Low-rank approximation of tensors and the statistical analysis of multi-indexed data

Lek-Heng Lim
Institute for Computational and Mathematical Engineering
Stanford University

NERSC Scientific Computing Seminars
Berkeley, CA
March 11, 2005

Thanks: Rasmus Bro, Gunnar Carlsson, Kim-Chuan Toh,
NSF DMS 01-01364

Collaborators

- **Vin de Silva**, Department of Mathematics, Stanford University (Manifold Learning)
- **Pierre Comon**, Laboratoire I3S, University of Nice Sophia-Antipolis (Independent Component Analysis)
- **Gene Golub**, Department of Computer Science, Stanford University (Numerical Linear Algebra)

Long Term Goal

Numerical Multilinear Algebra: Theory, Algorithms and Applications of Tensor Computations

- Develop a collection of standard computational methods for higher order tensors that parallel the methods that have been developed for order-2 tensors, ie. matrices.
- Develop the mathematical foundations to facilitate this goal.
- Applications (defer till three slides later).

Novelty

- Different from Computer Algebra
  - Interested in a numerical approach where inexpensive floating point operations, rather than expensive symbolic operations, play the central role.
  - Like other areas in numerical analysis, Numerical Multilinear Algebra will entail approximate solution of approximate multilinear problems with approximate data but under controllable error bounds.
- Different from Numerical Polynomial Algebra
  - Interested in systems of multilinear equations instead of polynomial equations.
  - Just as matrices are the main objects of interest in Numerical Linear Algebra, high order tensors will be the main

objects of interest in Numerical Multilinear Algebra. Multilinear systems of equations associated with tensors will play an auxiliary role, unlike Numerical Polynomial Algebra where systems of polynomial equations are the main objects of study.

- Different from Multilinear Algebra
  - Study of multilinear algebra in mathematics is concerned with algebraic properties of tensor products of modules or vector spaces. There is no interest in the mathematical properties of tensors per se — ie. notions such as ranks, decompositions, hyperdeterminants of a tensor — only the algebraic structure of the set of tensors as a vector space or module.
  - Will be interested in computations.

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 Now?

- Unavoidable in analyzing complex data arising from technological advancement in instruments and methodologies:
  - measurements from spectrophotometric fluorescence detector in high-performance chromatography — 3-way data
  - absorption spectra measurements at different wavelengths in a kinetic experiment — 4-way data
  - facial image database of human faces photographed under varying conditions of illumination, camera angle and facial expression — 5-way data
- Recent developments provide the right tools:
  - Algorithms: Semi-definite Programming (SDP)
  - Computing Technologies: better, cheaper
  - Theory: Algebraic Geometry, Invariant Theory

Motivation

Past 50 years, Numerical Linear Algebra played crucial 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 — develop 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.

Tensors

Tensor of order k and size $(d_1, \dots, d_k)$ is a k-way array of real numbers $A = [\![a_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ with two properties:

1. Vector Space Structure: $A = [\![a_{j_1 \cdots j_k}]\!], B = [\![b_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, $\alpha, \beta \in \mathbb{R}$,

$$\alpha A + \beta B := [\![\alpha a_{j_1 \cdots j_k} + \beta b_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}.$$

   ie. $\mathbb{R}^{d_1 \times \cdots \times d_k}$ is a vector space of dimension $d_1 \cdots d_k$ over $\mathbb{R}$.

2. Multilinear Structure: $A = [\![a_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ and matrices

$$L_1 = [\ell^1_{i_1 j_1}] \in \mathbb{R}^{r_1 \times d_1}, \quad \dots \quad , L_k = [\ell^k_{i_k j_k}] \in \mathbb{R}^{r_k \times d_k}.$$

   Then $(L_1, \dots, L_k)A = [\![c_{i_1 \cdots i_k}]\!] \in \mathbb{R}^{r_1 \times \cdots \times r_k}$ where

$$c_{i_1 \cdots i_k} = \sum_{j_1=1}^{d_1} \cdots \sum_{j_k=1}^{d_k} \ell^1_{i_1 j_1} \cdots \ell^k_{i_k j_k} a_{j_1 \cdots j_k}.$$

Property 2 distinguishes $\mathbb{R}^{d_1\times\cdots\times d_k}$ from being simply a vector space of dimension $d_1\cdots d_k$. It is the reason why, for instance, $\mathbb{R}^{l\times m\times n}$ (order-3 tensors) is different from $\mathbb{R}^{lm\times n}$ (matrices) or $\mathbb{R}^{lmn}$ (vectors).

