruff, ***Sketching as a Tool for Numerical Linear Algebra***, 2014

Aside: Uniform Sampling Okay for "Mixed" Dense Tensors (Inapplicable to Sparse)

Transform System: $\min_{\mathbf{B}\in\mathbb{R}^{r\times n}} \|\Phi\mathbf{Z}\mathbf{B}^\intercal - \Phi\mathbf{X}\|_F^2$



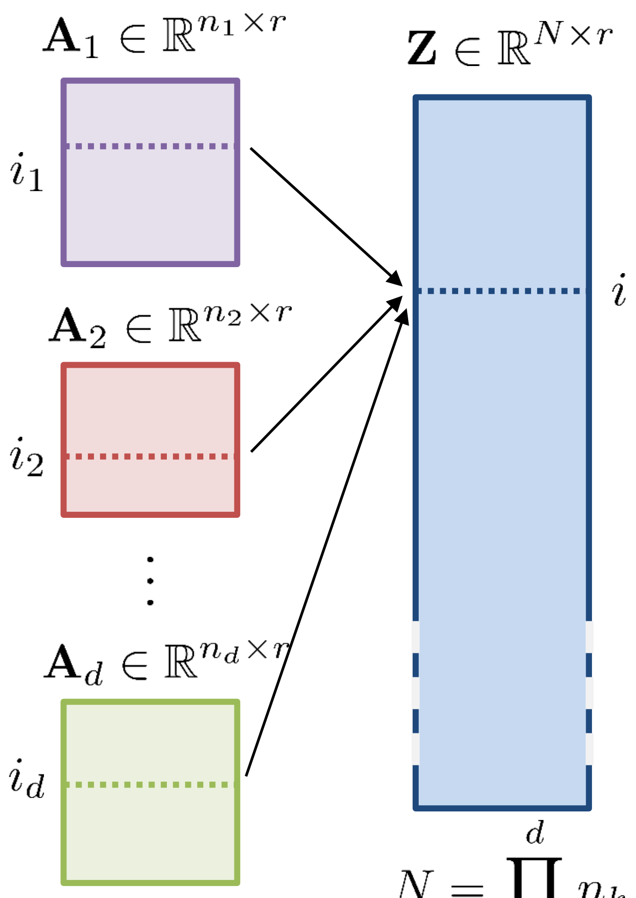
$N \gg r, n$

- Choose $\mathbf{\Phi}$ so that all leverage scores of $\mathbf{\Phi Z}$ approximately equal, then uniform sampling yields $\beta \approx 1$
  - "Uniformize" the leverage scores per Mahoney
  - Fast Johnson-Lindenstrauss Transform (FJLT) uses random rows of matrix transformed by FFT and Rademacher diagonal
  - FJLT cost per iteration: $O(rN \log N)$
- Gaining Efficiency for KRP matrices
  - Transform individual factor matrices *before* forming $\mathbf{Z}$
  - Sample rows of $\mathbf{Z}$ implicitly
  - Kronecker Fast Johnson-Lindenstrauss Transform (KFJLT)
  - Special handling of right-hand side with preprocessing costs
  - KFJLT cost per iteration: $O(r \sum_k n_k \log n_k + sr^2)$
- References

  - C. Battaglino, G. Ballard, T. G. Kolda. **A Practical Randomized CP Tensor Decomposition**. *SIAM Journal on Matrix Analysis and Applications*, Vol. 39, No. 2, pp. 876-901, 26 pages, 2018. https://doi.org/10.1137/17M1112303
  - R. Jin, T. G. Kolda, R. Ward. **Faster Johnson-Lindenstrauss Transforms via Kronecker Products,** 2019. http://arxiv.org/abs/1909.04801

Ingredient #3: Bound Leverage Scores

KRP: $\mathbf{Z} = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_1$



$$N = \prod_{k=1}^{d} n_k$$

Upper Bound on Leverage Score

Lemma (Cheng et al., NIPS 2016; Battaglino et al., SIMAX 2018):

$$\ell_i(\mathbf{Z}) \leq \prod_{k=1}^{d} \ell_{i_k}(\mathbf{A}_k)$$

Too expensive to calculate $O(Nr^2)$

Cheap to calculate individual leverage scores $O(r^2 \sum_k n_k)$

1-1 Correspondence between *linear index and multi index:*

$$i \in [N] \Leftrightarrow (i_1, \ldots, i_d) \in [n_1] \otimes \cdots \otimes [n_d]$$

