Numerical multilinear algebra in data analysis

(Ten ways to decompose a tensor)

Lek-Heng Lim

Stanford University

April 5, 2007

Ten ways to decompose a tensor

1. Complete triangular decomposition
2. Complete orthogonal decomposition
3. **Higher order singular value decomposition**
4. Higher order nonnegative matrix decomposition
5. **Outer product decomposition**
6. **Nonnegative outer product decomposition**
7. **Symmetric outer product decomposition**
8. **Block outer product decomposition**
9. Kronecker product decomposition
10. Coclustering decomposition

Idea

rank → *rank revealing decomposition* → *low-rank approximation* → *data analytic model*

Data mining in the olden days

- **Spectroscopy:** measure light absorption/emission of specimen as function of energy.
- Typical **specimen** contains 10^{13} to 10^{16} light absorbing entities or **chromophores** (molecules, amino acids, etc).

Fact (Beer's Law)

$A(\lambda) = -\log(I_1/I_0) = \varepsilon(\lambda)c$. $A =$ absorbance, $I_1/I_0 =$ fraction of intensity of light of wavelength λ that passes through specimen, $c =$ concentration of chromophores.

- Multiple chromophores ($k = 1, \dots, r$) and wavelengths ($i = 1, \dots, m$) and specimens/experimental conditions ($j = 1, \dots, n$),

$$A(\lambda_i, s_j) = \sum\nolimits_{k=1}^{r} \varepsilon_k(\lambda_i) c_k(s_j).$$

- Bilinear model aka **factor analysis**: $A_{m\times n} = E_{m\times r} C_{r\times n}$ rank-revealing factorization or, in the presence of noise, low-rank approximation $\min \|A_{m\times n} - E_{m\times r} C_{r\times n}\|$.

Modern data mining

- **Text mining** is the spectroscopy of documents.
- Specimens = **documents**.
- Chromophores = **terms**.
- Absorbance = inverse document frequency:

$$A(t_i) = -\log\left(\sum\nolimits_j \chi(f_{ij})/n\right).$$

- Concentration = term frequency: f_{ij}.
- $\sum_j \chi(f_{ij})/n$ = fraction of documents containing t_i.
- $A \in \mathbb{R}^{m\times n}$ term-document matrix. $A = QR = U\Sigma V^{\mathsf{T}}$ rank-revealing factorizations.
- Bilinear models:
  - Gerald Salton et. al.: **vector space model** (QR);
  - Sue Dumais et. al.: **latent sematic indexing** (SVD).

Bilinear models

- Bilinear models work on 'two-way' data:
  - measurements on object i (genomes, chemical samples, images, webpages, consumers, etc) yield a vector $\mathbf{a}_i \in \mathbb{R}^n$ where $n =$ number of features of i;
  - collection of m such objects, $A = [\mathbf{a}_1, \dots, \mathbf{a}_m]$ may be regarded as an m-by-n matrix, e.g. gene × microarray matrices in bioinformatics, terms × documents matrices in text mining, facial images × individuals matrices in computer vision.
- Various matrix techniques may be applied to extract useful information: QR, EVD, SVD, NMF, CUR, compressed sensing techniques, etc.
- Examples: vector space model, factor analysis, principal component analysis, latent semantic indexing, PageRank, EigenFaces.
- Some problems: **factor indeterminacy** — $A = XY$ rank-revealing factorization not unique; unnatural for k-**way data** when $k > 2$.

Ubiquity of multiway data

- **Batch data:** batch × time × variable
- **Time-series analysis:** time × variable × lag
- **Computer vision:** people × view × illumination × expression × pixel
- **Bioinformatics:** gene × microarray × oxidative stress
- **Phylogenetics:** codon × codon × codon
- **Analytical chemistry:** sample × elution time × wavelength
- **Atmospheric science:** location × variable × time × observation
- **Psychometrics:** individual × variable × time
- **Sensory analysis:** sample × attribute × judge
- **Marketing:** product × product × consumer

Fact (Inevitable consequence of technological advancement)

Increasingly sophisticated instruments, sensor devices, data collecting and experimental methodologies lead to increasingly complex data.

