Modewise methods for tensor dimension reduction

Liza Rebrova

Department of Mathematics
UCLA

Mathematics Graduate Seminar, SCU Channel Islands

March 16 2020

Joint work with Mark Iwen, Deanna Needell, and Ali Zare

Tensors and Kronecker/outer products

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

(for simplicity, in this talk, let's assume all $n_i = n$)

Rank 1 matrix can be defined as $\mathbf{x} \otimes \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{x} \otimes \mathbf{y} = \begin{bmatrix} \mathbf{x}(1)\mathbf{y}(1) & \ldots & \mathbf{x}(1)\mathbf{y}(n) \\ \mathbf{x}(2)\mathbf{y}(1) & \ldots & \mathbf{x}(2)\mathbf{y}(n) \\ \ldots & \ldots & \ldots \\ \mathbf{x}(n)\mathbf{y}(1) & \ldots & \mathbf{x}(n)\mathbf{y}(n) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(1) \\ \mathbf{x}(2) \\ \ldots \\ \mathbf{x}(n) \end{bmatrix} \circ \begin{bmatrix} \mathbf{y}(1) & \ldots & \mathbf{y}(n) \end{bmatrix}$$

By analogy, we define rank 1 tensor as $\mathcal{X} := \mathbf{x}_1 \otimes \ldots \otimes \mathbf{x}_d$,

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

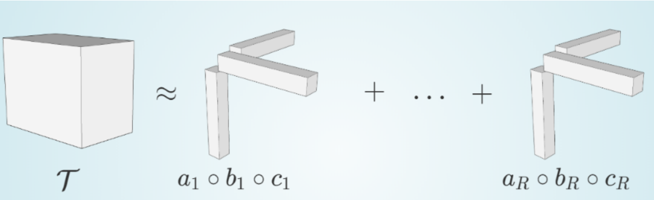
Tensor (CP) rank

(Candecomp-Parafac) rank r tensor as

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

Normalization: we always assume $\|\mathbf{x}_j^i\|_2 = 1$. Clearly, $r \leq n^d$.

For example, for a 3-way (3 modes) tensor,



Picture is taken from Medium post "Tensor Factorization for Graph Analysis in Python" by Claudio Stamile

Fitting problem

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

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

Tensor norm here is a generalization of the Frobenius matrix norm (sum of squares of all entries of the tensor)

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

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

Solving the fitting problem

Idea:

- 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

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

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

Dimension reduction for the fitting problem

Goal: reduce the size of this problem.

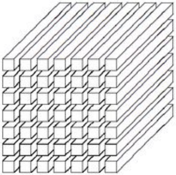
Preferably,

- in a subspace oblivious way (to have the same simple operation for the multiple applications in various bases)
  For example, classical dimension reduction lemma

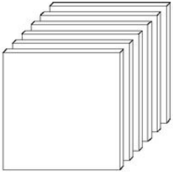
Lemma (Johnson-Lindenstrauss)

Take 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 $1 - \eta$.

- without vectorization of the tensors



(c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$



(c) Frontal slices: $\mathbf{X}_{::k}$ (or $\mathbf{X}_k$)

Picture is taken from Kolda&Bader paper

Modewise products: tensor $\times_j$ matrix

Definition (j-mode product)

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 any $j = 1, \ldots, n$.

For example, for a 2 way tensor (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

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

So, instead of

Fitting problem: $\|\mathcal{X} - \mathcal{Y}\|^2 \to \min$

$$\operatorname*{arg\,min}_{\alpha_1,\dots,\alpha_r \in \mathbb{R}} \| \sum_{i=1}^{r} \alpha_i \bigotimes_{j=1 \neq j'}^{d} \mathbf{x}_j^i - \mathcal{Y} \|^2$$

let us find

Reduced fitting problem: $\|\mathcal{X} \times_{j=1}^{d} \mathbf{A}_j - \mathcal{Y} \times_{j=1}^{d} \mathbf{A}_j\|^2 \to \min$

$$\operatorname*{arg\,min}_{\alpha_1,\dots,\alpha_r \in \mathbb{R}} \| \sum_{i=1}^{r} \alpha_i \bigotimes_{j=1}^{d} \mathbf{A}_j \mathbf{x}_j^i - \mathcal{Y} \bigtimes_{j=1}^{d} \mathbf{A}_j \|^2$$

Will it find us a good solution for the original (non-reduced) problem?