Ingredient #4: Use Factor Matrix Leverage Scores for Sampling Probabilities (Main Thm)

Given linear system: $\|\mathbf{Z}\mathbf{B}^\intercal - \mathbf{X}^\intercal\|^2$ with $\mathbf{Z} = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_1 \in \mathbb{R}^{N \times r}, \mathbf{X}^\intercal \in \mathbb{R}^{n \times N}$

Define sampling probabilities: $p_i = \frac{1}{r^d} \prod_{k=1}^{d} \ell_{i_k}(\mathbf{A}_k)$ for all $i \in [N]$

Leverage Scores

$\ell_{i_k}(\mathbf{A}_k) = \|\mathbf{Q}_k(i_k, :)\|_2$ where $\mathbf{Q}_k$ is orthonormal basis for column space of $\mathbf{A}_k$

And random sampling matrix: Pick a s random indices ξ_j such that $P(\xi_j = i) = p_i$ and define

$$\mathbf{\Omega} \in \mathbb{R}^{s \times N} \quad \text{with} \quad \omega(j,i) = \begin{cases} \frac{1}{\sqrt{s p_i}} & \text{if } \xi_j = i \\ 0 & \text{otherwise} \end{cases}$$

Solve sampled problem: $\tilde{\mathbf{B}}_* \equiv \arg \min_{\mathbf{B} \in \mathbb{R}^{r \times n}} \|\mathbf{\Omega}\mathbf{Z}\mathbf{B}^\intercal - \mathbf{\Omega}\mathbf{X}\|_F^2$

Get probabilistic error bound: With probability $1 - \delta$ for $\delta \in (0,1)$, we have

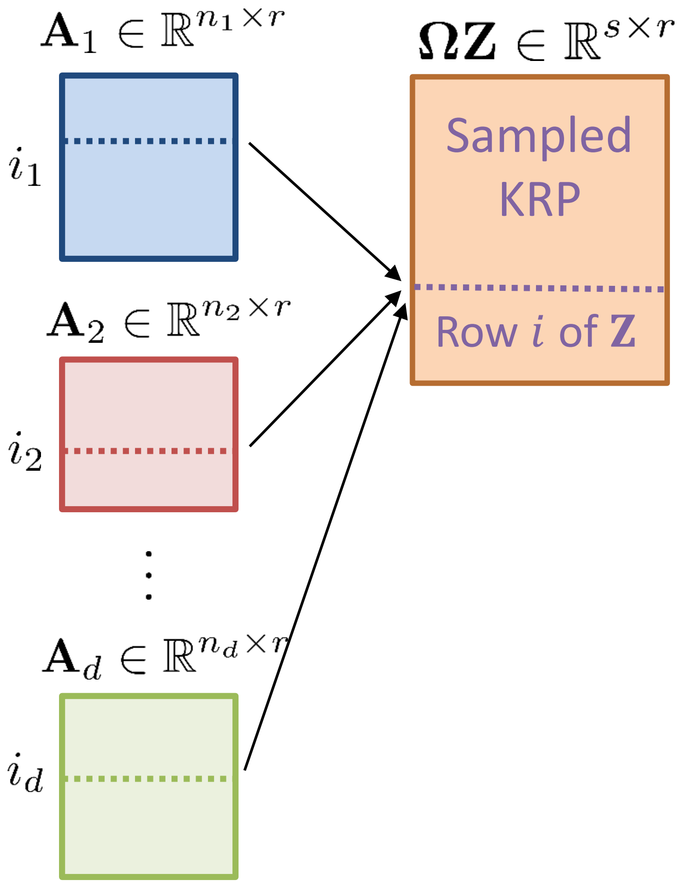
$$\|\mathbf{Z}\tilde{\mathbf{B}}_*^\intercal - \mathbf{X}^\intercal\|_F^2 \leq (1 + O(\epsilon))\|\mathbf{Z}\mathbf{B}_*^\intercal - \mathbf{X}^\intercal\|_F^2$$

when number of samples satisfies: $s = O(r^d \log(n/\delta)/\epsilon^2)$

1-1 Correspondence between *linear index and multi index:*
$i \in [N] \Leftrightarrow (i_1, \ldots, i_d) \in [n_1] \otimes \cdots \otimes [n_d]$

