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)**

$$A(\lambda) = -\log(l_1/l_0) = \varepsilon(\lambda)c.$$  

$A = \text{absorbance}$, $l_1/l_0 = \text{fraction of intensity of light of wavelength } \lambda \text{ that passes through specimen}$, $c = \text{concentration of chromophores}$.

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

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

- 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 $\min \|A_{m \times n} - E_{m \times r}C_{r \times n}\|$.  
  
Lek-Heng Lim (Stanford University)  
Numerical multilinear algebra in data analysis  
April 5, 2007  
3 / 33
Modern data mining

- **Text mining** is the spectroscopy of documents.
- Specimens = **documents**.
- Chromophores = **terms**.
- Absorbance = inverse document frequency:
  \[ A(t_i) = -\log \left( \frac{\sum_j \chi(f_{ij})}{n} \right). \]
- 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. \( A = QR = U\Sigma V^T \) rank-revealing factorizations.
- Bilinear models:
  - Gerald Salton et. al.: **vector space model** (QR);
  - Sue Dumais et. al.: **latent semantic 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, $A = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ may be regarded as an $m$-by-$n$ matrix, e.g. gene $\times$ microarray matrices in bioinformatics, terms $\times$ documents matrices in text mining, facial images $\times$ 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 \( [b_{ijk}] \in \mathbb{R}^{l \times m \times n}, \lambda \in \mathbb{R} \),
     \[
     [a_{ijk}] + [b_{ijk}] := [a_{ijk} + b_{ijk}] \quad \text{and} \quad \lambda [a_{ijk}] := [\lambda a_{ijk}] \in \mathbb{R}^{l \times m \times n}
     \]
  2. **Multilinear matrix multiplication:** for matrices
     \( L = [\lambda_{i'j'}] \in \mathbb{R}^{p \times l}, M = [\mu_{j'j}] \in \mathbb{R}^{q \times m}, N = [\nu_{k'k}] \in \mathbb{R}^{r \times n} \),
     \[
     (L, M, N) \cdot A := [c_{i'j'k'}] \in \mathbb{R}^{p \times q \times r}
     \]
     where
     \[
     c_{i'j'k'} := \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n} \lambda_{i'i} \mu_{j'j} \nu_{k'k} a_{ijk}.
     \]

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 \( (M, N) \cdot A = MAN^T \).
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 $u \otimes v \otimes w$,

\[
\sum \alpha u \otimes v \otimes w,
\]

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

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

\[
(\alpha u_1 + \beta u_2) \otimes v \otimes w = \alpha u_1 \otimes v \otimes w + \beta u_2 \otimes v \otimes w,
\]
\[
u \otimes (\alpha v_1 + \beta v_2) \otimes w = \alpha u \otimes v_1 \otimes w + \beta u \otimes v_2 \otimes w,
\]
\[
u \otimes v \otimes (\alpha w_1 + \beta w_2) = \alpha u \otimes v \otimes w_1 + \beta u \otimes v \otimes w_2.
\]

- Up to a choice of bases on $U$, $V$, $W$, $A \in U \otimes V \otimes W$ 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?” ≡ “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 \( A \) 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
  \[ u \otimes v \otimes w = [u_i v_j w_k]_{i,j,k=1}^{l,m,n}. \]

- A tensor $A \in \mathbb{R}^{l \times m \times n}$ is said to be decomposable if it can be written in the form
  \[ A = u \otimes v \otimes w \]
  for some $u \in \mathbb{R}^l$, $v \in \mathbb{R}^m$, $w \in \mathbb{R}^n$. For order 2, $u \otimes v = uv^T$.

- In general, any $A \in \mathbb{R}^{l \times m \times n}$ may be written as a sum of decomposable tensors
  \[ A = \sum_{i=1}^{r} \lambda_i u_i \otimes v_i \otimes w_i. \]

- 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.** $A \in \mathbb{R}^{m \times n}$
  
  \[
  \text{rank}(A) = \dim(\text{span}_{\mathbb{R}} \{ A_{\bullet 1}, \ldots, A_{\bullet n} \}) \quad \text{(column rank)}
  \]
  
  \[
  = \dim(\text{span}_{\mathbb{R}} \{ A_{1\bullet}, \ldots, A_{m\bullet} \}) \quad \text{(row rank)}
  \]
  
  \[
  = \min\{ r \mid A = \sum_{i=1}^{r} u_i v_i^T \} \quad \text{(outer product rank)}
  \]

