Using high-dimensional probability to study complex data: matrices, tensors, linear systems, and beyond

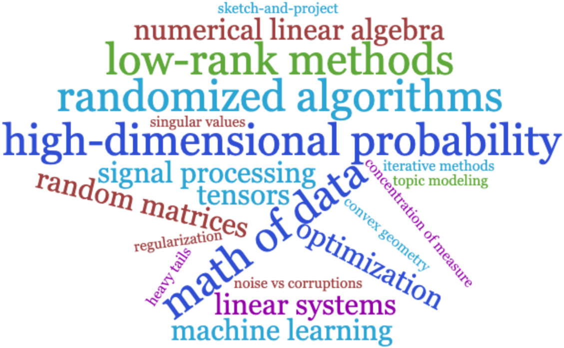
Liza Rebrova

UCLA, Department of Mathematics & LBL, Computational Research Division

February 5, 2021

What do I do?

- Study the structure of large high-dimensional objects in the presence of randomness
- Use this understanding to develop (randomized) methods that work with complex data efficiently

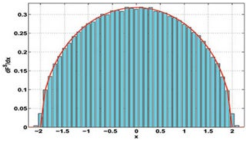


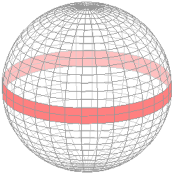
Large high-dimensional objects:

- Sets
- Matrices
- Tensors
- Graphs
- Systems of linear equations
- Neural nets
- ...

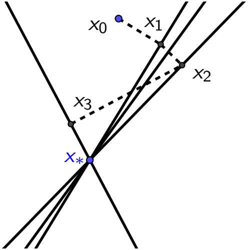
High-dimensional probability

helps revealing their structure:





Concentration of measure phenomenon



Research overview

1. Matrices: Condition numbers of i.i.d. heavy-tailed random matrices (UofMichigan)

Research overview

1. Matrices: Condition numbers of i.i.d. heavy-tailed random matrices (UofMichigan)
2. Linear systems:
   2.1 Scaling kernel ridge regression with HSS linear solvers (Lawrence Berkeley National Lab)
   2.2 Iterative methods for optimization (UCLA)
      - Guarantees for gaussian block sketching for Kaczmarz method
      - Corruption avoiding versions of Randomized Kaczmarz and SGD methods

Research overview

1. Matrices: Condition numbers of i.i.d. heavy-tailed random matrices (UofMichigan)
2. Linear systems:
   - 2.1 Scaling kernel ridge regression with HSS linear solvers (Lawrence Berkeley National Lab)
   - 2.2 Iterative methods for optimization (UCLA)
     - Guarantees for gaussian block sketching for Kaczmarz method
     - Corruption avoiding versions of Randomized Kaczmarz and SGD methods
3. Tensors: (UCLA)
   - Modewise (structure preserving) methods for tensor dimension reduction
   - Matrix and tensor low rank decomposition for interpretable machine learning

Why tensors?

$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \ldots \times n_d}$ – d-mode tensor

Why tensors?

$$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \ldots \times n_d} - d\text{-mode tensor}$$

Naturally multi-modal data is ubiquitous:

- datasets with many attributes
- datasets with temporal component
- color pictures, videos

Why tensors?

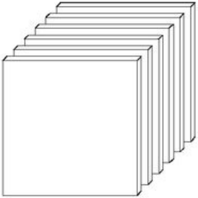
$$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \ldots \times n_d} \text{ -- } d\text{-mode tensor}$$

Naturally multi-modal data is ubiquitous:

- datasets with many attributes
- datasets with temporal component
- color pictures, videos

So,

- Converting it to a vector (vectorization) or to a matrix (matricization) destroys the structure of such data!
- Moreover, tensorized computations are memory- and time-efficient. For example, tensorized random projections



Tensor CP-rank

Dimension reduction often uses low rank property. What is a low rank tensor?

Tensor CP-rank

Dimension reduction often uses low rank property. What is a low rank tensor?

By analogy with matrices, rank 1 tensor $\mathcal{X} = \mathbf{x}_1 \bigcirc \ldots \bigcirc \mathbf{x}_d$ is

$$\mathcal{X}(i_1, \ldots, i_d) = \mathbf{x}_1(i_1)\mathbf{x}_2(i_2) \ldots \mathbf{x}_d(i_d).$$

Tensor CP-rank

Dimension reduction often uses low rank property. What is a low rank tensor?

By analogy with matrices, rank 1 tensor $\mathcal{X} = \mathbf{x}_1 \bigcirc \ldots \bigcirc \mathbf{x}_d$ is

$$\mathcal{X}(i_1, \ldots, i_d) = \mathbf{x}_1(i_1)\mathbf{x}_2(i_2)\ldots\mathbf{x}_d(i_d).$$

CP-rank r tensor is a smallest number of rank-one tensors that generate $\mathcal{X}$ as their sum:

$$\mathcal{X} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_1^i \bigcirc \ldots \bigcirc \mathbf{x}_d^i$$

Normalization: we assume $\|\mathbf{x}_j^i\|_2 = 1$. Clearly, $r \leq n^d$. Note that low-rank tensor has rnd degrees of freedom instead of n^d.

