Krylov methods for tensors

Lars Eldén and Berkant Savas

Department of Mathematics
Linköping University, Sweden

EMMDS 2009

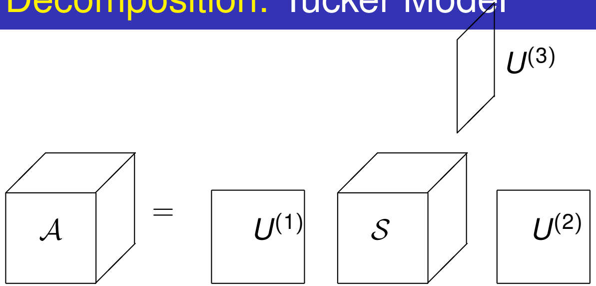
Outline

Tech Report

Download the tech report from

`http://www.mai.liu.se/~besav/tensor-krylov.pdf`

Tensor Decomposition: Tucker Model



- Tucker 1964, numerous papers in psychometrics and chemometrics
- De Lathauwer et al., SIMAX 2000: notation, theory.
- The matrices $U^{(i)}$ are usually orthogonal.

This talk: Tucker model for 3-tensors only! Generalization straightforward.

Mode$-I$ Multiplication of a Tensor by a Matrix

Assume that dimensions are such that all operations are well-defined.
Mostly 3-tensors. Lim's notation. (No standard notation yet)

$$\mathcal{B} = (X)_1 \cdot \mathcal{A}, \qquad \mathcal{B}(i,j,k) = \sum_{\nu=1}^{n} x_{i\nu} a_{\nu jk}.$$

All column vectors are multiplied by the matrix X.
Multiplication in all modes at the same time:

$$\mathcal{B} = (X, Y, Z) \cdot \mathcal{A}, \qquad \mathcal{B}(i,j,k) = \sum_{\nu,\mu,\lambda} x_{i\nu} y_{j\mu} z_{k\lambda} a_{\nu\mu\lambda}.$$

For convenience we write

$$\mathcal{B} = (X^T, Y^T, Z^T) \cdot \mathcal{A} = \mathcal{A} \cdot (X, Y, Z)$$

Inner Product and Norm

Inner product (contraction: $\mathbb{R}^{n \times n \times n} \to \mathbb{R}$)

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k} a_{ijk} b_{ijk}$$

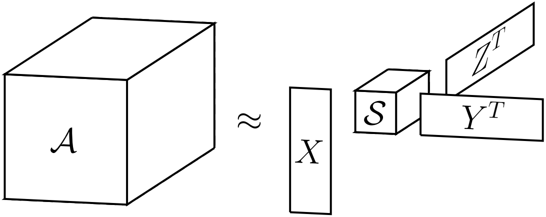
The Frobenius norm:

$$\|\mathcal{A}\| = \langle \mathcal{A}, \mathcal{A} \rangle^{1/2}$$

Matrix case

$$\langle A, B \rangle = \mathrm{tr}(A^T B)$$

Best Rank$-(r_1, r_2, r_3)$ Approximation



Best rank$-(r_1, r_2, r_3)$ approximation:

$$\min_{X,Y,Z,\mathcal{S}} \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{S}\|, \qquad X^T X = I, \quad Y^T Y = I, \quad Z^T Z = I$$

The problem is over-parameterized!

Best Approximation

$$\min_{\operatorname{rank}(\mathcal{B})=(r_1,r_2,r_3)} \|\mathcal{A} - \mathcal{B}\|$$

is equivalent to

$$\max_{X,Y,Z} \Phi(X,Y,Z) = \frac{1}{2}\|\mathcal{A} \cdot (X,Y,Z)\|^2$$
$$= \frac{1}{2}\sum_{j,k,l}\left(\sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu} x_{\lambda j} y_{\mu k} z_{\nu l}\right)^2,$$

subject to

$$X^T X = I_{r_1}, \qquad Y^T Y = I_{r_2}, \qquad Z^T Z = I_{r_3}$$

Grassmann Optimization

The Frobenius norm is invariant under orthogonal transformations:

$$\Phi(X, Y, Z) = \Phi(XU, YV, ZW) = \frac{1}{2}\|\mathcal{A} \cdot (XU, YV, ZW)\|^2$$

for orthogonal $U \in \mathbb{R}^{r_1 \times r_1}$, $V \in \mathbb{R}^{r_2 \times r_2}$, and $W \in \mathbb{R}^{r_3 \times r_3}$.
Maximize Φ over equivalence classes