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

  \[
  r_1(A) = \dim(\text{span}_{\mathbb{R}} \{ A_{1\bullet \bullet}, \ldots, A_{l\bullet \bullet} \})
  \]
  
  \[
  r_2(A) = \dim(\text{span}_{\mathbb{R}} \{ A_{\bullet 1 \bullet}, \ldots, A_{\bullet m \bullet} \})
  \]
  
  \[
  r_3(A) = \dim(\text{span}_{\mathbb{R}} \{ A_{\bullet \bullet 1}, \ldots, A_{\bullet \bullet n} \})
  \]

- **Outer product rank.** $A \in \mathbb{R}^{l \times m \times n}$.
  
  \[
  \text{rank}_{\otimes}(A) = \min\{ r \mid A = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i \}
  \]

- In general, \(\text{rank}_{\otimes}(A) \neq r_1(A) \neq r_2(A) \neq r_3(A)\).
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, i.e. $\text{rank}(A) = \min\{m, n\}$
5. $\text{rank}(A)$ from a non-orthogonal rank-revealing decomposition (e.g. $A = L_1 D L_2^T$) and $\text{rank}(A)$ from an orthogonal rank-revealing decomposition (e.g. $A = Q_1 R Q_2^T$) are equal
6. $\text{rank}(A)$ is base field independent, i.e. 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 $\text{rank}_\otimes(A)$ for $A \in \mathbb{R}^{l \times m \times n}$ is NP-hard [Håstad 1990]

2. For some $A \in \mathbb{R}^{l \times m \times n}$, $\text{argmin}_{\text{rank}_\otimes(B) \leq r} \|A - B\|_F$ does not have a solution

3. When $\text{argmin}_{\text{rank}_\otimes(B) \leq r} \|A - B\|_F$ 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 $\text{rank}_\otimes(A) = r$ 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 $\text{rank}_\otimes(A)$ number of terms

6. $\text{rank}_\otimes(A)$ is base field dependent, i.e. 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 \( \text{rank}_{\square}(A) \) for \( A \in \mathbb{R}^{l \times m \times n} \) is easy
2. Solution to \( \arg \min_{\text{rank}_{\square}(B) \leq (r_1, r_2, r_3)} \| A - B \|_F \) always exist
3. Solution to \( \arg \min_{\text{rank}_{\square}(B) \leq (r_1, r_2, r_3)} \| A - B \|_F \) easy to find
4. Pick \( A \in \mathbb{R}^{l \times m \times n} \) at random, then \( A \) has
   \[
   \text{rank}_{\square}(A) = (\min(l, mn), \min(m, ln), \min(n, lm))
   \]
   with probability \( 1 \)
5. If \( A \in \mathbb{R}^{l \times m \times n} \) has \( \text{rank}_{\square}(A) = (r_1, r_2, r_3) \). 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. \( \text{rank}_{\square}(A) \) 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 \( \lambda^\text{em}_j \) of \( i \)th sample excited with light at wavelength \( \lambda^\text{ex}_k \).
  - Get 3-way data \( A = [a_{ijk}] \in \mathbb{R}^{l \times m \times n} \).
  - Get outer product decomposition of \( A \)

\[
A = x_1 \otimes y_1 \otimes z_1 + \cdots + x_r \otimes y_r \otimes z_r.
\]

- Get the true chemical factors responsible for the data.
  - \( r \): number of pure substances in the mixtures,
  - \( x_\alpha = (x_{1\alpha}, \ldots, x_{l\alpha}) \): relative concentrations of \( \alpha \)th substance in specimens 1, \ldots, \( l \),
  - \( y_\alpha = (y_{1\alpha}, \ldots, y_{m\alpha}) \): excitation spectrum of \( \alpha \)th substance,
  - \( z_\alpha = (z_{1\alpha}, \ldots, z_{n\alpha}) \): emission spectrum of \( \alpha \)th substance.

- Noisy case: find best rank-\( r \) approximation (\text{CANDECOMP/PARAFAC}).
Multilinear decomposition in bioinformatics

- Application to cell cycle studies by Alter and Omberg.
- Collection of gene-by-microarray matrices $A_1, \ldots, A_l \in \mathbb{R}^{m \times n}$ obtained under varying oxidative stress.
  - $a_{ijk} =$ expression level of $j$th gene in $k$th microarray under $i$th stress.
  - Get 3-way data array $A = [a_{ijk}] \in \mathbb{R}^{l \times m \times n}$.
  - 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

\[
\text{argmin}_{\text{rank}(B) \leq r} \| A - B \|
\]

Examples

1. **Outer product rank**: \( A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \), find \( u_i, v_i, w_i \):

\[
\min \| A - u_1 \otimes v_1 \otimes w_1 - u_2 \otimes v_2 \otimes w_2 - \cdots - u_r \otimes v_r \otimes z_r \|.
\]