Tensors: computer scientist's definition

- **Data structure:** k-array $A = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$
- **Algebraic structure:**
  1. **Addition/scalar multiplication:** for $[\![b_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$, $\lambda \in \mathbb{R}$,
  $$[\![a_{ijk}]\!] + [\![b_{ijk}]\!] := [\![a_{ijk} + b_{ijk}]\!] \quad \textit{and} \quad \lambda [\![a_{ijk}]\!] := [\![\lambda a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$$
  2. **Multilinear matrix multiplication:** for matrices $L = [\lambda_{i'i}] \in \mathbb{R}^{p \times l}, M = [\mu_{j'j}] \in \mathbb{R}^{q \times m}, N = [\nu_{k'k}] \in \mathbb{R}^{r \times n}$,
  $$(L, M, N) \cdot A := [\![c_{i'j'k'}]\!] \in \mathbb{R}^{p \times q \times r}$$
  where
  $$c_{i'j'k'} := \sum\nolimits_{i=1}^{l} \sum\nolimits_{j=1}^{m} \sum\nolimits_{k=1}^{n} \lambda_{i'i} \mu_{j'j} \nu_{k'k} a_{ijk}.$$
- Think of A as 3-dimensional array of numbers. $(L, M, N) \cdot A$ as multiplication on '3 sides' by matrices L, M, N.
- Generalizes to arbitrary order k. If $k = 2$, ie. matrix, then $(M, N) \cdot A = MAN^{\mathsf{T}}$.

Tensors: mathematician's definition

- U, V, W vector spaces. Think of $U \otimes V \otimes W$ as the vector space of all formal linear combinations of terms of the form $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$,

$$\sum \alpha \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w},$$

  where $\alpha \in \mathbb{R}, \mathbf{u} \in U, \mathbf{v} \in V, \mathbf{w} \in W$.

- One condition: $\otimes$ decreed to have the multilinear property

$$\begin{aligned}(\alpha \mathbf{u}_1 + \beta \mathbf{u}_2) \otimes \mathbf{v} \otimes \mathbf{w} &= \alpha \mathbf{u}_1 \otimes \mathbf{v} \otimes \mathbf{w} + \beta \mathbf{u}_2 \otimes \mathbf{v} \otimes \mathbf{w},\\ \mathbf{u} \otimes (\alpha \mathbf{v}_1 + \beta \mathbf{v}_2) \otimes \mathbf{w} &= \alpha \mathbf{u} \otimes \mathbf{v}_1 \otimes \mathbf{w} + \beta \mathbf{u} \otimes \mathbf{v}_2 \otimes \mathbf{w},\\ \mathbf{u} \otimes \mathbf{v} \otimes (\alpha \mathbf{w}_1 + \beta \mathbf{w}_2) &= \alpha \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}_1 + \beta \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}_2.\end{aligned}$$

- Up to a choice of bases on U, V, W, $\mathbf{A} \in U \otimes V \otimes W$ can be represented by a 3-way array $A = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$.

Tensors: physicist's definition

- "What are tensors?" $\equiv$ "What kind of physical quantities can be represented by tensors?"
- Usual answer: if they satisfy some 'transformation rules' under a change-of-coordinates.

Theorem (Change-of-basis)

Two representations A, A' of $\mathbf{A}$ in different bases are related by

$$(L, M, N) \cdot A = A'$$

with L, M, N respective change-of-basis matrices (non-singular).

- Pitfall: tensor fields (roughly, tensor-valued functions on manifolds) often referred to as tensors — stress tensor, piezoelectric tensor, moment-of-inertia tensor, gravitational field tensor, metric tensor, curvature tensor.

Outer product

- If $U = \mathbb{R}^l$, $V = \mathbb{R}^m$, $W = \mathbb{R}^n$, $\mathbb{R}^l \otimes \mathbb{R}^m \otimes \mathbb{R}^n$ may be identified with $\mathbb{R}^{l \times m \times n}$ if we define $\otimes$ by
$$\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} = [\![u_i v_j w_k]\!]_{i,j,k=1}^{l,m,n}.$$
- A tensor $A \in \mathbb{R}^{l \times m \times n}$ is said to be decomposable if it can be written in the form
$$A = \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$$
for some $\mathbf{u} \in \mathbb{R}^l, \mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$. For order 2, $\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^{\mathsf{T}}$.
- In general, any $A \in \mathbb{R}^{l \times m \times n}$ may be written as a sum of decomposable tensors
$$A = \sum\nolimits_{i=1}^{r} \lambda_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i.$$
- May be written as a multilinear matrix multiplication:
$$A = (U, V, W) \cdot \Lambda.$$
$U \in \mathbb{R}^{l \times r}, V \in \mathbb{R}^{m \times r}, W \in \mathbb{R}^{n \times r}$ and diagonal $\Lambda \in \mathbb{R}^{r \times r \times r}$.