$$[X] = \{XU \mid U \text{ orthogonal}\}.$$

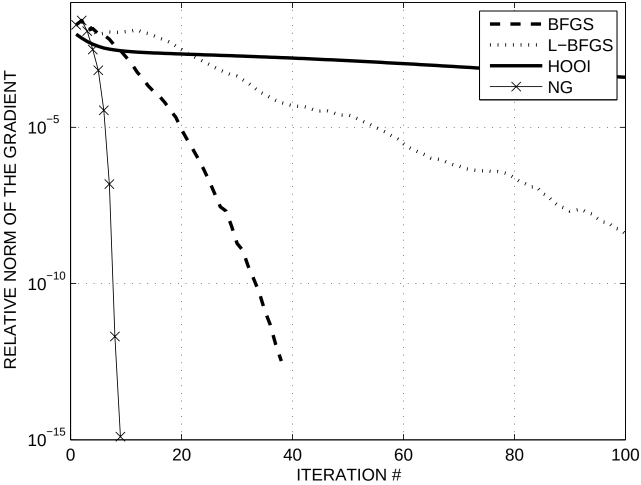
Product of manifolds: $\mathrm{Gr}^3 = \mathrm{Gr}(J, r_1) \times \mathrm{Gr}(K, r_2) \times \mathrm{Gr}(L, r_3)$

$$\max_{(X,Y,Z)\in \mathrm{Gr}^3} \Phi(X, Y, Z) = \max_{(X,Y,Z)\in \mathrm{Gr}^3} \frac{1}{2}\langle \mathcal{A} \cdot (X, Y, Z), \mathcal{A} \cdot (X, Y, Z)\rangle$$

Methods for Best Approximation

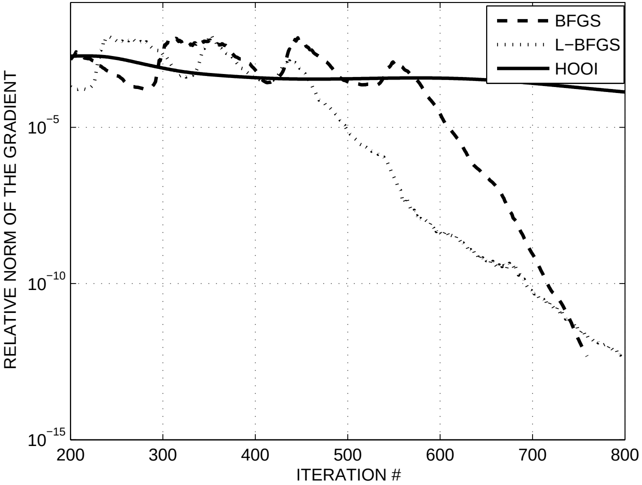
- Grassmann-based
  1. Newton (LE, B. Savas)
  2. Trust region/Newton (Ishteva, De Lathauwer et al.)
  3. BFGS quasi-Newton (Savas, Lim)
  4. Limited memory BFGS (Savas, Lim)
- Alternating
  1. HOOI (Kroonenberg, De Lathauwer)

Numerical Example I



A random tensor $\mathcal{A} \in \mathbb{R}^{20\times20\times20}$ with random entries N(0, 1) approximated with a rank $-(5, 5, 5)$ tensor.

Numerical Example II



A random tensor $\mathcal{A} \in \mathbb{R}^{100\times100\times100}$ with random entries N(0, 1) approximated with a rank $-(5, 10, 20)$ tensor.

Sparse Tensors in Information Sciences

In information sciences the tensors are often sparse:

- Term-document-author analysis (Dunlavy et al)
- Graphs, web link analysis (Kolda et al, PARAFAC model)

$$a_{ijk} = \begin{cases} 1 & \text{if page } i \text{ points to page } j \text{ using term } k \\ 0 & \text{otherwise} \end{cases}$$

Web page with links

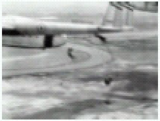
Welcome to Wikipedia,

the free encyclopedia that anyone can edit.