Ingredient #5: Efficient Sampling *without* Forming KRP

KRP: $\mathbf{Z} = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_1$



Upper Bound on Leverage Score

Lemma (Cheng et al., NIPS 2016; Battaglino et al., SIMAX 2018):

$$\ell_i(\mathbf{Z}) \leq \prod_{k=1}^{d} \ell_{i_k}(\mathbf{A}_k)$$

Too expensive to calculate $O(Nr^2)$

Cheap to calculate individual leverage scores $O(r^2 \sum_k n_k)$

Recall probability of sampling row i

$$p_i \equiv \frac{1}{r^d} \prod_{k=1}^{d} \ell_{i_k}(\mathbf{A}_k)$$

But still don't want to consider all N possible combinations corresponding to all rows of $\mathbf{Z}$!

1-1 Correspondence between *linear index* and *multi index:*

$$i \in [N] \Leftrightarrow (i_1, \ldots, i_d) \in [n_1] \otimes \cdots \otimes [n_d]$$

See Reference for Details of Other Specializations

$$\min_{\mathbf{B}} \|\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{X}^{\intercal}\|^2$$

$\mathbf{Z} \in \mathbb{R}^{N \times r}$ $\mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\mathbf{X}^{\intercal} \in \mathbb{R}^{N \times n}$

Khatri-Rao Product (KRP) Structure

Unknown

$-$

May Be Very Sparse

$$\min_{\mathbf{B}} \|\mathbf{\Omega}\mathbf{Z}\mathbf{B}^{\intercal} - \mathbf{\Omega}\mathbf{X}^{\intercal}\|^2$$

$\mathbf{\Omega}\mathbf{Z} \in \mathbb{R}^{s \times r}$ $\mathbf{B}^{\intercal} \in \mathbb{R}^{r \times n}$ $\mathbf{\Omega}\mathbf{X}^{\intercal} \in \mathbb{R}^{s \times n}$

