Numerical Multilinear Algebra I

Lek-Heng Lim

University of California, Berkeley

January 5–7, 2009

Hope

Past 50 years, Numerical Linear Algebra played indispensable role in

- the statistical analysis of two-way data,
- the numerical solution of partial differential equations arising from vector fields,
- the numerical solution of second-order optimization methods.

Next step — development of Numerical Multilinear Algebra for

- the statistical analysis of multi-way data,
- the numerical solution of partial differential equations arising from tensor fields,
- the numerical solution of higher-order optimization methods.

DARPA mathematical challenge eight

One of the twenty three mathematical challenges announced at DARPA Tech 2007.

Problem

Beyond convex optimization: *can linear algebra be replaced by algebraic geometry in a systematic way?*

- **Algebraic geometry in a slogan:** polynomials are to algebraic geometry what matrices are to linear algebra.
- Polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$ of degree d can be expressed as

$$f(\mathbf{x}) = a_0 + \mathbf{a}_1^\top \mathbf{x} + \mathbf{x}^\top A_2 \mathbf{x} + \mathcal{A}_3(\mathbf{x}, \mathbf{x}, \mathbf{x}) + \cdots + \mathcal{A}_d(\mathbf{x}, \dots, \mathbf{x}).$$

$a_0 \in \mathbb{R}, \mathbf{a}_1 \in \mathbb{R}^n, A_2 \in \mathbb{R}^{n \times n}, \mathcal{A}_3 \in \mathbb{R}^{n \times n \times n}, \dots, \mathcal{A}_d \in \mathbb{R}^{n \times \cdots \times n}$.

- Numerical linear algebra: $d = 2$.
- Numerical multilinear algebra: $d > 2$.

Motivation

Why multilinear:

- "*Classification of mathematical problems as linear and nonlinear is like classification of the Universe as bananas and non-bananas.*"
- Nonlinear — too general. Multilinear — next natural step.

Why numerical:

- Different from Computer Algebra.
- Numerical rather than symbolic: floating point operations — cheap and abundant; symbolic operations — expensive.
- Like other areas in numerical analysis, will entail the approximate solution of approximate multilinear problems with approximate data but under controllable and rigorous confidence bounds on the errors involved.

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-**hypermatrix** $A = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$.

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.

Tensors: data analyst'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 **hypermatrix**. $(L, M, N)\cdot A$ as multiplication on '3 sides' by matrices L, M, N.
- Generalizes to arbitrary order k. If $k = 2$, ie. matrix, then $(M, N)\cdot A = MAN^{\mathsf{T}}$.

Hypermatrices

Totally ordered finite sets: $[n] = \{1 < 2 < \cdots < n\}$, $n \in \mathbb{N}$.

- Vector or n-tuple
$$f : [n] \to \mathbb{R}.$$
If $f(i) = a_i$, then f is represented by $\mathbf{a} = [a_1, \dots, a_n]^\top \in \mathbb{R}^n$.
- Matrix
$$f : [m] \times [n] \to \mathbb{R}.$$
If $f(i,j) = a_{ij}$, then f is represented by $A = [a_{ij}]_{i,j=1}^{m,n} \in \mathbb{R}^{m \times n}$.
- Hypermatrix (order 3)
$$f : [l] \times [m] \times [n] \to \mathbb{R}.$$
If $f(i,j,k) = a_{ijk}$, then f is represented by $\mathcal{A} = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$.

Normally $\mathbb{R}^X = \{f : X \to \mathbb{R}\}$. Ought to be $\mathbb{R}^{[n]}, \mathbb{R}^{[m]\times[n]}, \mathbb{R}^{[l]\times[m]\times[n]}$.

Hypermatrices and tensors

Up to choice of bases

- $\mathbf{a} \in \mathbb{R}^n$ can represent a vector in V (contravariant) or a linear functional in V^* (covariant).
- $A \in \mathbb{R}^{m \times n}$ can represent a bilinear form $V^* \times W^* \to \mathbb{R}$ (contravariant), a bilinear form $V \times W \to \mathbb{R}$ (covariant), or a linear operator $V \to W$ (mixed).
- $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ can represent trilinear form $U \times V \times W \to \mathbb{R}$ (covariant), bilinear operators $V \times W \to U$ (mixed), etc.

A hypermatrix is the same as a tensor if

1. we give it coordinates (represent with respect to some bases);
2. we ignore covariance and contravariance.