Example. For $A\in\mathbb{R}^{m\times n}$, $(L_1,L_2)A$ is equivalent to multiplying every column vector of A by L_1 and then every row vector of the result by L_2 (or vice versa):

$$(L_1,L_2)A=L_1AL_2^t=L_1(AL_2^t)=(L_1A)L_2^t.$$

Caution: What physicists and geometers call tensors are really tensor fields (ie. tensor-valued functions on manifolds). E.g. stress tensor, moment-of-intertia tensor, Einstein tensor, metric tensor, curvature tensor, Ricci tensor, etc.

Properties of Multilinear Matrix Multiplication

- Let $A, B \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ and $\lambda, \mu \in \mathbb{R}$. Let $L_1 \in \mathbb{R}^{r_1 \times d_1}, \dots, L_k \in \mathbb{R}^{r_k \times d_k}$. Then

$$(L_1, \dots, L_k)(\lambda A + \mu B) = \lambda (L_1, \dots, L_k) A + \mu (L_1, \dots, L_k) B.$$

- Let $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. Let $L_1 \in \mathbb{R}^{r_1 \times d_1}, \dots, L_k \in \mathbb{R}^{r_k \times d_k}$, and $M_1 \in \mathbb{R}^{s_1 \times r_1}, \dots, M_k \in \mathbb{R}^{s_k \times r_k}$. Then

$$(M_1, \dots, M_k)(L_1, \dots, L_k) A = (M_1 L_1, \dots, M_k L_k) A$$

where $M_i L_i \in \mathbb{R}^{s_i \times d_i}$ is simply the matrix-matrix product of M_i and L_i.

- Let $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ and $\lambda, \mu \in \mathbb{R}$. Let $L_1 \in \mathbb{R}^{r_1 \times d_1}, \dots, L_j, M_j \in \mathbb{R}^{r_j \times d_j}, \dots, L_k \in \mathbb{R}^{r_k \times d_k}$. Then

$$(L_1, \dots, \lambda L_j + \mu M_j, \dots, L_k) A = \\ \lambda (L_1, \dots, L_j, \dots, L_k) A + \mu (L_1, \dots, M_j, \dots, L_k) A.$$

Outer Product

The outer product of k vectors, $\mathbf{x}^1 = (x_1^1, \dots, x_{d_1}^1)^t \in \mathbb{R}^{d_1}, \dots, \mathbf{x}^k = (x_1^k, \dots, x_{d_k}^k)^t \in \mathbb{R}^{d_k}$, is defined by

$$\mathbf{x}^1 \otimes \cdots \otimes \mathbf{x}^k := [\![x_{i_1}^1 \dots x_{i_k}^k]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}.$$

The outer product of k vector spaces, $\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_k}$, is simply

$$\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k} := \operatorname{span}_{\mathbb{R}}\{\mathbf{x}^1 \otimes \cdots \otimes \mathbf{x}^k \mid \mathbf{x}^1 \in \mathbb{R}^{d_1}, \dots, \mathbf{x}^k \in \mathbb{R}^{d_k}\}.$$

By definition, $\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k}$ is a subspace of the vector space $\mathbb{R}^{d_1 \times \cdots \times d_k}$. Counting dimensions, we see immediately that

$$\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k} = \mathbb{R}^{d_1 \times \cdots \times d_k}.$$

Property 2': Outer Product Structure

The fact that $\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k} = \mathbb{R}^{d_1 \times \cdots \times d_k}$ tells us that every $A \in \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k}$ may be written as

$$A = \sum_{\alpha=1}^{r} \mathbf{x}_\alpha^1 \otimes \cdots \otimes \mathbf{x}_\alpha^k$$

for some $\mathbf{x}_\alpha^j \in \mathbb{R}^{d_j}$ $(\alpha = 1, \ldots, r;\ j = 1, \ldots, k)$.

This is exactly what gives a tensor its multilinear structure. Given $L_1 \in \mathbb{R}^{r_1 \times d_1}, \ldots, L_k \in \mathbb{R}^{r_k \times d_k}$,

$$(L_1, \ldots, L_k)A = \sum_{\alpha=1}^{r} L_1\mathbf{x}_\alpha^1 \otimes \cdots \otimes L_k\mathbf{x}_\alpha^k.$$

So the multilinear structure (Property 2) and outer product structure (Property 2') are one and the same thing. We could have instead defined a tensor as one that satisfies Properties 1 and 2' — a k-array that can be decomposed into a sum of outer products of k vectors.