Sampled KRP

Unknown

$-$

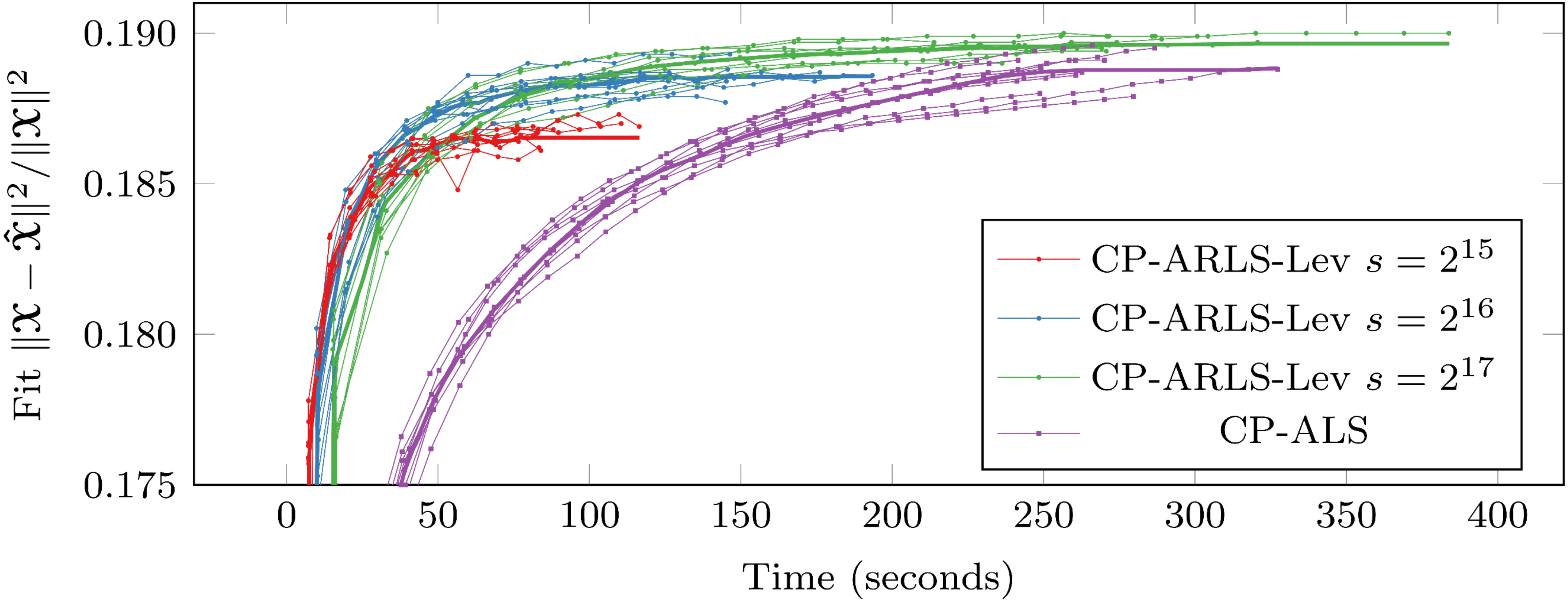
Sampled Data

- Combined repeated samples
- Deterministically include rows with high leverage scores, including specialized methods for finding such rows
- Never form $\mathbf{X}^{\mathrm{T}}$ explicitly, precompute linear indices for every nonzero and every mode
- Using epochs and estimating error using sampling

Numerical Results

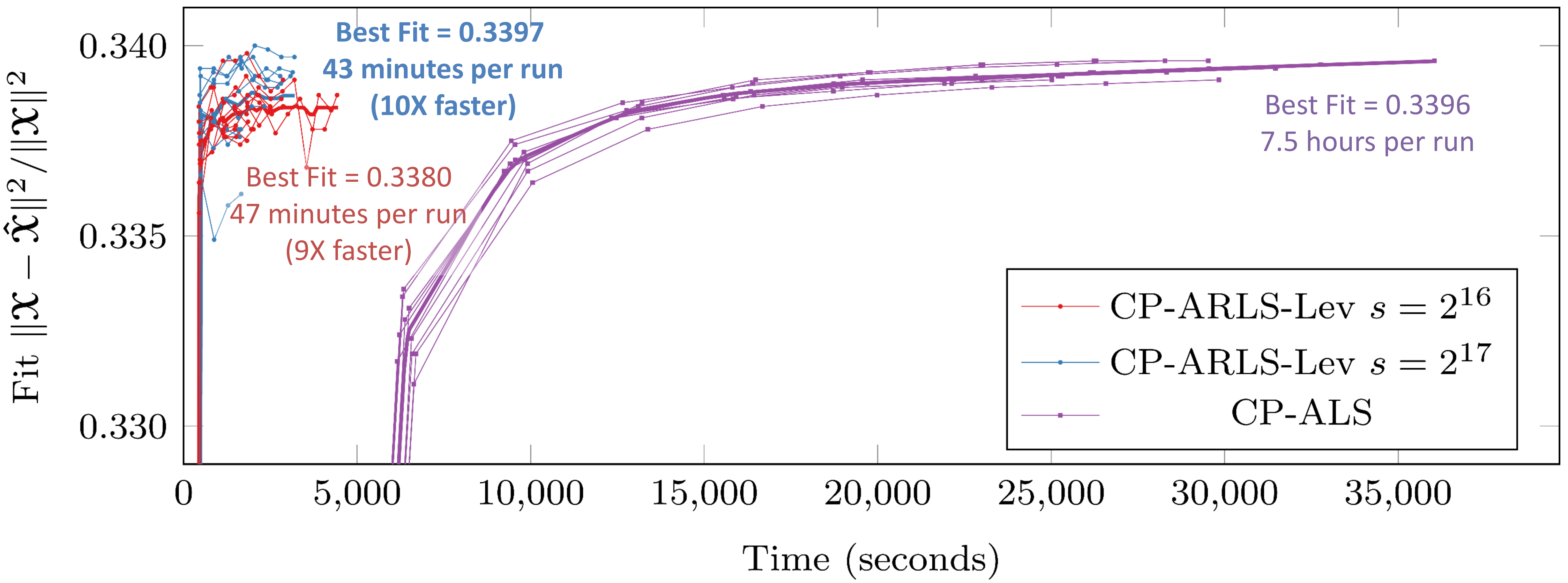
CP-ARLS-Lev Comparable to CP-ALS on Small Uber Problem

Uber Tensor: 183 x 24 x 1140 x 1717 Uber Tensor with 3M nonzeros (0.038% dense).
Rank r = 25 CP decomposition