2,696,663 articles in English

Overview · Editing · Questions · Help

Today's featured article



The **Battle of Dien Bien Phu** was the climactic battle of the First Indochina War between French Union forces and Viet Minh communist revolutionary forces. The battle occurred between March and May 1954, and culminated in a massive French defeat that effectively ended the war. The French undertook to create an air-supplied base at Dien Bien Phu, deep in the hills of Vietnam, in order to cut off Viet Minh supply lines into the neighboring French protectorate of Laos. The Viet Minh, under General Vo Nguyen Giap, surrounded and besieged the French, who were unaware of the Viet Minh's possession of heavy artillery. The Viet Minh occupied the highlands around Dien Bien Phu, and were able to fire down accurately onto French positions. Tenacious fighting on the ground ensued, reminiscent of the

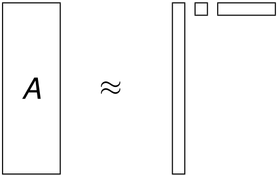
In t

- Th pl Sl in
- Fl di
- A m
- Ru de re
- Jc

Sparse Matrices

Krylov methods give low rank approximations:

$$AV_k = U_k H_k$$

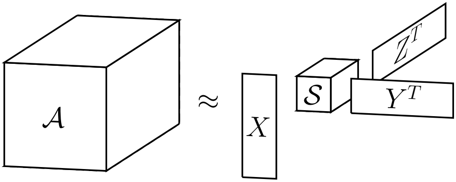


$$= U_k H_k V_k^T.$$

The matrix is only used as operator: $u = Av$

Sparse Tensors

Can we generalize Krylov methods to tensors and obtain low rank approximations?



Golub-Kahan Bidiagonalization for Rectangular Matrix

- $\beta_1 u_1 = b$, $v_0 = 0$
- **for** $i = 1 : k$
$$\alpha_i v_i = A^T u_i - \beta_i v_{i-1},$$
$$\beta_{i+1} u_{i+1} = A v_i - \alpha_i u_i$$
- **end**

The coefficients α_i and β_i are chosen to normalize the vectors.

Equivalent to Partial Least Squares (PLS)

Golub-Kahan Bidiagonalization for Rectangular Matrix

- $\beta_1 u_1 = b$, $v_0 = 0$
- **for** $i = 1 : k$

$$\alpha_i v_i = A^T u_i - \beta_i v_{i-1}, \qquad [\alpha_i v_i = A \cdot (u_i)_1 - \beta_i v_{i-1},]$$

$$\beta_{i+1} u_{i+1} = A v_i - \alpha_i u_i \qquad [\beta_{i+1} u_{i+1} = A \cdot (v_i)_2 - \alpha_i u_i]$$

- **end**

The coefficients α_i and β_i are chosen to normalize the vectors.

Krylov Method for Tensor Approximation

Arnoldi style (i.e., including Gram-Schmidt orthogonalization)

- Let u_1 and v_1 be given
- $h_{111} w_1 = \mathcal{A} \cdot (u_1, v_1)_{1,2}$
- **for** $\nu = 2 : m$

$$
\begin{aligned}
&h_u = \mathcal{A} \cdot (U_{\nu-1}, v_{\nu-1}, w_{\nu-1}) \\
&h_{\nu,\nu-1,\nu-1} u_\nu = \mathcal{A} \cdot (v_{\nu-1}, w_{\nu-1})_{2,3} - U_{\nu-1} h_u \\
&h_v = \mathcal{A} \cdot (u_\nu, V_{\nu-1}, w_{\nu-1}) \\
&h_{\nu,\nu,\nu-1} v_\nu = \mathcal{A} \cdot (u_\nu, w_{\nu-1})_{1,3} - V_{\nu-1} h_v \\
&h_w = \mathcal{A} \cdot (u_\nu, v_\nu, W_{\nu-1}) \\
&h_{\nu\nu\nu} w_\nu = \mathcal{A} \cdot (u_\nu, v_\nu)_{1,2} - W_{\nu-1} h_w
\end{aligned}
$$

- **end**

Approximate

$$
\mathcal{A} \approx (U_m, V_m, W_m) \cdot \mathcal{H}, \qquad \mathcal{H} = \left(U_m^T, V_m^T, W_m^T\right) \cdot \mathcal{A}
$$

Krylov Method

Arnoldi style (i.e., including Gram-Schmidt orthogonalization)