Multiway Data

- Psychometrics: *individual* × *variable* × *time*
- Sensory analysis: *sample* × *attribute* × *judge*
- Batch data: *batch* × *time* × *variable*
- Time-series analysis: *time* × *variable* × *lag*
- Facial image: *people* × *view* × *illumination* × *expression* × *pixels*
- Analytical chemistry: *sample* × *elution time* × *wavelength*
- Spectral data: *sample* × *emission* × *excitation* × *decay*
- Atmospheric science: *location* × *variable* × *time* × *observation*

New Mathematics for Data Analysis

- Gunnar Carlsson, Persi Diaconis, Joshua Tenenbaum: Topological Methods in Data Analysis
- Ronald Coifman: Harmonic Analysis on Data Sets
- David Donoho: High-Dimensional Data Analysis: the curses and blessings of dimensionality
- Peter Jones: The Traveling Salesman Meets Large Data Sets
- Tomasso Poggio, Steve Smale: The Mathematics of Learning: dealing with data

Treat data as points (vectors) in either subsets of $\mathbb{R}^n$ or $\mathbb{C}^n$ possibly with additional structures (e.g. Riemannian manifolds, symmetric spaces, rectifiable sets).

Problem: The intrinsic multilinear structure in many types of data is simply discarded.

Application: Analytical Chemistry

Fluorescence spectra measurements performed by Rasmus Bro.

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 $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l\times m\times n}$.

Decomposing A into a sum of outer products (a unique decomposition under mild assumptions),

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

The vectors $\mathbf{x}_\alpha, \mathbf{y}_\alpha, \mathbf{z}_\alpha$, $\alpha = 1, \dots, r$, yield the true chemical factors responsible for the data:

- there are r pure substances in the mixtures,
- $\mathbf{x}_\alpha = (x_{1\alpha}, \dots, x_{l\alpha})$ holds the relative concentrations of the αth substance in samples $1, \dots, l$,

- $\mathbf{y}_\alpha = (y_{1\alpha}, \ldots, y_{m\alpha})$ holds the excitation spectrum of the αth substance,
- $\mathbf{z}_\alpha = (z_{1\alpha}, \ldots, z_{n\alpha})$ holds the emission spectrum of the αth substance.

Recovering such information will not be possible if we 'forget' the trilinear structure, ie. vectorize A into a vector $\mathbf{a} \in \mathbb{R}^{lmn}$, and work only with $\mathbf{a}$ with no information on its original 3-way structure.

Matrix Rank

$A \in \mathbb{R}^{m \times n}$. rank$(A)$ may be defined in either one of the three (among other) ways:

- outer-product rank: rank$(A) = r$ iff there exists $\mathbf{x}_1, \dots, \mathbf{x}_r \in \mathbb{R}^m$, $\mathbf{y}_1, \dots, \mathbf{y}_r \in \mathbb{R}^n$ such that

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

  and r is minimal over all such decompositions.

- row rank: rank$(A) = r$ iff

$$\operatorname{dim}(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet}, \dots, A_{m\bullet}\}) = r$$

  where $A_{i\bullet} \in \mathbb{R}^n$ denotes the ith row vector of A.

- column rank: rank$(A) = r$ iff

$$\operatorname{dim}(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1}, \dots, A_{\bullet n}\}) = r$$

  where $A_{\bullet j} \in \mathbb{R}^m$ denotes the jth column vector of A.

Tensor Rank

$A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. Different notions of tensor ranks:

- outer product rank: $\operatorname{rank}_{\otimes}(A) = r$ iff there exists $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$, $j = 1, \ldots, k$, such that
$$A = \sum_{i=1}^{r} \mathbf{x}_i^1 \otimes \cdots \otimes \mathbf{x}_i^k$$
and r is minimal over all such decompositions.
- p-slab rank ($p = 1, \ldots, k$): $\operatorname{rank}_p(A) = r_p$ iff
$$\operatorname{dim}(\operatorname{span}_{\mathbb{R}}\{A_{\bullet\cdots\bullet 1\bullet\cdots\bullet}, \ldots, A_{\bullet\cdots\bullet d_p \bullet\cdots\bullet}\}) = r_p$$
where $A_{\bullet\cdots\bullet i \bullet\cdots\bullet} \in \mathbb{R}^{d_1 \times \cdots \times \widehat{d_p} \times \cdots \times d_k}$ denotes the ith p-slab of A, an order-$(k-1)$ tensor.
- multilinear rank of A is defined as
$$\operatorname{rank}_{\boxplus}(A) = (\operatorname{rank}_1(A), \ldots, \operatorname{rank}_k(A))$$