Tensors are harder than matrices

However, tensor rank issue:

- CP CANDECOMP/PARAFAC (canonical decomposition/parallel factors) rank is quite natural, but has important issues:
    - It is NP-hard to compute the rank
    - Uniqueness question

Tensors are harder than matrices

However, tensor rank issue:

- CP CANDECOMP/PARAFAC (canonical decomposition/parallel factors) rank is quite natural, but has important issues:
  - It is NP-hard to compute the rank
  - Uniqueness question
- There are other rank notions: HOSVD (Tucker decomposition), TT, hierarchichal versions ...

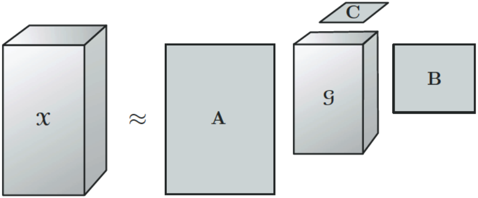


Fig. 4.1 *Tucker decomposition of a three-way array.*

Picture is taken from "Tensor Decompositions and Applications paper by Kolda and Bader

Tensors are more delicate than matrices

Lemma (Classical Johnson-Lindenstrauss lemma)

Take small $\eta > 0$. *Random projection from* $\mathbb{R}^{m'} \to \mathbb{R}^m$ ε*-preserves distances between* K *points with probability* $1 - \eta$ *for* $m \geq \frac{c_\eta \ln K}{\varepsilon^2}$.

Tensors are more delicate than matrices

Lemma (Classical Johnson-Lindenstrauss lemma)

Take small $\eta > 0$. Random projection from $\mathbb{R}^{m'} \to \mathbb{R}^m$ ε-preserves distances between K points with probability $1-\eta$ for $m \geq \frac{c_\eta \ln K}{\varepsilon^2}$.

The strength of JL Lemma is that projection matrix can be taken from a large class of so-called JL-embeddings (including Gaussian, Fast Fourier, as well as sparse matrices and more)

Tensors are more delicate than matrices

Lemma (Classical Johnson-Lindenstrauss lemma)

Take small $\eta > 0$. Random projection from $\mathbb{R}^{m'} \to \mathbb{R}^m$ ε-preserves distances between K points with probability $1 - \eta$ for $m \geq \frac{c_\eta \ln K}{\varepsilon^2}$.

The strength of JL Lemma is that projection matrix can be taken from a large class of so-called JL-embeddings (including Gaussian, Fast Fourier, as well as sparse matrices and more)

These random projections are typically constructed as

$$m \times m' \text{ (random) matrices.}$$

If the data is vectorization of a $n_1 \times n_2 \times \ldots \times n_d$-dimensional tensor, then $m' = \prod n_i$, resulting in a huge $m \times \prod n_i$ projection matrix.

Modewise products: tensor $\times_j$ matrix

Definition (j-mode product, $j = 1, \ldots, d$)

A tensor $\mathcal{X} \in \mathbb{R}^{n^d}$ can be multiplied by a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ to get a tensor $(\mathcal{X} \times_j \mathbf{A}) \in \mathbb{R}^{n \times \ldots \times m \times \ldots \times n}$ with the coordinates

$$(\mathcal{X} \times_j \mathbf{A})(\ldots, i_{j-1}, \ell, i_{j+1}, \ldots) = \sum_{i_j=1}^{n} \mathbf{A}(\ell, i_j)\mathcal{X}(\ldots, i_j, \ldots).$$

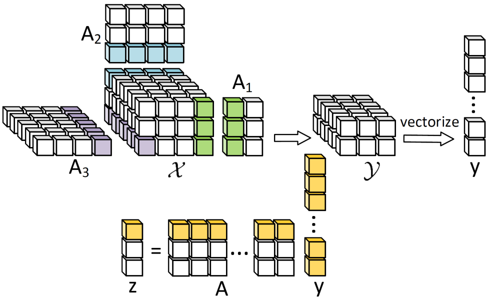
- For a 2 way tensor (a matrix)

$$\mathcal{X} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 = \mathbf{A}_1 \mathcal{X} \mathbf{A}_2^T$$

- For the CP representation, it is equivalent to

$$\mathcal{X} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \ldots \times_d \mathbf{A}_d = \sum_{i=1}^{r} \alpha_i (\mathbf{A}_1 \mathbf{x}_1^i) \bigcirc \ldots \bigcirc (\mathbf{A}_d \mathbf{x}_d^i)$$

Modewise dimension reduction: $L(\mathcal{X}) = \mathbf{A}(vect(\mathcal{X} \times_{j=1}^{d} \mathbf{A}_j))$



Combined size of dimension reduction matrices is

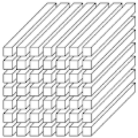
$$m \times \prod_{i=1}^{d} n_i \quad \to \quad \sum_{i=1}^{d} m_i n_i + m' \prod_{i=1}^{d} m_i \quad \text{total}$$

Here, $n_1 = 3, n_2 = 4, n_3 = 5$. Then, $m_1 = 2, m_2 = 3, m_4 = 4$.