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**: \( A \in S^k(\mathbb{C}^n) \), find \( u_i \):

\[
\min \| A - u_1 \otimes^k - u_2 \otimes^k - \cdots - u_r \otimes^k \|.
\]

4. **Nonnegative rank**: \( 0 \leq A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \), find \( u_i \geq 0, v_i \geq 0, w_i \geq 0 \).
Feature revelation

More generally, $\mathcal{D} = \text{dictionary}$. 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**: $\mathcal{D} = \{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \text{rank}_\otimes (A) \leq 1 \}$.
2. **Tucker**: $\mathcal{D} = \{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \text{rank}_\boxtimes (A) \leq (1, 1, 1) \}$.
3. **De Lathauwer**: $\mathcal{D} = \{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \text{rank}_\boxplus (A) \leq (r_1, r_2, r_3) \}$.
4. **ICA**: $\mathcal{D} = \{ A \in S^k(\mathbb{C}^n) \mid \text{rank}_S (A) \leq 1 \}$.
5. **NTF**: $\mathcal{D} = \{ A \in \mathbb{R}_+^{d_1 \times d_2 \times d_3} \mid \text{rank}_+ (A) \leq 1 \}$.
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 $S_r(d_1, \ldots, d_k) := \{ A \mid \text{rank}_\otimes(A) \leq r \}$ is not closed.

2. There exists $B$, $\text{rank}_\otimes(B) > r$, that may be approximated arbitrarily closely by tensors of strictly lower rank, ie.

   $$\inf\{ \|B - A\| \mid \text{rank}_\otimes(A) \leq r \} = 0.$$

3. There exists $C$, $\text{rank}_\otimes(C) > r$, that does not have a best rank-$r$ approximation, ie.

   $$\inf\{ \|C - A\| \mid \text{rank}_\otimes(A) \leq r \}$$

   is not attained (by any $A$ with $\text{rank}_\otimes(A) \leq r$).
Non-existence of best low-rank approximation

Let \( x_i, y_i \in \mathbb{R}^{d_i}, i = 1, 2, 3 \). Let

\[
A := x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3
\]

and for \( n \in \mathbb{N} \),

\[
A_n := x_1 \otimes x_2 \otimes (y_3 - nx_3) + \left( x_1 + \frac{1}{n}y_1 \right) \otimes \left( x_2 + \frac{1}{n}y_2 \right) \otimes nx_3.
\]

**Lemma (de Silva and Lim)**

\( \text{rank}_\otimes(A) = 3 \) iff \( x_i, y_i \) linearly independent, \( i = 1, 2, 3 \). Furthermore, it is clear that \( \text{rank}_\otimes(A_n) \leq 2 \) and

\[
\lim_{n \to \infty} A_n = A.
\]

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**
   - all orders \( k > 2 \),
   - all norms and Brègman divergences,
   - all ranks \( r = 2, \ldots, \min\{d_1, \ldots, d_k\} \).

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 \( r + s \).
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 $C \in \{A \mid \text{rank}_\otimes(A) \leq r\}$ attaining
  \[
  \inf\{\|C - A\| \mid \text{rank}_\otimes(A) \leq r\}
  \]
  a **weak solution**. In particular, we must have \(\text{rank}_\otimes(C) > r\).

**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 \(\text{rank}_\otimes(A_n) \leq 2\) 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 \(\text{rank}_\otimes(A)\) must be exactly 3 and there exist pairs of linearly independent vectors $x_1, y_1 \in \mathbb{R}^{d_1}$, $x_2, y_2 \in \mathbb{R}^{d_2}$, $x_3, y_3 \in \mathbb{R}^{d_3}$ such that

\[
A = x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3.
\]
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

\[
\text{rank}_+(A) := \min\{ r \mid \sum_{i=1}^r u_i \otimes v_i \otimes \cdots \otimes z_i, \ u_i, \ldots, z_i \geq 0 \}
\]

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

\[
\inf \{ \| A - \sum_{i=1}^r u_i \otimes v_i \otimes \cdots \otimes z_i \| \mid u_i, \ldots, z_i \geq 0 \}
\]

is always attained.