Tensor ranks

- **Matrix rank.** $A \in \mathbb{R}^{m\times n}$

$$\begin{aligned}\operatorname{rank}(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1}, \dots, A_{\bullet n}\}) && \text{(column rank)}\\ &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet}, \dots, A_{m\bullet}\}) && \text{(row rank)}\\ &= \min\{r \mid A = \textstyle\sum_{i=1}^{r} \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}\} && \text{(outer product rank).}\end{aligned}$$

- **Multilinear rank.** $A \in \mathbb{R}^{l\times m\times n}$. $\operatorname{rank}_{\boxplus}(A) = (r_1(A), r_2(A), r_3(A))$ where

$$\begin{aligned}r_1(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet\bullet}, \dots, A_{l\bullet\bullet}\})\\ r_2(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1\bullet}, \dots, A_{\bullet m\bullet}\})\\ r_3(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet\bullet 1}, \dots, A_{\bullet\bullet n}\})\end{aligned}$$

- **Outer product rank.** $A \in \mathbb{R}^{l\times m\times n}$.

$$\operatorname{rank}_{\otimes}(A) = \min\{r \mid A = \textstyle\sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i\}$$

- In general, $\operatorname{rank}_{\otimes}(A) \neq r_1(A) \neq r_2(A) \neq r_3(A)$.

Properties of matrix rank

1. Rank of $A \in \mathbb{R}^{m \times n}$ **easy to determine** (Gaussian elimination)
2. Best rank-r approximation to $A \in \mathbb{R}^{m \times n}$ **always exist** (Eckart-Young theorem)
3. Best rank-r approximation to $A \in \mathbb{R}^{m \times n}$ **easy to find** (singular value decomposition)
4. Pick $A \in \mathbb{R}^{m \times n}$ at random, then A has **full rank with probability** 1, ie. $\operatorname{rank}(A) = \min\{m, n\}$
5. $\operatorname{rank}(A)$ from a **non-orthogonal** rank-revealing decomposition (e.g. $A = L_1 D L_2^{\mathsf{T}}$) and $\operatorname{rank}(A)$ from an **orthogonal** rank-revealing decomposition (e.g. $A = Q_1 R Q_2^{\mathsf{T}}$) are **equal**
6. $\operatorname{rank}(A)$ is **base field independent**, ie. same value whether we regard A as an element of $\mathbb{R}^{m \times n}$ or as an element of $\mathbb{C}^{m \times n}$

Properties of outer product rank

1. Computing $\operatorname{rank}_{\otimes}(A)$ for $A \in \mathbb{R}^{l \times m \times n}$ is **NP-hard** [Håstad 1990]
2. For some $A \in \mathbb{R}^{l \times m \times n}$, $\operatorname{argmin}_{\operatorname{rank}_{\otimes}(B) \leq r} \|A - B\|_F$ **does not have a solution**
3. When $\operatorname{argmin}_{\operatorname{rank}_{\otimes}(B) \leq r} \|A - B\|_F$ does have a solution, computing the solution is an **NP-complete** problem in general
4. For some l, m, n, if we sample $A \in \mathbb{R}^{l \times m \times n}$ at random, there is **no** r such that $\operatorname{rank}_{\otimes}(A) = r$ **with probability** 1
5. An outer product decomposition of $A \in \mathbb{R}^{l \times m \times n}$ with **orthogonality constraints** on X, Y, Z will in general require a sum with **more than** $\operatorname{rank}_{\otimes}(A)$ number of terms
6. $\operatorname{rank}_{\otimes}(A)$ is **base field dependent**, ie. value depends on whether we regard $A \in \mathbb{R}^{l \times m \times n}$ or $A \in \mathbb{C}^{l \times m \times n}$