Note: when we wish to emphasize the dependence of multilinear rank on the order k, we will use the term k-linear rank instead. For example, when $k = 2$, then 1-slab = row, 2-slab = column. Bilinear rank of a matrix $A \in \mathbb{R}^{m \times n}$ is simply

$$\operatorname{rank}_{\boxplus}(A) = (\operatorname{rowrank}(A), \operatorname{colrank}(A)) = (\operatorname{rank}(A), \operatorname{rank}(A)).$$

When $k \geq 3$, $\operatorname{rank}_p(A) \neq \operatorname{rank}_q(A) \neq \operatorname{rank}_{\otimes}(A)$ in general (for $p \neq q$).

Useful notation: a hat over an index like $\widehat{d_p}$ means that the index is to be omitted, eg. $l \times \widehat{m} \times n = l \times n$, $(i, j, \widehat{k}, l) = (i, j, l)$.

Example

For an order-3 tensor $A \in \mathbb{R}^{l \times m \times n}$, we have

- outer product rank: $\operatorname{rank}_{\otimes}(A) = r$ iff there exists $\mathbf{x}_1, \dots, \mathbf{x}_r \in \mathbb{R}^l$, $\mathbf{y}_1, \dots, \mathbf{y}_r \in \mathbb{R}^m$, $\mathbf{y}_1, \dots, \mathbf{y}_r \in \mathbb{R}^n$ such that

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

  and r is minimal over all such decompositions.

- 1-slab rank: $\operatorname{rank}_1(A) = r_1$ iff

  $$\dim(\operatorname{span}_{\mathbb{R}}\{A_{1\bullet\bullet}, \dots, A_{l\bullet\bullet}\}) = r_1$$

  where $A_{i\bullet\bullet} \in \mathbb{R}^{m \times n}$ denotes the ith 1-slab of A.

- 2-slab rank: $\operatorname{rank}_2(A) = r_2$ iff

  $$\dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet 1 \bullet}, \dots, A_{\bullet m \bullet}\}) = r_2$$

  where $A_{\bullet j \bullet} \in \mathbb{R}^{l \times n}$ denotes the jth 2-slab of A.

- 3-slab rank: $\operatorname{rank}_3(A) = r_3$ iff

$$\dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet\bullet 1}, \dots, A_{\bullet\bullet n}\}) = r_3$$

  where $A_{\bullet\bullet k} \in \mathbb{R}^{l\times m}$ denotes the kth 3-slab of A.

- trilinear rank: $\operatorname{rank}_{\boxplus}(A) = (r_1, r_2, r_3)$.

Outer Product Decomposition

Let $A \in \mathbb{R}^{l\times m\times n}$ and $\operatorname{rank}_{\otimes}(A) = r$. The outer product or Candecomp/Parafac decomposition of A is

$$A = \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha.$$

In other words,

$$a_{ijk} = \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha}$$

for some $\mathbf{x}_\alpha = (x_{1\alpha}, \dots, x_{l\alpha})^t \in \mathbb{R}^l$, $\mathbf{y}_\alpha = (y_{1\alpha}, \dots, y_{m\alpha})^t \in \mathbb{R}^m$, $\mathbf{z}_\alpha = (z_{1\alpha}, \dots, z_{n\alpha})^t \in \mathbb{R}^n$, $\alpha = 1, \dots, r$.

The vectors $\mathbf{x}_\alpha, \mathbf{y}_\alpha, \mathbf{z}_\alpha$ are sometimes regarded as column vectors of matrices $X = [\mathbf{x}_1, \dots, \mathbf{x}_r] \in \mathbb{R}^{l\times r}$, $Y = [\mathbf{y}_1, \dots, \mathbf{y}_r] \in \mathbb{R}^{m\times r}$, $Z = [\mathbf{z}_1, \dots, \mathbf{z}_r] \in \mathbb{R}^{n\times r}$.