Basic operation on a hypermatrix

- A matrix can be multiplied on the left and right: $A \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{p \times m}$, $Y \in \mathbb{R}^{q \times n}$,

$$(X, Y) \cdot A = XAY^\top = [c_{\alpha\beta}] \in \mathbb{R}^{p \times q}$$

  where

$$c_{\alpha\beta} = \sum\nolimits_{i,j=1}^{m,n} x_{\alpha i} y_{\beta j} a_{ij}.$$

- A hypermatrix can be multiplied on three sides: $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$, $X \in \mathbb{R}^{p \times l}$, $Y \in \mathbb{R}^{q \times m}$, $Z \in \mathbb{R}^{r \times n}$,

$$(X, Y, Z) \cdot \mathcal{A} = [\![c_{\alpha\beta\gamma}]\!] \in \mathbb{R}^{p \times q \times r}$$

  where

$$c_{\alpha\beta\gamma} = \sum\nolimits_{i,j,k=1}^{l,m,n} x_{\alpha i} y_{\beta j} z_{\gamma k} a_{ijk}.$$

Basic operation on a hypermatrix

- Covariant version:

$$\mathcal{A} \cdot (X^\top, Y^\top, Z^\top) := (X, Y, Z) \cdot \mathcal{A}.$$

- Gives convenient notations for multilinear functionals and multilinear operators. For $\mathbf{x} \in \mathbb{R}^l, \mathbf{y} \in \mathbb{R}^m, \mathbf{z} \in \mathbb{R}^n$,

$$\mathcal{A}(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathcal{A} \cdot (\mathbf{x}, \mathbf{y}, \mathbf{z}) = \sum_{i,j,k=1}^{l,m,n} a_{ijk} x_i y_j z_k,$$

$$\mathcal{A}(I, \mathbf{y}, \mathbf{z}) := \mathcal{A} \cdot (I, \mathbf{y}, \mathbf{z}) = \sum_{j,k=1}^{m,n} a_{ijk} y_j z_k.$$

Segre 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 $\mathbf{u} \in \mathbb{R}^l, \mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$.

The set of all decomposable tensors is known as the **Segre variety** in algebraic geometry. It is a closed set (in both the Euclidean and Zariski sense) as it can be described algebraically:

$$\operatorname{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n) = \{A \in \mathbb{R}^{l \times m \times n} \mid a_{i_1 i_2 i_3} a_{j_1 j_2 j_3} = a_{k_1 k_2 k_3} a_{l_1 l_2 l_3}, \{i_\alpha, j_\alpha\} = \{k_\alpha, l_\alpha\}\}$$

Symmetric hypermatrices

- Cubical hypermatrix $[\![a_{ijk}]\!] \in \mathbb{R}^{n\times n\times n}$ is **symmetric** if

$$a_{ijk} = a_{ikj} = a_{jik} = a_{jki} = a_{kij} = a_{kji}.$$

- Invariant under all permutations $\sigma \in \mathfrak{S}_k$ on indices.
- $\mathsf{S}^k(\mathbb{R}^n)$ denotes set of all order-k symmetric hypermatrices.

Example

Higher order derivatives of multivariate functions.

Example

Moments of a random vector $\mathbf{x} = (X_1, \dots, X_n)$:

$$m_k(\mathbf{x}) = \left[E(x_{i_1} x_{i_2} \cdots x_{i_k})\right]_{i_1,\dots,i_k=1}^{n} = \left[\int \cdots \int x_{i_1} x_{i_2} \cdots x_{i_k} \; d\mu(x_{i_1}) \cdots d\mu(x_{i_k})\right]_{i_1,\dots,i_k=1}^{n}.$$

Symmetric hypermatrices

Example

Cumulants of a random vector $\mathbf{x} = (X_1, \dots, X_n)$:

$$\kappa_k(\mathbf{x}) = \left[\sum_{A_1 \sqcup \cdots \sqcup A_p = \{i_1, \dots, i_k\}} (-1)^{p-1}(p-1)! E\left(\prod_{i \in A_1} x_i\right) \cdots E\left(\prod_{i \in A_p} x_i\right) \right]_{i_1, \dots, i_k = 1}^{n}.$$

For $n = 1$, $\kappa_k(x)$ for $k = 1, 2, 3, 4$ are the expectation, variance, skewness, and kurtosis.

- Important in Independent Component Analysis (ICA).