- Let u_1 and v_1 be given
- $h_{111} w_1 = \mathcal{A} \cdot (u_1, v_1)_{1,2}$
- **for** $\nu = 2 : m$

$$
\begin{aligned}
&h_u = \mathcal{A} \cdot (U_{\nu-1}, v_{\nu-1}, w_{\nu-1}) \\
&h_{\nu,\nu-1,\nu-1} u_\nu = \mathcal{A} \cdot (v_{\nu-1}, w_{\nu-1})_{2,3} - U_{\nu-1} h_u \\
&h_v = \mathcal{A} \cdot (u_\nu, V_{\nu-1}, w_{\nu-1}) \\
&h_{\nu,\nu,\nu-1} v_\nu = \mathcal{A} \cdot (u_\nu, w_{\nu-1})_{1,3} - V_{\nu-1} h_v \\
&h_w = \mathcal{A} \cdot (u_\nu, v_\nu, W_{\nu-1}) \\
&h_{\nu\nu\nu} w_\nu = \mathcal{A} \cdot (u_\nu, v_\nu)_{1,2} - W_{\nu-1} h_w
\end{aligned}
$$

- **end**

Gram-Schmidt, closer look

- **for** $\nu = 2 : m$

  $h_u = \mathcal{A} \cdot (U_{\nu-1}, v_{\nu-1}, w_{\nu-1})$

  $h_{\nu,\nu-1,\nu-1} u_\nu = \mathcal{A} \cdot (v_{\nu-1}, w_{\nu-1})_{2,3} - U_{\nu-1} h_u$

  ...

  ...

- **end**

The algebra is straightforward:

- h_u is a vector
- u-vectors live in first mode, $U_{\nu-1} = (u_1, u_2, \ldots, u_{\nu-1})$
- Multiply by $U_{\nu-1}$ in first mode:

$$h_{\nu,\nu-1,\nu-1} U_{\nu-1}^T u_\nu = \mathcal{A} \cdot (U_{\nu-1}, v_{\nu-1}, w_{\nu-1}) - h_u = 0$$

Minimal Krylov Method

- Let u_1 and v_1 be given
- $h_{111} w_1 = \mathcal{A} \cdot (u_1, v_1)_{1,2}$
- **for** $\nu = 2 : m$

$$h_u = \mathcal{A} \cdot (U_{\nu-1}, v_{\nu-1}, w_{\nu-1})$$
$$h_{\nu,\nu-1,\nu-1} u_\nu = \mathcal{A} \cdot (v_{\nu-1}, w_{\nu-1})_{2,3} - U_{\nu-1} h_u$$
$$h_v = \mathcal{A} \cdot (u_\nu, V_{\nu-1}, w_{\nu-1})$$
$$h_{\nu,\nu,\nu-1} v_\nu = \mathcal{A} \cdot (u_\nu, w_{\nu-1})_{1,3} - V_{\nu-1} h_v$$
$$h_w = \mathcal{A} \cdot (u_\nu, v_\nu, W_{\nu-1})$$
$$h_{\nu\nu\nu} w_\nu = \mathcal{A} \cdot (u_\nu, v_\nu)_{1,2} - W_{\nu-1} h_w$$

- **end**

Richer combinatorial structure:
Let $\mu \leq \nu - 1$ and $\lambda \leq \nu - 1$:

$$h_u = \mathcal{A} \cdot (U_{\nu-1}, v_\mu, w_\lambda)$$
$$h u_\nu = \mathcal{A} \cdot (v_\mu, w_\lambda)_{2,3} - U_{\nu-1} h_u$$

Maximal Krylov Method

u_1 and v_1 are given, all possible combinations are formed

1. $\{u_1\} \times \{v_1\} \longrightarrow w_1$
2. $\{v_1\} \times \{w_1\} \longrightarrow u_2$
3. $\begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} \times \{w_1\} \longrightarrow \begin{Bmatrix} v_2 \\ v_3 \end{Bmatrix}$
4. $\begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} \times \begin{Bmatrix} v_1 \\ v_2 \\ v_3 \end{Bmatrix} \longrightarrow \begin{Bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{Bmatrix}$
5. $\begin{Bmatrix} v_1 \\ v_2 \\ v_3 \end{Bmatrix} \times \begin{Bmatrix} w_1 \\ w_2 \\ \vdots \\ w_6 \end{Bmatrix} \longrightarrow \begin{Bmatrix} u_2 \\ u_3 \\ \vdots \\ u_{19} \end{Bmatrix}$