Takeaway

Vectorizing tensor data is

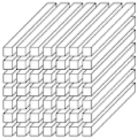
- non-compact
- destroys the data structure
- results in a clumsier object to work with (respective projection matrix must be huge comparing to any of the initial tensor dimensions)



Takeaway

Vectorizing tensor data is

- non-compact
- destroys the data structure
- results in a clumsier object to work with (respective projection matrix must be huge comparing to any of the initial tensor dimensions)



We propose simple, efficient and provable modewise framework for tensor data.

Tensor dimension reduction and low-rank tensor fitting problem[1]

[1] M. Iwen, D. Needell, E. Rebrova, A. Zare, *Lower Memory Oblivious (Tensor) Subspace Embeddings with Fewer Random Bits: Modewise Methods for Least Squares*, accepted to SIMAX

Tensor fitting problem

Fitting problem: For an arbitrary tensor $\mathcal{Y}$, find the closest rank r tensor $\mathcal{X}$:

$$\arg\min_{\mathcal{X}} \|\mathcal{X} - \mathcal{Y}\|_F^2.$$

Tensor fitting problem

Fitting problem: For an arbitrary tensor $\mathcal{Y}$, find the closest rank r tensor $\mathcal{X}$:

$$\arg\min_{\mathcal{X}} \|\mathcal{X} - \mathcal{Y}\|_F^2.$$

Recall that

- Rank-r tensor has rnd degrees of freedom instead of n^d
- exact CP form is NP hard to find.

Tensor fitting problem

Fitting problem: For an arbitrary tensor $\mathcal{Y}$, find the closest rank r tensor $\mathcal{X}$:

$$\arg\min_{\mathcal{X}} \|\mathcal{X} - \mathcal{Y}\|_F^2.$$

Recall that

- Rank-r tensor has *rnd* degrees of freedom instead of n^d
- exact CP form is NP hard to find.

This problem includes finding the best set of unit norm vectors $\{\mathbf{x}_j^i\}$ (basis) and the best set of coefficients $\{\alpha_i\}_{i=1}^r$:

$$\arg\min_{\mathcal{X}} \|\mathcal{X} - \mathcal{Y}\|^2 = \arg\min_{\mathbf{x}_j^i \in \mathbb{R}^n, \alpha_i \in \mathbb{R}} \left\| \sum_{i=1}^{r} \alpha_i \bigcirc_{j=1}^{d} \mathbf{x}_j^i - \mathcal{Y} \right\|_F^2$$

Dimension reduction for tensor fitting problem

Goal: use the geometry preserving modewise dimension reduction to fit a smaller tensor.

Directly reusing the previous result would not work:

- Here, tensor $\mathcal{Y}$ is not low rank
- Tensor $\mathcal{X}$ is low rank, but not from a fixed low-rank subspace

Solving the fitting problem

Idea (ALS: alternating least squares):

- Start with random basis for $\mathcal{X}$: take random unit vectors $\mathbf{x}_j^i \in \mathbb{R}^n$ for $j = 1, \ldots, d$, $i = 1, \ldots, r$
- Fix all but one mode $j \in [d]$, namely, $\mathbf{x}_j^1, \ldots, \mathbf{x}_j^r$
- Optimize over j-th mode
- Repeat for the other modes until some error threshold

Solving the fitting problem

Idea (ALS: alternating least squares):

- Start with random basis for $\mathcal{X}$: take random unit vectors $\mathbf{x}_j^i \in \mathbb{R}^n$ for $j = 1, \ldots, d$, $i = 1, \ldots, r$
- Fix all but one mode $j \in [d]$, namely, $\mathbf{x}_j^1, \ldots, \mathbf{x}_j^r$
- Optimize over j-th mode
- Repeat for the other modes until some error threshold

This turns out to be equivalent to solving n_j separate problems of the form:

Find

$$\underset{\alpha_1,\ldots,\alpha_r \in \mathbb{R}}{\arg\min} \|\mathcal{Z}\| := \underset{\alpha_1,\ldots,\alpha_r \in \mathbb{R}}{\arg\min} \left\| \sum_{i=1}^{r} \alpha_i \bigcirc_{j=1 \neq j'}^{d} \mathbf{x}_j^i - \mathcal{Y}' \right\|^2$$

That is, looking for the best fit in some fixed basis

Subspace oblivious dimension reduction for tensors

What dimension reduction do we need?