Subspace oblivious dimension reduction for tensors

For now: let $\mathcal{Y} = 0$.

We want

$$\left| \|\mathcal{X}\|^2 - \|\mathcal{X} \bigtimes_{j=1 \neq j}^{d} \mathbf{A}_j\|^2 \right| \leq \varepsilon \|\mathcal{X}\|^2$$

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

Johnson-Lindenstrauss embeddings

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

Definition (η-optimal family of JL embeddins)

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) ...

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 $1-\eta$.

Main theorem -1

Theorem (Iwen-Needell-R.-Zare)

Let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{R}^{n^d}$ spanned by a basis $\mathcal{B} := \left\{ \bigcirc_{\ell=1}^{d} \mathbf{x}_k^{(\ell)} \mid k \in [r] \right\}$. If all $\mathbf{A}_j \in \mathbb{R}^{m \times n}$ from an (η/d)-optimal family of JL embeddings, $m \gtrsim \varepsilon^{-2} r^{2/d} d^2$, then with probability at least $1-\eta$

$$\left| \|\mathcal{X}\|^2 - \|\mathcal{X} \bigtimes_{j=1}^{d} \mathbf{A}_j\|^2 \right| \le \varepsilon \, \|\mathbf{a}\|_2^2 \, ,$$

for all $\mathcal{X} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_1^i \otimes \ldots \otimes \mathbf{x}_d^i \in \mathcal{L}$.

Total number of entries $N = n^d \to M \sim \varepsilon^{-2d} r^2 d^{2d}$.

Main theorem-1

Theorem (Iwen-Needell-R.-Zare)

Let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{R}^{n^d}$ spanned by a basis $\mathcal{B} := \left\{ \bigcirc_{\ell=1}^{d} \mathbf{x}_k^{(\ell)} \right\}_{k\in[r]}$ with modewise coherence $\mu_{\mathcal{B}}^{d-1} < 1/2r$.
If all $\mathbf{A}_j \in \mathbb{R}^{m\times n}$ from an (η/d)-optimal family of JL embeddings with $m \gtrsim \varepsilon^{-2} r^{2/d} d^2$, then with probability at least $1-\eta$

$$\left| \|\mathcal{X}\|^2 - \|\mathcal{X} \bigtimes_{j=1}^{d} \mathbf{A}_j\|^2 \right| \le \varepsilon \|\mathcal{X}\|^2,$$

for all $\mathcal{X} \in \mathcal{L}$.

Total number of entries $N = n^d \to M \sim \varepsilon^{-2d} r^2 d^{2d}$.

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.

Theorem 2: Fitting an arbitrary $\mathcal{X}$

Theorem (Iwen-Needell-R.-Zare)

Let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{R}^{n^d}$ spanned by a basis $\mathcal{B} := \left\{ \bigcirc_{\ell=1}^{d} \mathbf{x}_k^{(\ell)} \right\}_{k \in [r]}$ with $\mu_{\mathcal{B}}^{d-1} < 1/2r$ and $\mathcal{Y} \notin \mathcal{L}$.
If all $\mathbf{A}_j \in \mathbb{R}^{m \times n}$ are from an (η/d)-optimal family of JL embeddings with $m \gtrsim \varepsilon^{-2} r d^3$, then with probability at least $1 - \eta$

$$\left| \|\mathcal{Y} - \mathcal{X}\|^2 - \|(\mathcal{Y} - \mathcal{X}) \bigtimes_{j=1}^{d} \mathbf{A}_j\|^2 \right| \leq \varepsilon \|\mathcal{Y}\|^2 ,$$

for all $\mathcal{X} \in \mathcal{L}$.

Total number of entries $N = n^d \to M \sim \varepsilon^{-2d} r^d d^{3d}$.
Reason: we need to additionally compress a subspace spanned by $\{P_{\mathcal{L}^\perp}(\mathcal{Y}) \pm \mathcal{B}\}$, this basis is NOT low rank.