Inner products and norms

- $\ell^2([n])$: $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b} = \sum_{i=1}^n a_i b_i$.
- $\ell^2([m] \times [n])$: $A, B \in \mathbb{R}^{m \times n}$, $\langle A, B \rangle = \operatorname{tr}(A^\top B) = \sum_{i,j=1}^{m,n} a_{ij} b_{ij}$.
- $\ell^2([l] \times [m] \times [n])$: $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{l \times m \times n}$, $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i,j,k=1}^{l,m,n} a_{ijk} b_{ijk}$.
- In general,

$$\ell^2([m] \times [n]) = \ell^2([m]) \otimes \ell^2([n]),$$
$$\ell^2([l] \times [m] \times [n]) = \ell^2([l]) \otimes \ell^2([m]) \otimes \ell^2([n]).$$

- Frobenius norm

$$\|\mathcal{A}\|_F^2 = \sum_{i,j,k=1}^{l,m,n} a_{ijk}^2.$$

- Norm topology often more directly relevant to engineering applications than Zariski toplogy.

Other norms

- Let $\|\cdot\|_{\alpha_i}$ be a norm on $\mathbb{R}^{d_i}$, $i = 1, \dots, k$. Then **operator norm** of multilinear functional $A : \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_k} \to \mathbb{R}$ is
$$\|A\|_{\alpha_1,\dots,\alpha_k} := \sup \frac{|A(\mathbf{x}_1, \dots, \mathbf{x}_k)|}{\|\mathbf{x}_1\|_{\alpha_1} \cdots \|\mathbf{x}_k\|_{\alpha_k}}.$$
- Deep and important results about such norms in functional analysis.
- *E-norm* and *G-norm*:
$$\|A\|_E = \sum\nolimits_{i_1,\dots,i_k=1}^{d_1,\dots,d_k} |a_{j_1\cdots j_k}|$$
and
$$\|A\|_G = \max\{|a_{j_1\cdots j_k}| \mid j_1 = 1, \dots, d_1; \dots ; j_k = 1, \dots, d_k\}.$$
- Multiplicative on rank-1 tensors:
$$\|\mathbf{u} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{z}\|_E = \|\mathbf{u}\|_1 \|\mathbf{v}\|_1 \cdots \|\mathbf{z}\|_1,$$
$$\|\mathbf{u} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{z}\|_F = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \cdots \|\mathbf{z}\|_2,$$
$$\|\mathbf{u} \otimes \mathbf{v} \otimes \cdots \otimes \mathbf{z}\|_G = \|\mathbf{u}\|_\infty \|\mathbf{v}\|_\infty \cdots \|\mathbf{z}\|_\infty.$$

Tensor ranks (Hitchcock, 1927)

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

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

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

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

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

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

where $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} := [\![u_i v_j w_k]\!]_{i,j,k=1}^{l,m,n}$.

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 $A \in \mathbb{C}^{l \times m \times n}$

Algebraic computational complexity

- For $A = (a_{ij}), B = (b_{jk}) \in \mathbb{R}^{n\times n}$,

$$AB = \sum\nolimits_{i,j,k=1}^{n} a_{ik} b_{kj} E_{ij} = \sum\nolimits_{i,j,k=1}^{n} \varphi_{ik}(A)\varphi_{kj}(B)E_{ij}$$

  where $E_{ij} = \mathbf{e}_i\mathbf{e}_j^\top \in \mathbb{R}^{n\times n}$. Let

$$T = \sum\nolimits_{i,j,k=1}^{n} \varphi_{ik} \otimes \varphi_{kj} \otimes E_{ij}.$$

- $O(n^{2+\varepsilon})$ algorithm for multiplying two $n \times n$ matrices gives $O(n^{2+\varepsilon})$ algorithm for solving system of n linear equations [Strassen 1969].
- **Conjecture.** $\log_2(\operatorname{rank}_\otimes(T)) \le 2 + \varepsilon$.
- **Best known result.** $O(n^{2.376})$ [Coppersmith-Winograd 1987; Cohn-Kleinberg-Szegedy-Umans 2005].