- in a geometry preserving and modewise way
- in a subspace oblivious way (to have the same simple operation for the multiple applications in various bases)
  For example, in classical Johnson-Lindenstrauss lemma random matrices are taken from general models

Theorem: general model

Let $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$,

- for a fixed tensor $\mathcal{Y}$
- and any low r-rank tensor $\mathcal{X}$ from a fixed CP subspace (basis)
- for $m \times n$ matrices $\mathbf{A}_j$'s taken from some general (subspace oblivious!) model

we want

$$\left| \|\mathcal{Z}\|^2 - \left\| \mathcal{Z} \bigtimes_{j=1 \neq j}^{d} \mathbf{A}_j \right\|^2 \right| \leq \varepsilon \|\mathcal{Z}\|^2 . \qquad (1)$$

Theorem

If $\mathbf{A}_j \in \mathbb{R}^{m \times n}$ are (η/d)-optimal JL embeddings and $\mathcal{L}$ is spanned by r rank-1 tensors with $\mu_{\mathcal{L}}^{d-1} < \frac{1}{2r}$, and $m \gtrsim \varepsilon^{-2} r d^3$, then (1) is satisfied with probability at least $1 - \eta$.

Tensor norm

We consider $\|\mathcal{X}\|$ = sum of squares of the elements (generalization of the Frobenius norm)

For a rank r tensor,

$$\begin{aligned}\|\mathcal{X}\|^2 &= \sum_{i,j=1}^{r} \alpha_i \alpha_j \left\langle \bigcirc_{\ell=1}^{d} \mathbf{x}_i^{\ell},\ \bigcirc_{\ell=1}^{d} \mathbf{x}_j^{\ell} \right\rangle \\ &= \sum_{i \neq j}^{r} \alpha_i \alpha_j \prod_{\ell=1}^{d} \left\langle \mathbf{x}_i^{\ell},\ \mathbf{x}_j^{\ell} \right\rangle + \|\boldsymbol{\alpha}\|_2^2\end{aligned}$$

Using Cauchy-Swartz, one can estimate

$$\left(1 - \mu'_{\mathcal{X}}\right) \|\boldsymbol{\alpha}\|_2^2 \leq \|\mathcal{X}\|^2 \leq \left(1 + \mu'_{\mathcal{X}}\right) \|\boldsymbol{\alpha}\|_2^2.$$

Johnson-Lindenstrauss embeddings

We are going to consider matrices $\mathbf{A}_j$ such that

Definition (η-optimal family of JL embeddings)

A $m \times n$ matrix $\mathbf{A}$ is an η-optimal JL embedding if for any $\varepsilon \in (0, 1)$ and $\mathcal{S} \subset \mathbb{R}^n$ of cardinality $|\mathcal{S}| \leq \eta e^{\varepsilon^2 m/C}$,

$$\left| \|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \varepsilon \|\mathbf{x}\|_2^2 \text{ for any } \mathbf{x} \in \mathcal{S}$$

with probability at least $1 - \eta$.

Gaussian, Fourier matrices, random projection matrices (to a subspace uniformly selected from the Grassmanian) ...

Johnson-Lindenstrauss embeddings

We are going to consider matrices $\mathbf{A}_j$ such that

Definition (η-optimal family of JL embeddings)

A $m \times n$ matrix $\mathbf{A}$ is an η-optimal JL embedding if for any $\varepsilon \in (0,1)$ and $\mathcal{S} \subset \mathbb{R}^n$ of cardinality $|\mathcal{S}| \leq \eta e^{\varepsilon^2 m/C}$,

$$\left| \|\mathbf{Ax}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \varepsilon \|\mathbf{x}\|_2^2 \text{ for any } \mathbf{x} \in \mathcal{S}$$

with probability at least $1 - \eta$.

Gaussian, Fourier matrices, random projection matrices (to a subspace uniformly selected from the Grassmanian) ...

Definition is inspired by Johnson-Lindenstrauss Lemma:
for any small $\eta > 0$, random projection from $\mathbb{R}^n \to \mathbb{R}^m$ ε-preserves distances between $e^{c(\eta)\varepsilon^2 m}$ points with probability $1 - \eta$.

Modewise (in)coherence

$$\mu_{\mathcal{B}} := \max_{\ell \in [d]} \max_{\substack{k,h \in [r] \\ k \neq h}} \left| \left\langle \mathbf{x}_k^\ell ,\ \mathbf{x}_h^\ell \right\rangle \right| ,$$

- measures angles between all basis vectors (from the same subspaces)
- orthogonal bases have coherence zero

Modewise (in)coherence

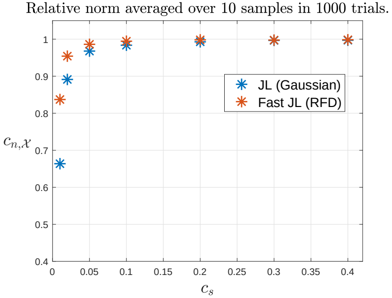
$$\mu_{\mathcal{B}} := \max_{\ell \in [d]} \max_{\substack{k,h \in [r] \\ k \neq h}} \left| \left\langle \mathbf{x}_k^\ell, \ \mathbf{x}_h^\ell \right\rangle \right|,$$