Maximal Krylov recursion

$\hat{h}w_1 = \mathcal{A} \cdot (u_1, v_1)_{1,2}$
$\nu = \mu = \lambda = 1$
while $\nu \leq \nu_{\max}$ and $\mu \leq \mu_{\max}$ and $\lambda \leq \lambda_{\max}$ **do**
  %—— *u-loop* ——%
  for all $(\bar{\mu}, \bar{\lambda})$ such that $\bar{\mu} \leq \mu$ and $\bar{\lambda} \leq \lambda$ **do**
    if the pair $(\bar{\mu}, \bar{\lambda})$ has not been used before **then**
      $\nu = \nu + 1$
      $h_u = \mathcal{A} \cdot (U_{\nu-1}, v_{\bar{\mu}}, w_{\bar{\lambda}})$
      $h_{\nu\bar{\mu}\bar{\lambda}} u_\nu = \mathcal{A} \cdot (v_{\bar{\mu}}, w_{\bar{\lambda}})_{2,3} - U_{\nu-1} h_u$
      $\mathcal{H}(:, \bar{\mu}, \bar{\lambda}) = \begin{pmatrix} h_u \\ h_{\nu\bar{\mu}\bar{\lambda}} \end{pmatrix}$ % Mode 1
    end if
  end for

Maximal Krylov Method II

%—— *v-loop* ——

ANALOGOUS

%—— *w-loop* ——%

ANALOGOUS

end while

Krylov factorization for maximal recursion

Theorem (Tensor Krylov factorizations)

After a complete u-loop:

$$\mathcal{A} \cdot (V_k, W_l)_{2,3} = \left(U_j\right)_1 \cdot \mathcal{H}_{jkl}. \tag{1}$$

After a complete v-loop:

$$\mathcal{A} \cdot \left(U_j, W_l\right)_{1,3} = (V_m)_2 \cdot \mathcal{H}_{jml}. \tag{2}$$

After a complete w-loop:

$$\mathcal{A} \cdot \left(U_j, V_m\right)_{1,2} = (W_n)_3 \cdot \mathcal{H}_{jmn}. \tag{3}$$

Krylov-Schur Method for Refining Eigenvalues

Arnoldi factorization:

$$AV = VH + vc^{\mathsf{T}}, \qquad H = XSX^{\mathsf{T}}, \tag{4}$$

where S is triangular.
Put $U = VX$ $\Longrightarrow$ Krylov-Schur factorization:

$$AU = US + ub^{\mathsf{T}}, \tag{5}$$

Partition:

$$A(U_1 \; U_2) = (U_1 \; U_2) \begin{pmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{pmatrix} + u(b_1^{\mathsf{T}} \; b_2^{\mathsf{T}}); \tag{6}$$

Discard the part that contains unwanted eigenvalues:

$$AU_1 = U_1 S_{11} + ub_1^{\mathsf{T}}, \tag{7}$$

Restart!

Krylov-Schur-like Method for Best Approximation

1. Compute a Krylov factorization:

$$\mathcal{A} \cdot (V, W)_{2,3} = (U)_1 \cdot \mathcal{H}.$$

2. Compute a best low-rank approximation (or HOSVD) of $\mathcal{H}$, make a change of bases, and truncate the factorization.
3. Restart!

Applications: are these methods useful?

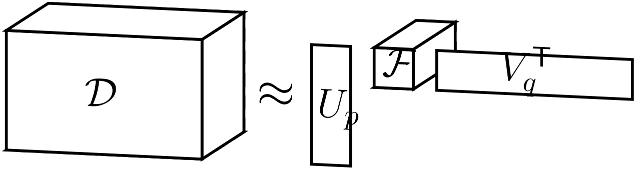
Handwritten digit classification on US Postal Service database digits

16 $\times$ 16 pixel, gray scale images

0	1	2	3	4	5	6	7	8	9
1194	1005	731	658	652	556	664	645	542	644
359	264	198	166	200	160	170	147	166	177