More tensor ranks

- For $\mathbf{u} \in \mathbb{R}^l, \mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$,

$$\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} := [\![u_i v_j w_k]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}.$$

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

$$\operatorname{rank}_{\otimes}(\mathcal{A}) = \min\{r \mid \mathcal{A} = \textstyle\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i, \quad \sigma_i \in \mathbb{R}\}.$$

- **Symmetric outer product rank.** $\mathcal{A} \in \mathsf{S}^k(\mathbb{R}^n)$,

$$\operatorname{rank}_{\mathsf{S}}(\mathcal{A}) = \min\{r \mid \mathcal{A} = \textstyle\sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i, \quad \lambda_i \in \mathbb{R}\}.$$

- **Nonnegative outer product rank.** $\mathcal{A} \in \mathbb{R}_+^{l \times m \times n}$,

$$\operatorname{rank}_{+}(\mathcal{A}) = \min\{r \mid \mathcal{A} = \textstyle\sum_{i=1}^{r} \delta_i \mathbf{x}_i \otimes \mathbf{y}_i \otimes \mathbf{z}_i, \quad \delta_i \in \mathbb{R}_+\}.$$

SVD, EVD, NMF of a matrix

- **Singular value decomposition** of $A \in \mathbb{R}^{m \times n}$,

$$A = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i$$

where $\operatorname{rank}(\mathcal{A}) = r$, $U \in \mathrm{O}(m)$ left singular vectors, $V \in \mathrm{O}(n)$ right singular vectors, Σ singular values.

- **Symmetric eigenvalue decomposition** of $A \in \mathsf{S}^2(\mathbb{R}^n)$,

$$A = V\Lambda V^\top = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i,$$

where $\operatorname{rank}(A) = r$, $V \in \mathrm{O}(n)$ eigenvectors, Λ eigenvalues.

- **Nonnegative matrix factorization** of $A \in \mathbb{R}_+^{n \times n}$,

$$A = X\Delta Y^\top = \sum_{i=1}^{r} \delta_i \mathbf{x}_i \otimes \mathbf{y}_i$$

where $\operatorname{rank}_+(A) = r$, $X, Y \in \mathbb{R}_+^{m \times r}$ unit column vectors (in the 1-norm), Δ positive values.

SVD, EVD, NMF of a hypermatrix

- **Outer product decomposition** of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

$$\mathcal{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i$$

where $\operatorname{rank}_{\otimes}(\mathcal{A}) = r$, $\mathbf{u}_i \in \mathbb{R}^l, \mathbf{v}_i \in \mathbb{R}^m, \mathbf{w}_i \in \mathbb{R}^n$ unit vectors, $\sigma_i \in \mathbb{R}$.

- **Symmetric outer product decomposition** of $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$,

$$\mathcal{A} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i$$

where $\operatorname{rank}_{\mathsf{S}}(A) = r$, $\mathbf{v}_i$ unit vector, $\lambda_i \in \mathbb{R}$.

- **Nonnegative outer product decomposition** for hypermatrix $\mathcal{A} \in \mathbb{R}_+^{l \times m \times n}$ is

$$\mathcal{A} = \sum_{i=1}^{r} \delta_i \mathbf{x}_i \otimes \mathbf{y}_i \otimes \mathbf{z}_i$$

where $\operatorname{rank}_{+}(A) = r$, $\mathbf{x}_i \in \mathbb{R}_+^l, \mathbf{y}_i \in \mathbb{R}_+^m, \mathbf{z}_i \in \mathbb{R}_+^n$ unit vectors, $\delta_i \in \mathbb{R}_+$.

Best low rank approximation of a matrix

- Given $A \in \mathbb{R}^{m \times n}$. Want

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

- More precisely, find $\sigma_i, \mathbf{u}_i, \mathbf{v}_i$, $i = 1, \dots, r$, that minimizes

$$\|\mathcal{A} - \sigma_1 \mathbf{u}_1 \otimes \mathbf{v}_1 - \sigma_2 \mathbf{u}_2 \otimes \mathbf{v}_2 - \cdots - \sigma_r \mathbf{u}_r \otimes \mathbf{v}_r\|.$$

Theorem (Eckart–Young)

Let $A = U\Sigma V^\top = \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ be singular value decomposition. For $r \leq \operatorname{rank}(A)$, let

$$A_r := \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.$$

Then

$$\|A - A_r\|_F = \min_{\operatorname{rank}(B) \leq r} \|A - B\|_F.$$

- No such thing for hypermatrices of order 3 or higher.