Properties of multilinear rank

1. Computing $\operatorname{rank}_{\boxplus}(A)$ for $A \in \mathbb{R}^{l \times m \times n}$ is **easy**
2. Solution to $\operatorname{argmin}_{\operatorname{rank}_{\boxplus}(B) \leq (r_1, r_2, r_3)} \|A - B\|_F$ **always exist**
3. Solution to $\operatorname{argmin}_{\operatorname{rank}_{\boxplus}(B) \leq (r_1, r_2, r_3)} \|A - B\|_F$ **easy to find**
4. Pick $A \in \mathbb{R}^{l \times m \times n}$ at random, then A has

   $$\operatorname{rank}_{\boxplus}(A) = (\min(l, mn), \min(m, ln), \min(n, lm))$$

   with probability 1
5. If $A \in \mathbb{R}^{l \times m \times n}$ has $\operatorname{rank}_{\boxplus}(A) = (r_1, r_2, r_3)$. Then there exist full-rank matrices $X \in \mathbb{R}^{l \times r_1}$, $Y \in \mathbb{R}^{m \times r_2}$, $Z \in \mathbb{R}^{n \times r_3}$ and core tensor $C \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ such that $A = (X, Y, Z) \cdot C$. X, Y, Z **may be chosen to have orthonormal columns**
6. $\operatorname{rank}_{\boxplus}(A)$ is **base field independent**, ie. same value whether we regard $A \in \mathbb{R}^{l \times m \times n}$ or $A \in \mathbb{C}^{l \times m \times n}$

Outer product decomposition in spectroscopy

- Application to fluorescence spectral analysis by Rasmus Bro.
- Specimens with a number of pure substances in different concentration
  - a_{ijk} = fluorescence emission intensity at wavelength λ_j^{em} of ith sample excited with light at wavelength λ_k^{ex}.
  - Get 3-way data $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
  - Get outer product decomposition of A

$$A = \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \cdots + \mathbf{x}_r \otimes \mathbf{y}_r \otimes \mathbf{z}_r.$$

- Get the true chemical factors responsible for the data.
  - r: number of pure substances in the mixtures,
  - $\mathbf{x}_\alpha = (x_{1\alpha}, \dots, x_{l\alpha})$: relative concentrations of αth substance in specimens $1, \dots, l$,
  - $\mathbf{y}_\alpha = (y_{1\alpha}, \dots, y_{m\alpha})$: excitation spectrum of αth substance,
  - $\mathbf{z}_\alpha = (z_{1\alpha}, \dots, z_{n\alpha})$: emission spectrum of αth substance.
- Noisy case: find best rank-r approximation (CANDECOMP/PARAFAC).

Multilinear decomposition in bioinformatics

- Application to cell cycle studies by Alter and Omberg.
- Collection of gene-by-microarray matrices $A_1, \dots, A_l \in \mathbb{R}^{m \times n}$ obtained under varying oxidative stress.
  - a_{ijk} = expression level of jth gene in kth microarray under ith stress.
  - Get 3-way data array $A = \llbracket a_{ijk} \rrbracket \in \mathbb{R}^{l \times m \times n}$.
  - Get multilinear decomposition of A

    $$A = (X, Y, Z) \cdot C,$$

    to get orthogonal matrices X, Y, Z and core tensor C by applying SVD to various 'flattenings' of A.
- Column vectors of X, Y, Z are 'principal components' or 'parameterizing factors' of the spaces of stress, genes, and microarrays; C governs interactions between these factors.
- Noisy case: approximate by discarding small c_{ijk} (Tucker Model).

Fundamental problem of multiway data analysis

$$\operatorname{argmin}_{\operatorname{rank}(B)\leq r}\|A - B\|$$

Examples

1. **Outer product rank:** $A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i$:
$$\min\|A - \mathbf{u}_1\otimes\mathbf{v}_1\otimes\mathbf{w}_1 - \mathbf{u}_2\otimes\mathbf{v}_2\otimes\mathbf{w}_2 - \cdots - \mathbf{u}_r\otimes\mathbf{v}_r\otimes\mathbf{z}_r\|.$$
2. **Multilinear rank:** $A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $C \in \mathbb{R}^{r_1\times r_2\times r_3}$, $L_i \in \mathbb{R}^{d_i\times r_i}$:
$$\min\|A - (L_1, L_2, L_3)\cdot C\|.$$
3. **Symmetric rank:** $A \in \mathsf{S}^k(\mathbb{C}^n)$, find $\mathbf{u}_i$:
$$\min\|A - \mathbf{u}_1^{\otimes k} - \mathbf{u}_2^{\otimes k} - \cdots - \mathbf{u}_r^{\otimes k}\|.$$
4. **Nonnegative rank:** $0 \leq A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $\mathbf{u}_i \geq 0, \mathbf{v}_i \geq 0, \mathbf{w}_i \geq 0$.