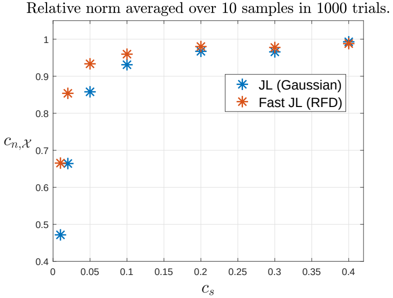
- measures angles between all basis vectors (from the same subspaces)
- orthogonal bases have coherence zero
- random (sub)gaussian tensors are incoherent enough with exponentially high probability:

Lemma

If all components of all vectors $\mathbf{x}_k^{(j)}$ are normalized independent mean zero K-subgaussian random variables, with probability at least $1 - 2r^2 d \exp\left(-c\mu^2 n\right)$ maximum modewise coherence parameter of the tensor $\mathcal{X}$ is at most μ.

Experiments: gaussian and coherent tensors compression





$c_s = m/n$ – compression ratio

$c_{n,\mathcal{X}} = \|\mathcal{X} \times_1 \mathbf{A_1} \ldots \times_d \mathbf{A_d}\| / \|\mathcal{X}\|$ – relative norm

Both data sets contain 10 tensors with $d = 4$, $r = 10$, $n = 100$

Coherent tensors constructed as $1 + \sqrt{0.1} \cdot g$, $g \sim N(0, 1)$

Theorem: general model, optimality

Theorem (informal statement)

If $\mathbf{A}_j \in \mathbb{R}^{m\times n}$ are (η/d)-optimal JL embeddings and $\mathcal{L}$ is spanned by r rank-1 tensors with $\mu_{\mathcal{L}}^{d-1} < \frac{1}{2r}$, and $m \gtrsim \varepsilon^{-2} r d^3$, then (1) is satisfied with probability at least $1-\eta$.

Total number of entries $n^d \to \varepsilon^{-2d} r^d d^{3d}$. Is this optimal?

Theorem: general model, optimality

Theorem (informal statement)

If $\mathbf{A}_j \in \mathbb{R}^{m\times n}$ are (η/d)-optimal JL embeddings and $\mathcal{L}$ is spanned by r rank-1 tensors with $\mu_{\mathcal{L}}^{d-1} < \frac{1}{2r}$, and $m \gtrsim \varepsilon^{-2} r d^3$, then (1) is satisfied with probability at least $1-\eta$.

Total number of entries $n^d \to \varepsilon^{-2d} r^d d^{3d}$. Is this optimal?

The best dependence on ε (distortion) and r (rank) can be estimated as:

- (Larsen, Nelson, 2016) ε^{-2} is optimal for vectors
- A set of all rank r matrices of the size $n \times n$ can be recovered from $O(rn)$ linear measurements.

Theorem: KFJL operators, log-optimal in ε and r

Define (inspired by Jin, Kolda, Ward, 2019):

$$L_{\mathrm{KFJL}}(\mathcal{X}) \;:=\; \mathbf{R}\left(\mathrm{vect}\left(\mathcal{X} \times_1 \mathbf{F}_1\mathbf{D}_1 \cdots \times_d \mathbf{F}_d\mathbf{D}_d\right)\right),$$

$\mathbf{R}$ is a matrix containing m random rows from $Id_{n^d \times n^d}$,
$\mathbf{F}_i \in \mathbb{R}^{n\times n}$ is a unitary discrete Fourier transform matrix,
$\mathbf{D}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix with n random ± 1 entries.

Theorem

Let $\mathcal{L}$ be an r-dimensional subspace of d order tensor space $\mathbb{R}^{n\times n \ldots n}$. Assume that $n^d \gtrsim \eta^{-1}$ and $2r^2 < n^d$. Let $L^1_{\mathrm{KFJL}} : \mathbb{R}^{n\times n\ldots n} \to \mathbb{R}^{m_1}$ and $L^2_{\mathrm{KFJL}} : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$, then for $m_1 \gtrsim_{logs} c^d r^2 \varepsilon^{-2}$ and $m_2 \gtrsim_{logs} c^d \mathbf{r} \cdot \varepsilon^{\mathbf{-2}}$, we have

$$\left| \|\mathcal{Z}\|^2 - \|L^2_{\mathrm{KFJL}}(L^1_{\mathrm{KFJL}}(\mathcal{Z}))\|^2 \right| \leq \varepsilon \|\mathcal{Z}\|^2 ,$$

for $\mathcal{Z} = \mathcal{Y} - \mathcal{X}$ and all $\mathcal{X} \in \mathcal{L}$ with probability at least $1 - \eta$.