Segre variety and its secant varieties

- The set of all rank-1 hypermatrices is known as the Segre variety in algebraic geometry.
- It is a closed set (in both the Euclidean and Zariski sense) as it can be described algebraically:

$$\operatorname{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n) = \{\mathcal{A} \in \mathbb{R}^{l\times m\times n} \mid \mathcal{A} = \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}\} =$$
$$\{\mathcal{A} \in \mathbb{R}^{l\times m\times n} \mid a_{i_1 i_2 i_3} a_{j_1 j_2 j_3} = a_{k_1 k_2 k_3} a_{l_1 l_2 l_3}, \{i_\alpha, j_\alpha\} = \{k_\alpha, l_\alpha\}\}$$

- Hypermatrices that have rank > 1 are elements on the higher secant varieties of $\mathscr{S} = \operatorname{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n)$.
- E.g. a hypermatrix has rank 2 if it sits on a secant line through two points in $\mathscr{S}$ but not on $\mathscr{S}$, rank 3 if it sits on a secant plane through three points in $\mathscr{S}$ but not on any secant lines, etc.
- Minor technicality: should really be secant *quasiprojective variety*.

Scientific data mining

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

$$A(\lambda_i, s_j) = \sum\nolimits_{f=1}^{r} \varepsilon_f(\lambda_i) c_f(s_j).$$

- 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 model aka **vector space model**.
- Due to Gerald Salton and colleagues: SMART (system for the mechanical analysis and retrieval of text).

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

Fact (Inevitable consequence of technological advancement)

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

Fundamental problem of multiway data analysis

- $\mathcal{A}$ hypermatrix, symmetric hypermatrix, or nonnegative hypermatrix.
- Solve

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

- rank may be outer product rank, multilinear rank, symmetric rank (for symmetric hypermatrix), or nonnegative rank (nonnegative hypermatrix).

Example

Given $\mathcal{A}\in\mathbb{R}^{d_1\times d_2\times d_3}$, find $\mathbf{u}_i,\mathbf{v}_i,\mathbf{w}_i$, $i=1,\dots,r$, that minimizes

$$\|\mathcal{A}-\mathbf{u}_1\otimes\mathbf{v}_1\otimes\mathbf{w}_1-\mathbf{u}_2\otimes\mathbf{v}_2\otimes\mathbf{w}_2-\cdots-\mathbf{u}_r\otimes\mathbf{v}_r\otimes\mathbf{z}_r\|$$

or $\mathcal{C}\in\mathbb{R}^{r_1\times r_2\times r_3}$ and $U\in\mathbb{R}^{d_1\times r_1}, V\in\mathbb{R}^{d_2\times r_2}, W\in\mathbb{R}^{d_3\times r_3}$, that minimizes

$$\|\mathcal{A}-(U,V,W)\cdot\mathcal{C}\|.$$

Fundamental problem of multiway data analysis

Example

Given $\mathcal{A} \in \mathsf{S}^k(\mathbb{C}^n)$, find $\mathbf{u}_i$, $i = 1, \dots, r$, that minimizes

$$\|\mathcal{A} - \mathbf{u}_1^{\otimes k} - \mathbf{u}_2^{\otimes k} - \cdots - \mathbf{u}_r^{\otimes k}\|$$

or $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and $U \in \mathbb{R}^{n \times r_i}$ that minimizes

$$\|\mathcal{A} - (U, U, U) \cdot \mathcal{C}\|.$$

Outer product decomposition in spectroscopy

- Application to fluorescence spectral analysis by [Bro; 1997].
- 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 $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
  - Get outer product decomposition of $\mathcal{A}$

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

- 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).

Uniqueness of tensor decompositions

- $M \in \mathbb{R}^{m\times n}$, $\operatorname{spark}(M)$ = size of minimal linearly dependent subset of column vectors [Donoho, Elad; 2003].

Theorem (Kruskal)

$X = [\mathbf{x}_1, \dots, \mathbf{x}_r]$, $Y = [\mathbf{y}_1, \dots, \mathbf{y}_r]$, $Z = [\mathbf{z}_1, \dots, \mathbf{z}_r]$. *Decomposition is unique up to scaling if*

$$\operatorname{spark}(X) + \operatorname{spark}(Y) + \operatorname{spark}(Z) \geq 2r + 5.$$

- May be generalized to arbitrary order [Sidiroupoulos, Bro; 2000].
- Avoids factor indeterminacy under mild conditions.

Multilinear decomposition in bioinformatics

- Application to cell cycle studies [Omberg, Golub, Alter; 2008].
- 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 $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
  - Get multilinear decomposition of $\mathcal{A}$

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

  to get orthogonal matrices X, Y, Z and core tensor $\mathcal{C}$ 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; $\mathcal{C}$ governs interactions between these factors.
- Noisy case: approximate by discarding small c_{ijk} (Tucker Model).

Code of life is a 3-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, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA: Stop UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Tensors in algebraic statistical biology

Problem (Salmon conjecture)

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 $\mathbb{R}_+$.