Feature revelation

- More generally, $\mathcal{D}$ = dictionary. Minimal r with

$$A \approx \alpha_1 B_1 + \cdots + \alpha_r B_r \in \mathcal{D}_r.$$

$B_i \in \mathcal{D}$ often reveal features of the dataset A.

Examples

1. **PARAFAC:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_\otimes(A) \leq 1\}$.
2. **Tucker:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_\boxplus(A) \leq (1, 1, 1)\}$.
3. **De Lathauwer:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_\boxplus(A) \leq (r_1, r_2, r_3)\}$.
4. **ICA:** $\mathcal{D} = \{A \in \mathsf{S}^k(\mathbb{C}^n) \mid \operatorname{rank}_\mathsf{S}(A) \leq 1\}$.
5. **NTF:** $\mathcal{D} = \{A \in \mathbb{R}_+^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_+(A) \leq 1\}$.

A simple result

Lemma (de Silva and Lim)

Let $r \geq 2$ *and* $k \geq 3$*. Given the norm-topology on* $\mathbb{R}^{d_1 \times \cdots \times d_k}$*, the following statements are equivalent:*

1. *The set* $\mathcal{S}_r(d_1, \dots, d_k) := \{A \mid \operatorname{rank}_\otimes(A) \leq r\}$ *is not closed.*
2. *There exists* B, $\operatorname{rank}_\otimes(B) > r$*, that may be approximated arbitrarily closely by tensors of strictly lower rank, ie.*
$$\inf\{\|B - A\| \mid \operatorname{rank}_\otimes(A) \leq r\} = 0.$$
3. *There exists* C, $\operatorname{rank}_\otimes(C) > r$*, that does not have a best rank-r approximation, ie.*
$$\inf\{\|C - A\| \mid \operatorname{rank}_\otimes(A) \leq r\}$$
is not attained (by any A *with* $\operatorname{rank}_\otimes(A) \leq r$*).*

Non-existence of best low-rank approximation

Let $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{d_i}$, $i = 1, 2, 3$. Let

$$A := \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{y}_3 + \mathbf{x}_1 \otimes \mathbf{y}_2 \otimes \mathbf{x}_3 + \mathbf{y}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3$$

and for $n \in \mathbb{N}$,

$$A_n := \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes (\mathbf{y}_3 - n\mathbf{x}_3) + \left(\mathbf{x}_1 + \frac{1}{n}\mathbf{y}_1\right) \otimes \left(\mathbf{x}_2 + \frac{1}{n}\mathbf{y}_2\right) \otimes n\mathbf{x}_3.$$

Lemma (de Silva and Lim)

$\operatorname{rank}_{\otimes}(A) = 3$ *iff* $\mathbf{x}_i, \mathbf{y}_i$ *linearly independent,* $i = 1, 2, 3$. *Furthermore, it is clear that* $\operatorname{rank}_{\otimes}(A_n) \leq 2$ *and*

$$\lim_{n \to \infty} A_n = A.$$

Exercise 62, Section 4.6.4, in: D. Knuth, *The art of computer programming*, **2**, 3rd Ed., Addison-Wesley, Reading, MA, 1997.

Bad news: outer product approximations are ill-behaved

Theorem (de Silva and Lim)

1. *Tensors failing to have a best rank-r approximation exist for*
   1. *all* **orders** $k > 2$,
   2. *all* **norms** *and Brègman divergences,*
   3. *all* **ranks** $r = 2, \dots, \min\{d_1, \dots, d_k\}$.
2. *Tensors that fail to have best low-rank approximations occur with* **non-zero probability** *and sometimes with certainty — all* $2 \times 2 \times 2$ *tensors of rank* 3 *fail to have a best rank-*2 *approximation.*
3. *Tensor rank can* **jump arbitrarily large gaps**. *There exists sequence of rank-r tensor converging to a limiting tensor of rank* $r + s$.

Message

- That the best rank-r approximation problem for tensors has no solution poses serious difficulties.
- Incorrect to think that if we just want an 'approximate solution', then this doesn't matter.
- If there is no solution in the first place, then what is it that are we trying to approximate? ie. what is the 'approximate solution' an approximate of?
- Problems near an ill-posed problem are generally **ill-conditioned**.
- Current way to deal with such difficulties — pretend that it doesn't matter.