Full compression process

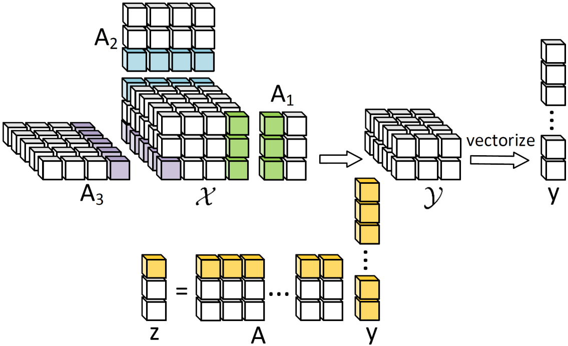
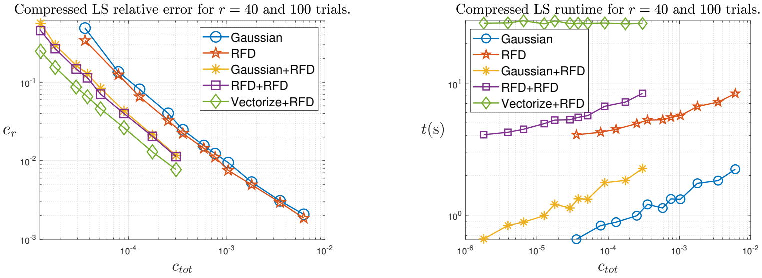


Figure: An example of 2-stage JL embedding applied to a 3-dimensional tensor $\mathcal{X} \in \mathbb{R}^{3\times4\times5}$. Next, the resulting tensor is vectorized (leading to $\mathbf{y} \in \mathbb{R}^{24}$), and a 2^{nd}-stage JL is then performed to obtain $\mathbf{z} = \mathbf{A}\mathbf{y}$ where $\mathbf{A} \in \mathbb{R}^{3\times24}$, and $\mathbf{z} \in \mathbb{R}^{3}$.

Comparing various compression models



(a) error (b) time

Figure: Effect of JL embeddings on the relative reconstruction error of least squares estimation of CPD coefficients. In the 2-stage cases, $c_2 = 0.05$ has been used, $r = 40$.

Connected research-1: compressive sensing

Follow-up work, joint with D. Needell, M. Iwen, M. Perlmutter:

Connected research-1: compressive sensing

Follow-up work, joint with D. Needell, M. Iwen, M. Perlmutter:

Give JL-type guarantees for all rank r tensors with high probability: get (T)RIP restricted isometry property type results - used in compressive sensing algorithms for recovery of a tensor from a few modewise samples (such as Tensor Iterative Hard Thresholding)

Connected research-1: compressive sensing

Follow-up work, joint with D. Needell, M. Iwen, M. Perlmutter:

Give JL-type guarantees for all rank r tensors with high probability: get (T)RIP restricted isometry property type results - used in compressive sensing algorithms for recovery of a tensor from a few modewise samples (such as Tensor Iterative Hard Thresholding)

- Based on supremum of chaos concentration inequality (cf [Krahmer, Mendelson, Rauhut, 2012])
- Preliminary results for low HOSVD rank: partial vectorization seems required, but modewise approach still crucial for memory saving
- Second stage deals with nearly orthogonal decomposition ("generalized HOSVD"), our key lemma proves new complexity estimate for such tensors

Connected research-2: machine learning

Topic modeling on text data[2]

[2] L. Kassab, A. Kryschenko, H. Lyu, D. Molitor, D. Needell, E. Rebrova, *On Nonnegative Matrix and Tensor Decompositions for COVID-19 Twitter Dynamics*, submitted

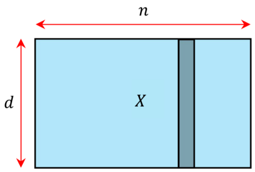
Twitter data related to COVID-19 (Feb-May 2020)

Tweet $\rightarrow$ vector (bag-of-words/TFIDF/word embeddings)

- $d = 5000$ words/terms (dictionary size)
- $n = 90K$ tweets (1000 top retweeted tweets $\times$ 90 days)

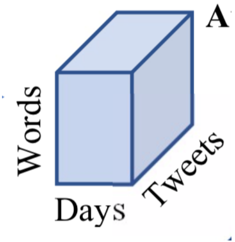
All data . . .

. . . as a matrix:

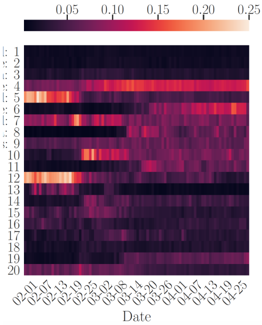


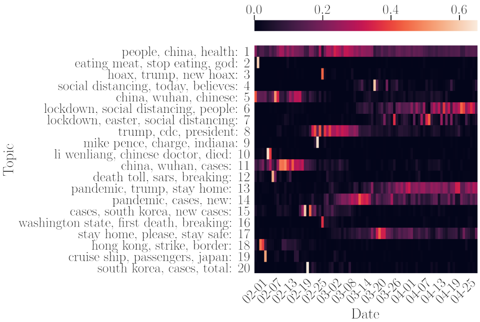
Data

. . . as a tensor with temporal component:



Dynamic topic modeling on matrix/tensor data





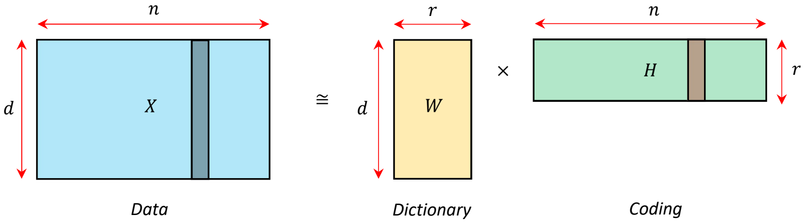
Date	News
2 Feb	‘No Meat, No Coronavirus’ (Wire 2020)
3 Feb	COVID-19 cruise ship outbreak (Kakimoto et al. 2020)
7 Feb	Death of Dr. Li Wenliang (BBC 2020)
8 Feb	COVID-19 death toll overtakes SARS (CNBC 2020)

18 Feb	Spike of cases in South Korea (Statista 2020)
26 Feb	Mike Pence appointed to lead coronavirus task force (Politico 2020)
28 Feb	‘Trump calls Coronavirus Democrats’ ‘new hoax” (NBCNews 2020a)
29 Feb	First COVID-19 death in the U.S. (NBCNews 2020b)
11 Mar	WHO declares a pandemic (WHO 2020)

Non-negative Matrix Factorization

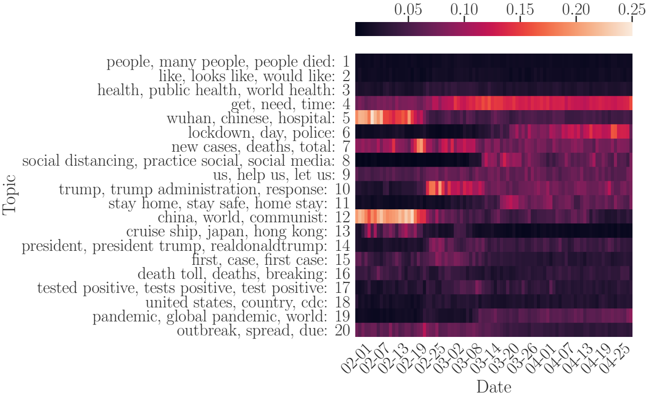
$$\mathbf{X} \in \mathbb{R}_{+}^{d \times n} \approx \mathbf{W} \cdot \mathbf{H}, \quad \mathbf{W} \in \mathbb{R}_{+}^{d \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n}$$

assuming that X is approximately low rank (r)



Provides soft interpretable clustering of the data into r "topics"

Dynamic topic modeling with NMF



H matrix is split into blocks per day and averaged over the rows of the blocks, showing prevalent topics for each day

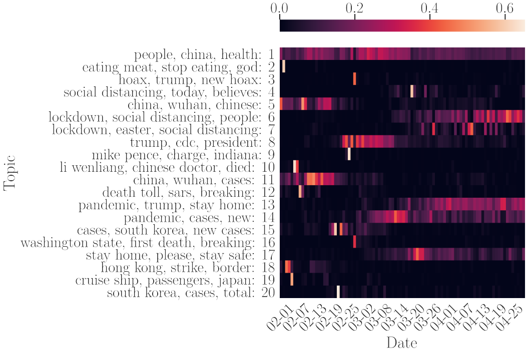
Non-negative low rank CP decomposition (NCPD)

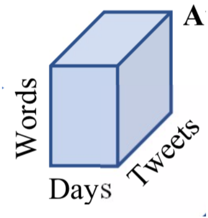
Find non-negative factor matrices $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r] \in \mathbb{R}_+^{n_1 \times r}$, $\mathbf{B} \in \mathbb{R}_+^{n_2 \times r}$, $\mathbf{C} \in \mathbb{R}_+^{n_3 \times r}$ minimizing the reconstruction error

$$arg \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathcal{X} - \sum_{k=1}^{r} \mathbf{a}_k \bigcirc \mathbf{b}_k \bigcirc \mathbf{c}_k \right\|_F^2$$

- **A** is time representation of topics, i.e. the prevalence of each topic through time (emerging, trending, fading away, etc.)
- **B** is term representation of the topics, i.e. words that characterize each topic
- **C** is tweet representation of the topics, i.e. tweets associated with each topic.