Can we do better?

Is our dependence on r and on ε (and on d) good?

Lemma (Larsen, Nelson, 2016)

For any $n, d \geq 2$, there exists a set of n vectors in $\mathbb{R}^d$ so that any linear map $\mathbb{R}^d \to \mathbb{R}^m$, ε-preserving distances between them, must have

$$m \gtrsim \varepsilon^{-2} \ln n.$$

Modewise Fourier JL for a finite set

Consider a special modewise operator:

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

$\mathrm{vect} : \mathbb{R}^{n\times\cdots\times n} \to \mathbb{R}^{n^d}$ is the vectorization operator,
$\mathbf{R}$ 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 (Yin, Kolda, Ward, 2019)

Let $\eta \gtrsim n^{-d}$. Consider $\mathcal{S} \subset \mathbb{R}^{n^d}$ of cardinality $|\mathcal{S}| = p$. Then with probability at least $1-\eta$ the linear operator L_{FJL} is an ε-JL embedding of $\mathcal{S}$ into $\mathbb{R}^m$, where

$$m \;\gtrsim\; \varepsilon^{-2} \cdot \log^{2d-1}\left(\frac{\max(p, n^d)}{\eta}\right) \cdot \log n^d.$$

Moreover, if $d = 1$, then we may replace $\max(p, n^1)$ with p.

Yin–Kolda–Ward Theorem/Theorem 2:

Let us compare these two modewise JL-type embedding results:

- For a fixed finite set S / for a fixed subset $\mathcal{L}$
- Special Fourier modewise transform /large class of JL-type modewise maps
- $m \gtrsim \varepsilon^{-2}$ / $m \gtrsim \varepsilon^{-2d}$
- for any subset of tensors / only for incoherent bases

Idea: using Yin–Kolda–Ward Theorem to improve ε-dependence and to get rid of the incoherence assumption

How can this help with subspace embeddings?

Two ways to apply JL-type results to a low r-dimensional subspace (it is enough to approximate unit norm tensors only!):

- To an ε-net on $\mathcal{S}^{r-1}$:

Lemma (JL discretization)

Fix $\varepsilon \in (0,1)$. Let $\mathcal{L}$ be an r-dimensional subspace of $\mathbb{R}^n$, and let $\mathcal{N} \subset \mathcal{L}$ be an $(\epsilon/16)$-net of the unit sphere $\mathcal{S}^{r-1} \subset \mathcal{L}$. Then, if $\mathbf{A} \in \mathbb{R}^{m \times n}$ is an $(\varepsilon/2)$-JL embedding of $\mathcal{N}$ it will also satisfy

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2 \text{ for all } \mathbf{x} \in \mathcal{L}.$$

There exists an $(\varepsilon/16)$-net such that $|\mathcal{N}| \leq \left(\frac{47}{\varepsilon}\right)^r$.

- To a set of r basis vectors:
  Recall Theorem 1 above

Using Yin–Kolda–Ward theorem: wrong way

1. Apply Yin–Kolda–Ward Theorem to the approximation net $S = \mathcal{N}$ of cardinality $\left(\frac{47}{\varepsilon}\right)^r$
2. Use JL Discretization Lemma

Resulting dimension is at least

$$m \gtrsim \varepsilon^{-2} r^{2d-1} \cdot \log^{2d-1}\left(\frac{47}{\eta^{1/r}\varepsilon}\right) \cdot \log n^d.$$

So, ε dependence improves, but dependence on the rank even become worse: r^{2d-1} instead of r^d (Theorem 2)

Using Yin–Kolda–Ward theorem: right way

1. Apply Yin–Kolda–Ward Theorem to the set of r basis vectors
2. Proceed like we did for Theorem 2 to get the estimate for all others

Resulting dimension (since $r < n^d$):

$$m \gtrsim \varepsilon^{-2} r^2 \cdot \log^{2d-1}\left(\frac{n^d}{\eta}\right) \cdot \log n^d.$$

Much better! :)
Still quadratic dependence on rank...