Some good news: weak solutions may be characterized

- For a tensor A that has no best rank-r approximation, we will call a $C \in \overline{\{A \mid \operatorname{rank}_{\otimes}(A) \leq r\}}$ attaining
$$\inf\{\|C - A\| \mid \operatorname{rank}_{\otimes}(A) \leq r\}$$
a **weak solution**. In particular, we must have $\operatorname{rank}_{\otimes}(C) > r$.

Theorem (de Silva and Lim)

Let $d_1, d_2, d_3 \geq 2$. *Let* $A_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *be a sequence of tensors with* $\operatorname{rank}_{\otimes}(A_n) \leq 2$ *and*
$$\lim_{n \to \infty} A_n = A,$$
where the limit is taken in any norm topology. If the limiting tensor A *has rank higher than* 2*, then* $\operatorname{rank}_{\otimes}(A)$ *must be exactly* 3 *and there exist pairs of linearly independent vectors* $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{R}^{d_1}$, $\mathbf{x}_2, \mathbf{y}_2 \in \mathbb{R}^{d_2}$, $\mathbf{x}_3, \mathbf{y}_3 \in \mathbb{R}^{d_3}$ *such that*
$$A = \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{y}_3 + \mathbf{x}_1 \otimes \mathbf{y}_2 \otimes \mathbf{x}_3 + \mathbf{y}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3.$$

More good news: nonnegative tensors are better behaved

- Let $0 \leq A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. The nonnegative rank of A is

$$\operatorname{rank}_+(A) := \min\bigl\{r \mid \sum\nolimits_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i,\ \mathbf{u}_i, \dots, \mathbf{z}_i \geq 0\bigr\}$$

  Clearly, such a decomposition exists for any $A \geq 0$.

Theorem (Lim)

Let $A = [\![a_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ be nonnegative. Then

$$\inf\bigl\{\bigl\|A - \sum\nolimits_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i\bigr\| \bigm| \mathbf{u}_i, \dots, \mathbf{z}_i \geq 0\bigr\}$$

is always attained.

Corollary

Nonnegative tensor approximation always have solutions.

Algorithms

- Even when an optimal solution B_* to $\operatorname{argmin}_{\operatorname{rank}_\otimes(B)\le r}\|A-B\|_F$ exists, B_* is not easy to compute since the objective function is non-convex.
- A widely used strategy is a nonlinear Gauss-Seidel algorithm, better known as the Alternating Least Squares algorithm:

```
Algorithm: ALS for optimal rank-r approximation
initialize $X^{(0)} \in \mathbb{R}^{l\times r}, Y^{(0)} \in \mathbb{R}^{m\times r}, Z^{(0)} \in \mathbb{R}^{n\times r}$;
initialize $s^{(0)}, \varepsilon > 0, k = 0$;
while $\rho^{(k+1)}/\rho^{(k)} > \varepsilon$;
    $X^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{X}\in\mathbb{R}^{l\times r}} \|T - \sum_{\alpha=1}^{r} \bar{x}_\alpha^{(k+1)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}\|_F^2$;
    $Y^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Y}\in\mathbb{R}^{m\times r}} \|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes \bar{y}_\alpha^{(k+1)} \otimes z_\alpha^{(k)}\|_F^2$;
    $Z^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Z}\in\mathbb{R}^{n\times r}} \|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes \bar{z}_\alpha^{(k+1)}\|_F^2$;
    $\rho^{(k+1)} \leftarrow \|\sum_{\alpha=1}^{r} [x_a^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes z_\alpha^{(k+1)} - x_\alpha^{(k)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}]\|_F^2$;
    $k \leftarrow k+1$;
```

Convex relaxation

- Joint work with Kim-Chuan Toh.
- $F(x_{11}, \dots, z_{nr}) = \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$ is a polynomial.
- **Lasserre/Parrilo strategy:** Find largest λ^* such that $F - \lambda^*$ is a sum of squares. Then λ^* is often min $F(x_{11}, \dots, z_{nr})$.