- 7291 training digits
- 2007 test digits

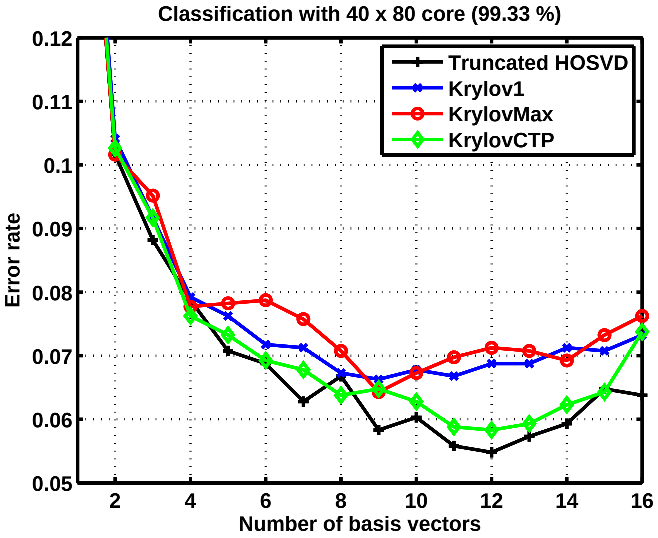
Comparison: relative error



Relative error $= \|\mathcal{D} - \mathcal{D}_{p,q,10}\| / \|\mathcal{D}\|$ in %

$p \times q$	20×40	30×60	40×80
THOSVD	29.6	24.7	21.1
$\langle \mathcal{D}, \mathcal{D} \rangle_{-i}$	33.9	28.8	24.9
$\mathcal{K}_{min}(\mathcal{D}, u, v, w)$	39.5	32.8	27.4
$\mathcal{K}_{max}(\mathcal{D}, u, v, w)$	35.5	31.5	28.4

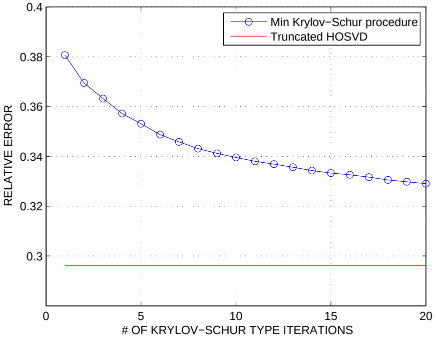
Error rate

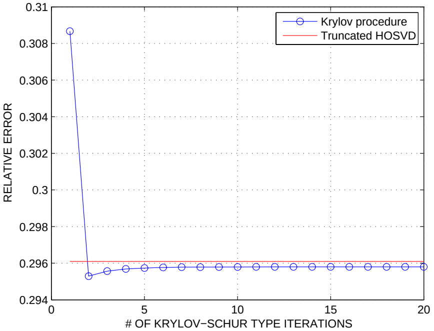


Classification with 40 x 80 core (99.33 %)
0.12
0.11
0.1
0.09
0.08
0.07
0.06
0.05
Error rate
2
4
6
8
10
12
14
16
Number of basis vectors
Truncated HOSVD
Krylov1
KrylovMax
KrylovCTP

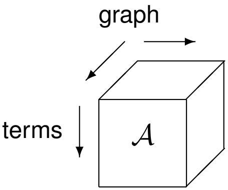
Restarted Krylov approach

$$\mathcal{D} \approx (U_{20}, V_{40})_{1,2} \cdot \mathcal{F}.$$





Text-graph analysis



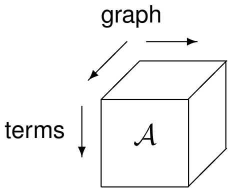
Query: q, starting vector for Krylov
Low rank approximation:

$$\mathcal{A} \approx (Q, V, W) \cdot \mathcal{S}$$

Find relevant documents:

$$\left(q^T\right)_1 \cdot \mathcal{A} \approx \left(q^T Q, V, W\right) \cdot \mathcal{S} = \left(e_1^T, V, W\right) \cdot \mathcal{S}$$

Collapses the term mode: graph matrix



Query: q, starting vector for Krylov
Low rank approximation:

$$\mathcal{A} \approx (Q, V, W) \cdot \mathcal{S}$$

Find relevant documents:

$$\left(q^T\right)_1 \cdot \mathcal{A} \approx \left(q^T Q, V, W\right) \cdot \mathcal{S} = \left(e_1^T, V, W\right) \cdot \mathcal{S}$$

Collapses the term mode: graph matrix

Conclusions I

- Tensor methods/algorithms without index-wrestling
  - Indices hidden using matrix-inspired notation and object-oriented software
  - Generalization to higher order tensors is straightforward
- Grassmann optimization (for Tucker model)
  - Needed because tensors cannot be deflated like matrices
  - Unconstrained optimization
  - Newton: Quadratic convergence
- Sparse tensors: Krylov methods
  - Suitable for
    - sparse tensors
    - tensors whose dimensions vary rapidly (new data)
  - What are the convergence properties?
  - How to construct a method that gives an intermediate between the minimal and maximal sequence?

Conclusions II

- Which variant is more economical in terms of the number of tensor-vector operations, taking into account the convergence rate?
- If one of the modes is of small dimension, so that a complete basis is quickly computed, how can one modify the recursion so that one can use the already computed information as efficiently as possible?
- .....

- Many fundamental mathematical and algorithmic problems remain
- Numerous new applications in information sciences
- Tensor algorithms and computations can be (easily) managed if we define the right abstractions!