Improved two step dimension reduction

Let us vectorize the result of Step 2 to get a vector (tensor with $d = 1$) in $\mathbb{R}^m$,

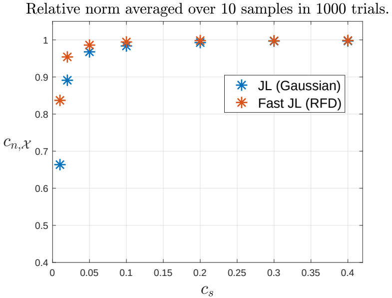
3. Now, apply Yin–Kolda–Ward Theorem to the approximation net $S = \mathcal{N}$ of cardinality $\left(\frac{47}{\varepsilon}\right)^r$ in $\mathbb{R}^m$

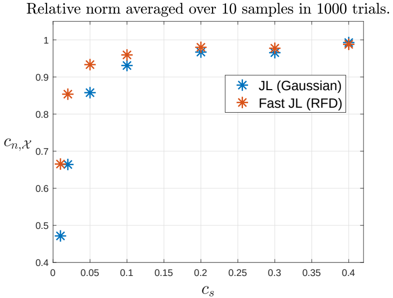
to get

$$\tilde{m} \gtrsim \varepsilon^{-2} r \cdot \log\left(\frac{47}{\varepsilon \eta^{1/r}}\right) \cdot \log m.$$

Optimal dependence on both ε and r! (and a bit of logarithmic multiples...)

Experiments-1





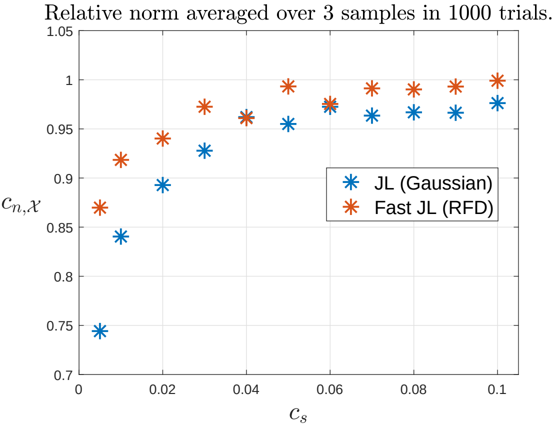
$c_s = m/n$ – compression ratio

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

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

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

Experiments-2

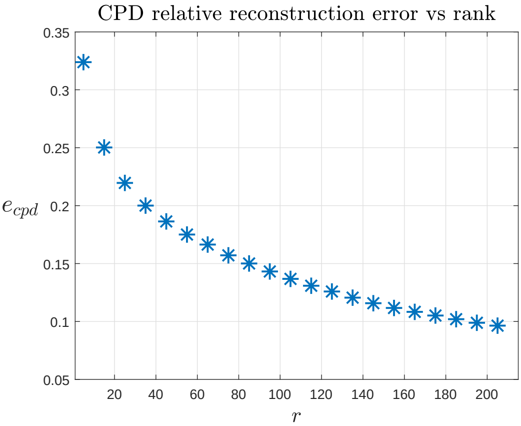


The same for MRI data: three 3-mode MRI images of size

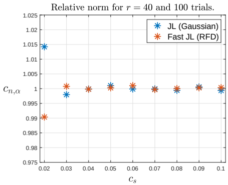
$240 \times 240 \times 155$

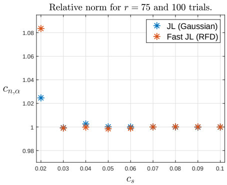
What was the rank r?

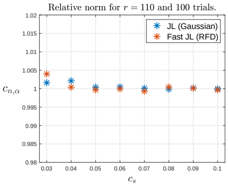
Experiments-3



Experiments-4







Some future directions

- Remove theoretical incoherence assumption in Theorem 2 (which is still the most general model for modewise compression!)
- Give JL-type guarantees for all CP-rank r tensors with high probability: get RIP (restricted isometry property) type results.

Thanks for your attention!

QUESTIONS?