1. Let $\mathbf{v}$ be the D-tuple of monomials of degree ≤ 6. Since deg(F) is even, $F - \lambda$ may be written as

   $$F(x_{11}, \dots, z_{nr}) - \lambda = \mathbf{v}^\mathsf{T}(M - \lambda E_{11})\mathbf{v}$$

   for some $M \in \mathbb{R}^{D \times D}$.
2. Note RHS is a sum of squares iff $M - \lambda E_{11}$ is positive semi-definite (since $M - \lambda E_{11} = B^\mathsf{T} B$).
3. Get convex problem

   $$\begin{aligned} \text{minimize} \quad & -\lambda \\ \text{subjected to} \quad & \mathbf{v}^\mathsf{T}(S + \lambda E_{11})\mathbf{v} = F, \\ & S \succeq 0. \end{aligned}$$

Convex relaxation

- **Complexity:** for rank-r approximations to order-k tensors $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, $D = \binom{r(d_1+\cdots+d_k)+k}{k}$ — large even for moderate d_i, r and k.
- **Sparsity:** our polynomials are always sparse (eg. for $k = 3$, only terms of the form xyz or $x^2y^2z^2$ or $uvwxyz$ appear). This can be exploited.

Theorem (Reznick)

If $f(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x})^2$, then the powers of the monomials in p_i must lie in $\frac{1}{2}\operatorname{Newton}(f)$.

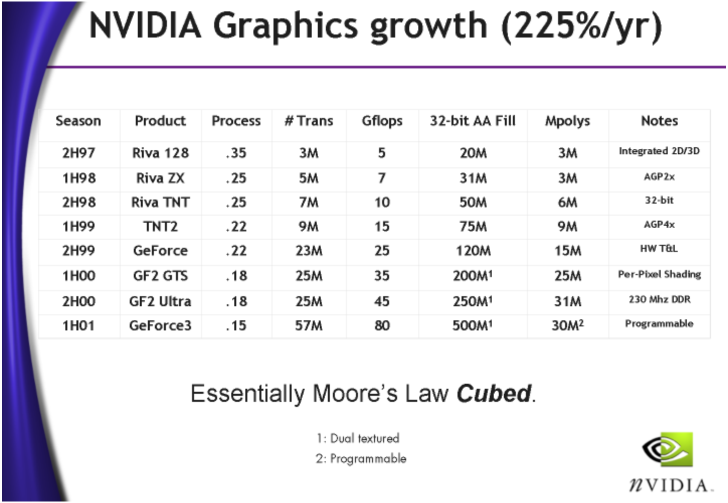
- So if $f(x_{11}, \dots, z_{nr}) = \sum_{j=1}^{N} p_j(x_{11}, \dots, z_{nr})^2$, then only 1 and monomials of the form $x_{i\alpha}y_{j\alpha}z_{k\alpha}$ may occur in $p_1, \dots, p_N$.
- Complexity is reduced to $rlmn + 1$ from $\binom{r(l+m+n)+3}{3}$.

Exploiting semiseparability

- Joint work with Ming Gu.
- **Gauss-Newton Method:** $g(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2$. Approximate Hessian using Jacobian: $H_g \approx J_{\mathbf{f}}^{\mathsf{T}} J_{\mathbf{f}}$.
- The Hessian of $F(X, Y, Z) = \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$ can be approximated by a semiseparable matrix.
- This is the case even when X, Y, Z are required to be nonnegative.
- **Goal:** Exploit this in optimization algorithms.

Basic multilinear algebra subroutines?

- Multilinear matrix multiplication $(L_1, \dots, L_k) \cdot A$ is **data parallel**.
- **GPGPU:** general purpose computations on graphics hardware.
- **Kirk's Law:** GPU speed behaves like Moore's Law cubed.



NVIDIA Graphics growth (225%/yr)

Season	Product	Process	# Trans	Gflops	32-bit AA Fill	Mpolys	Notes
2H97	Riva 128	.35	3M	5	20M	3M	Integrated 2D/3D
1H98	Riva ZX	.25	5M	7	31M	3M	AGP2x
2H98	Riva TNT	.25	7M	10	50M	6M	32-bit
1H99	TNT2	.22	9M	15	75M	9M	AGP4x
2H99	GeForce	.22	23M	25	120M	15M	HW T&L
1H00	GF2 GTS	.18	25M	35	200M[1]	25M	Per-Pixel Shading
2H00	GF2 Ultra	.18	25M	45	250M[1]	31M	230 Mhz DDR
1H01	GeForce3	.15	57M	80	500M[1]	30M[2]	Programmable

Essentially Moore's Law ***Cubed***.