Picking up short-term topics from tensor data





Topic 1:	people (0.29)	china (0.22)	health (0.17)	outbreak (0.17)	like (0.15)
Topic 2:	eating meat (0.55)	stop eating (0.13)	god (0.12)	ji maharaj (0.10)	sin (0.10)
Topic 3:	hoax (0.55)	trump (0.18)	new hoax (0.10)	called hoax (0.10)	democrats (0.08)
Topic 4:	social distancing (0.83)	today (0.04)	believes (0.04)	practice social (0.04)	currently (0.04)
Topic 5:	china (0.51)	wuhan (0.24)	chinese (0.14)	pakistan (0.06)	pakstandswithchina (0.05)
Topic 6:	lockdown (0.70)	social distancing (0.12)	people (0.07)	government (0.06)	trump (0.06)
Topic 7:	lockdown (0.75)	easter (0.08)	social distancing (0.07)	day (0.06)	stayhome (0.05)
Topic 8:	trump (0.43)	cdc (0.18)	president (0.18)	realdonaldtrump (0.11)	administration (0.10)
Topic 9:	mike pence (0.46)	charge (0.20)	indiana (0.13)	hiv (0.13)	response (0.08)
Topic 10:	li wenliang (0.34)	chinese doctor (0.23)	died (0.17)	dr li (0.16)	warn (0.10)
Topic 11:	china (0.42)	wuhan (0.18)	cases (0.14)	chinese (0.13)	new (0.13)

Further directions

Modewise tensor dimension reduction can help here:

- Dimension reduction for non-negative tensor fitting problem would speed up finding NCPD decomposition significantly

  Warning: practically, non-negative decompositions are frequently done via multiplicative updates, not alternating least squares with thresholding

- Modewise dimension reduction on the data itself can reduce memory and enforce privacy on specific modes (such as, user/tweet mode)

Connected research-3: numerical linear algebra

Solving the fitting problem, we deal with many least square problems

$$\underset{\alpha_1,\ldots,\alpha_r\in\mathbb{R}}{\arg\min}\left\|\sum_{i=1}^{r}\alpha_i\bigcirc_{j=1\neq j'}^{d}\mathbf{x}_j^i-\mathcal{Y}'\right\|^2$$

Essentially, we solve many inconsistent linear systems of the type $A\mathbf{x}=\mathbf{b}$. Let $\tilde{\mathbf{b}}=A\alpha$, where α is its least square solution. Two natural cases are:

- Noise: $A\mathbf{x}=\tilde{\mathbf{b}}$ and $\left\|\tilde{\mathbf{b}}-\mathbf{b}\right\|\leq\varepsilon$
- Corruptions: $\mathbf{b}$ is obtained by large changes on some of the entries of $\tilde{\mathbf{b}}$

See appendix!

Current directions

1. **Matrices**:
   - Delocalization of eigenvectors of graph Laplacians with the applications to signal processing on graphs (uncertainty principle, with P. Salanevich)
2. **Optimization beyond linear systems**:
   - Theoretical guarantees for stochastic gradient methods (with H.Lyu, W. Swartworth, D. Needell)
   - Algorithms for more general noise/corruption models, randomization of other projection-based algorithms (e.g., Douglas-Rachford method)
3. **Tensors**:
   - Tensor restricted isometry property (with M.Iwen, W. Swartsworth, M. Perlmutter)
   - Tensor fitting for scientific data (with Y.H. Tang)
4. **Machine learning beyond topic modeling**:
   - Non-negative low rank methods for regression problems, guided clustering, etc (collaborators from UCLA)

Literature

Low-rank methods

- Structure preserving tensor dimension reduction
- Matrix and tensor low rank non-negative decompositions for interpretable machine learning

Lower Memory Oblivious (Tensor) Subspace Embeddings with Fewer Random Bits: Modewise Methods for Least Squares
... M. Iwen, D. Needell, E. Rebrova, A. Zare
... SIAM Journal on Matrix Analysis and Applications (SIMAX), 2020

On Nonnegative Matrix and Tensor Decompositions for COVID-19 Twitter Dynamics
... L. Kassab, A. Kryschenko, H.Lyu, D. Molitor, D. Needell, E. Rebrova

COVID-19 Literature Topic-Based Search via Hierarchical NMF
... R. Grotheer, K. Ha, L. Huang, Y. Huang, A. Kryshchenko, O. Kryshchenko, P. Li, X. Li, D. Needell, E. Rebrova
... Proc. NLP-COVID19-EMNLP (2020)

On A Guided Nonnegative Matrix Factorization
... J. Vendrow, J. Haddock, E. Rebrova, D. Needell

Iterative methods for optimization

- Solving linear systems with randomized iterative methods (Randomized Kaczmarz and SGD)
- Sketch-and-project framework: Gaussian block (matrix) sketches
- Avoiding adversarial corruptions

Quantile-based Iterative Methods for Corrupted Systems of Linear Equations
... J. Haddock, D. Needell, E. Rebrova, W. Swartworth

On block Gaussian sketching for the Kaczmarz method
... E. Rebrova, D. Needell
... Numerical algorithms (NUMA), 2019

Stochastic Gradient Descent Methods for Corrupted Systems of Linear Equations
... J. Haddock, D. Needell, E. Rebrova, W. Swartworth
... Proc. Conference on Information Sciences and Systems, 2020

In almost all papers the authors have equal contribution and are listed alphabetically

Thanks for your attention!
Questions?