**Corollary**

Nonnegative tensor approximation always have solutions.
Algorithms

- Even when an optimal solution $B_*$ to $\text{argmin}_{\text{rank} \leq r} \| A - B \|_F$ 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:

\begin{center}
\textbf{Algorithm: ALS for optimal rank-r approximation}
\end{center}

\begin{center}
\begin{tabular}{|l|}
\hline
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 $\rho^{(k+1)}/\rho^{(k)} > \varepsilon$;  
\begin{align*}
X^{(k+1)} &\leftarrow \text{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 \text{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 \text{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_{\alpha}^{(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;
\end{align*}
\hline
\end{tabular}
\end{center}
Convex relaxation

- Joint work with Kim-Chuan Toh.
- \( F(x_{11}, \ldots, z_{nr}) = \| A - \sum_{\alpha=1}^{r} x_\alpha \otimes y_\alpha \otimes z_\alpha \|_F^2 \) is a polynomial.

**Lasserre/Parrilo strategy:** Find largest \( \lambda^* \) such that \( F - \lambda^* \) is a sum of squares. Then \( \lambda^* \) is often \( \min F(x_{11}, \ldots, z_{nr}) \).

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

\[
F(x_{11}, \ldots, z_{nr}) - \lambda = \mathbf{v}^T(M - \lambda E_{11})\mathbf{v}
\]

for some \( M \in \mathbb{R}^{D \times D} \).

2. Note \( \text{RHS} \) is a sum of squares iff \( M - \lambda E_{11} \) is positive semi-definite (since \( M - \lambda E_{11} = B^T B \)).

3. Get convex problem

\[
\begin{align*}
\text{minimize} & \quad -\lambda \\
\text{subjected to} & \quad \mathbf{v}^T(S + \lambda E_{11})\mathbf{v} = F, \\
& \quad S \succeq 0.
\end{align*}
\]
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 \(f(x) = \sum_{i=1}^{m} p_i(x)^2\), then the powers of the monomials in \(p_i\) must lie in \(\frac{1}{2} \text{Newton}(f)\).

- So if \(f(x_{i1}, \ldots, z_{nr}) = \sum_{j=1}^{N} p_j(x_{i1}, \ldots, z_{nr})^2\), then only \(1\) and monomials of the form \(x_{i\alpha}y_{j\alpha}z_{k\alpha}\) may occur in \(p_1, \ldots, p_N\).
- Complexity is reduced to \(rlmn + 1\) from \(\binom{r(l+m+n)+3}{3}\).
Exploiting semiseparability

- Joint work with Ming Gu.

**Gauss-Newton Method:** $g(x) = \|f(x)\|^2$. Approximate Hessian using Jacobian: $H_g \approx J^T_f J_f$.

The Hessian of $F(X, Y, Z) = \|A - \sum_{\alpha=1}^{r} x_\alpha \otimes y_\alpha \otimes z_\alpha \|_F^2$ 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 \((L_1, \ldots, L_k) \cdot A\) 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)

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

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.

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>Third letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td>UG</td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td>UG</td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>UG</td>
</tr>
<tr>
<td>C</td>
<td>UUC</td>
<td>UG</td>
</tr>
<tr>
<td></td>
<td>UCC</td>
<td>UG</td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>UG</td>
</tr>
<tr>
<td></td>
<td>UCG</td>
<td>UG</td>
</tr>
<tr>
<td>A</td>
<td>CAU</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CCG</td>
<td>CG</td>
</tr>
<tr>
<td>G</td>
<td>CGU</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CGC</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>CG</td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>CG</td>
</tr>
</tbody>
</table>

- **Codons:**
  - UUU: Phe
  - UUC: Leu
  - UUA: Leu
  - UUG: Leu
  - CUA: Leu
  - CUG: Leu
  - AUA: Met
  - AUG: Met
  - GUA: Val
  - GUG: Val

- **Second letter:**
  - U: Phe, Leu
  - C: Ser, Pro
  - A: Thr
  - G: Ala

- **Third letter:**
  - U: Tyr
  - C: His, Gln
  - A: Asp
  - G: Glu
  - GGA, GGG: Gly
Tensors in algebraic statistical biology

- Joint work with Bernd Sturmfels.

Problem

*Find the polynomial equations that defines the set*

\[
\{ P \in \mathbb{C}^{4 \times 4 \times 4} \mid \text{rank}_{\otimes}(P) \leq 4 \}.
\]

- 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

\[
P = \rho_A \otimes \sigma_A \otimes \theta_A + \rho_C \otimes \sigma_C \otimes \theta_C + \rho_G \otimes \sigma_G \otimes \theta_G + \rho_T \otimes \sigma_T \otimes \theta_T,
\]

in other words,

\[
p_{ijk} = \rho_A \sigma_A \theta_A + \rho_C \sigma_C \theta_C + \rho_G \sigma_G \theta_G + \rho_T \sigma_T \theta_T.
\]

- 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.