Over 9X Speed-up for Amazon Tensor with 1.7 Billion Nonzeros

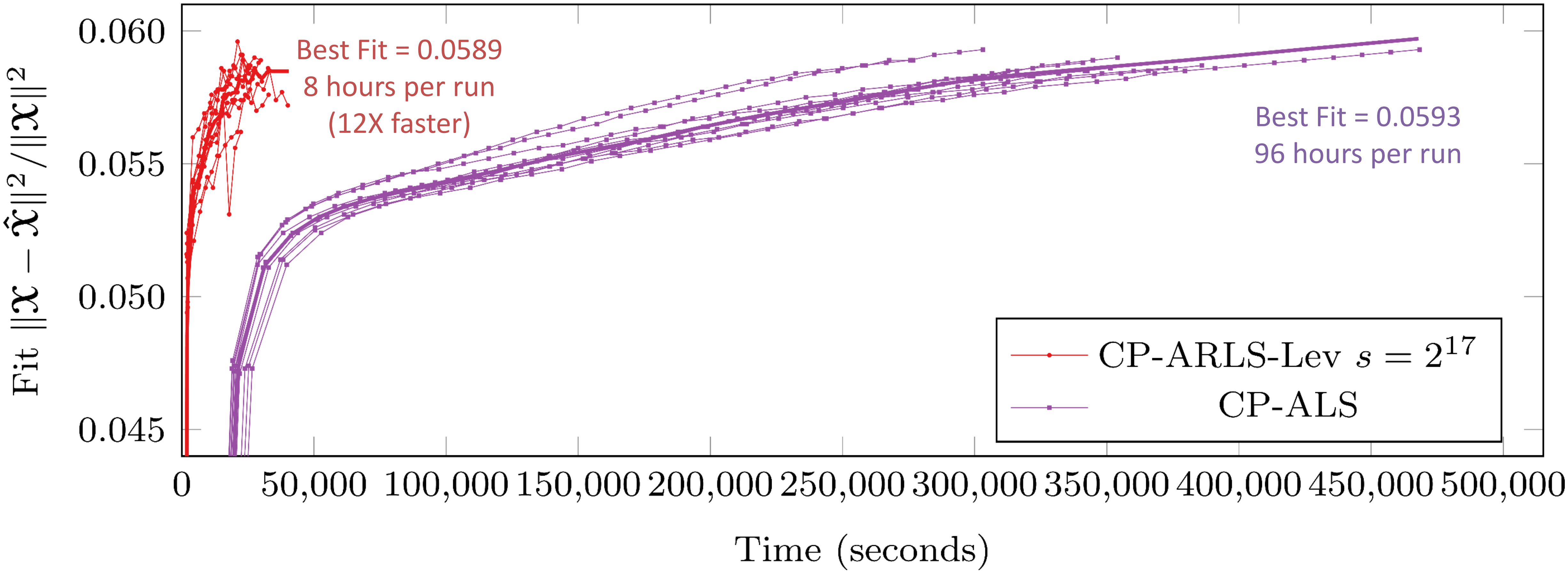
Amazon Tensor: 4.8M x 1.8M x 1.8M Amazon Tensor with 1.7B nonzeros.
Rank r = 25 CP decomposition



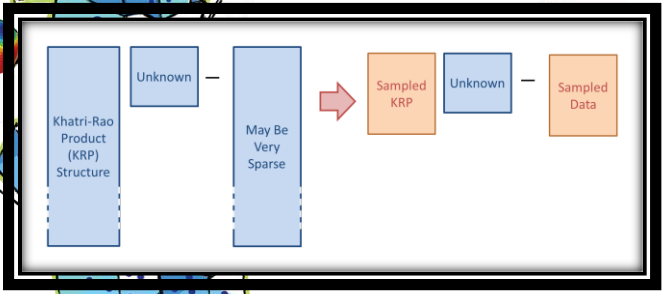
Over 12X Speed-up for Reddit Tensor with 4.7 Billion Nonzeros (106 GB)

Amazon Tensor: 8.2M x 0.2M x 8.1M Reddit Tensor with 4.7B nonzeros.

Rank r = 25 CP decomposition



Conclusions & Future Work



Larsen and Kolda, ***Practical Leverage-Based Sampling for Tensor Decomposition***, arXiv:2006.16438, 2020

- Tensor decomposition: unsupervised machine learning
- Many applications, including social discussion analysis
- Model fit via alternating optimization, resulting in series of least squares subproblems
- Subproblems are “tall and skinny”, amenable to sketching
- Leverage-score sampling ideal for sparse data tensors
- Can estimate leverage scores cheaply using leverage scores of factor matrices
- Results in huge speedups f

Contact Info: Brett bwlarsen@stanford.edu, Tammy tgkolda@sandia.gov