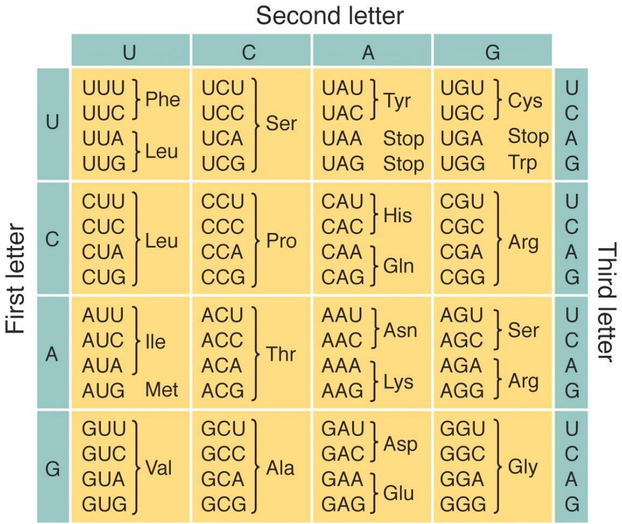
1: Dual textured
2: Programmable

Survey: some other results and work in progress

- **Symmetric tensors**
  - symmetric rank can leap arbitrarily large gap [with Comon & Mourrain]
- **Multilinear spectral theory**
  - Perron-Frobenius theorem for tensors
  - spectral hypergraph theory
- **New tensor decompositions**
  - Kronecker product decomposition
  - coclustering decomposition [with Dhillon]
- **Applications**
  - approximate simultaneous eigenvectors [with Alter & Sturmfels]
  - nonnegative tensors in algebraic statistical biology [with Sturmfels]
  - tensor decompositions for model reduction [with Pereyra]

Code of life is a $4 \times 4 \times 4$ tensor

- **Codons:** triplets of nucleotides, (i, j, k) where $i, j, k \in \{A, C, G, U\}$.
- **Genetic code:** these $4^3 = 64$ codons encode the 20 amino acids.



Second letter

First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Tensors in algebraic statistical biology

- Joint work with Bernd Sturmfels.

Problem

Find the polynomial equations that defines the set

$$\{P \in \mathbb{C}^{4\times 4\times 4} \mid \operatorname{rank}_{\otimes}(P) \leq 4\}.$$

- Why interested? Here $P = [\![p_{ijk}]\!]$ is understood to mean 'complexified' probability density values with $i, j, k \in \{A, C, G, T\}$ and we want to study tensors that are of the form

$$P = \boldsymbol{\rho}_A \otimes \boldsymbol{\sigma}_A \otimes \boldsymbol{\theta}_A + \boldsymbol{\rho}_C \otimes \boldsymbol{\sigma}_C \otimes \boldsymbol{\theta}_C + \boldsymbol{\rho}_G \otimes \boldsymbol{\sigma}_G \otimes \boldsymbol{\theta}_G + \boldsymbol{\rho}_T \otimes \boldsymbol{\sigma}_T \otimes \boldsymbol{\theta}_T,$$

  in other words,

$$p_{ijk} = \rho_{Ai}\sigma_{Aj}\theta_{Ak} + \rho_{Ci}\sigma_{Cj}\theta_{Ck} + \rho_{Gi}\sigma_{Gj}\theta_{Gk} + \rho_{Ti}\sigma_{Tj}\theta_{Tk}.$$

- Why over $\mathbb{C}$? Easier to deal with mathematically.
- Ultimately, want to study this over $\mathbb{R}_+$.

Conclusion

- Floating point computing is powerful and cheap
  - 1 million fold increase in the last 50 years,
  - potentially our best tool for analyzing massive datasets.
- Last 50 years, Numerical Linear Algebra played crucial role in:
  - statistical analysis of **two-way data**,
  - numerical solution of partial differential equations of **vector fields**,
  - numerical solution of **second-order optimization** methods.
- Next step — develop Numerical Multilinear Algebra for:
  - statistical analysis of **multi-way data**,
  - numerical solution of partial differential equations of **tensor fields**,
  - numerical solution of **higher-order optimization** methods.
- **Goal:** develop a collection of standard algorithms for higher order tensors that parallel algorithms developed for order-2 tensors.