Multilinear Decomposition

Let $A \in \mathbb{R}^{l\times m\times n}$ and $\operatorname{rank}_{\boxplus}(A) = (r_1, r_2, r_3)$. Multilinear or Tucker decomposition of A is

$$A = (X, Y, Z)C.$$

In other words,

$$a_{ijk} = \sum_{\alpha=1}^{r_1}\sum_{\beta=1}^{r_2}\sum_{\gamma=1}^{r_3} x_{i\alpha}y_{j\beta}z_{k\gamma}c_{\alpha\beta\gamma}$$

for some full-rank matrices $X = [x_{i\alpha}] \in \mathbb{R}^{l\times r_1}$, $Y = [y_{j\beta}] \in \mathbb{R}^{m\times r_2}$, $Z = [z_{k\gamma}] \in \mathbb{R}^{n\times r_3}$, and core tensor $C = [\![c_{\alpha\beta\gamma}]\!] \in \mathbb{R}^{r_1\times r_2\times r_3}$. Again, X, Y, Z may be chosen to have orthonormal columns.

Observe that for matrices, this reduces to the $L_1DL_2^t$ or $Q_1RQ_2^t$ decompositions.

Norms and Inner Products

In order to discuss approximations, we need to define a norm on $\mathbb{R}^{d_1\times\cdots\times d_k}$.

The most convenient one to use is the Frobenius norm, $\|\cdot\|_F$, defined by

$$\|[\![a_{j_1\dots j_k}]\!]\|_F^2 = \sum_{j_1=1}^{d_1}\cdots\sum_{j_k=1}^{d_k} a_{j_1\dots j_k}^2.$$

for $[\![a_{j_1\dots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k}$.

It is the norm associated with the trace inner product, $\langle\cdot,\cdot\rangle_{\mathrm{tr}}$, defined by

$$\langle [\![a_{j_1\dots j_k}]\!] \mid [\![b_{j_1\dots j_k}]\!]\rangle_{\mathrm{tr}} := \sum_{j_1=1}^{d_1}\cdots\sum_{j_k=1}^{d_k} a_{j_1\dots j_k} b_{j_1\dots j_k}$$

for $[\![a_{j_1\dots j_k}]\!], [\![b_{j_1\dots j_k}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_k}$. Thus $\|A\|_F^2 = \langle A \mid A\rangle_{\mathrm{tr}}$.

Outer Product Approximation

A Candecomp/Parafac or outer product model has the following form

$$a_{ijk} = \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} + e_{ijk}$$

where $E = [\![e_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$ denotes the (unknown) error.

To minimize the error, we want an outer product approximation

$$\text{argmin} \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha \|_F$$

where the minimum is taken over all matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_r] \in \mathbb{R}^{l \times r}$, $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_r] \in \mathbb{R}^{m \times r}$, $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_r] \in \mathbb{R}^{n \times r}$.

In short, we want an optimal solution

$$B_\otimes^* = \underset{\text{rank}_\otimes(B) \le r}{\text{argmin}} \|A - B\|_F.$$

Multilinear Approximation

A Tucker or multilinear model has the following form

$$a_{ijk} = \sum_{\alpha=1}^{r_1} \sum_{\beta=1}^{r_2} \sum_{\gamma=1}^{r_3} x_{i\alpha} y_{j\beta} z_{k\gamma} c_{\alpha\beta\gamma} + e_{ijk}$$

where $E = [\![e_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$ denotes the (unknown) error.

To minimize the error, we want a multilinear approximation

$$\text{argmin} \| A - (X, Y, Z) C \|_F$$

where minimum is taken over all 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 tensor $C \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

In short, we want an optimal solution

$$B_{\boxplus}^* = \underset{\text{rank}_{\boxplus}(B) \le (r_1, r_2, r_3)}{\text{argmin}} \| A - B \|_F .$$

Outer Product Decomposition: Analytical Chemistry

Fluorescence spectra measurements by Burdick et. al. on one sample. The sample is a mixture of two pure components: benzo[*b*]fluoranthene and benzo[*k*]fluoranthene.

a_{ijk} = fluorescence emission intensity at wavelength λ_j^{em} excited with light at wavelength λ_k^{ex} and modulation frequency ν_i. Get 3-way data array $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l\times m\times n}$.

As before, we decompose A into a sum of outer products,

$$A = \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2.$$

The vectors $\mathbf{x}_\alpha, \mathbf{y}_\alpha, \mathbf{z}_\alpha$, $\alpha = 1, \dots, r$, yield the true chemical factors responsible for the data:

- there are 2 chemical components in the sample,

- $\mathbf{x}_\alpha = (x_{1\alpha}, \ldots, x_{l\alpha})$ holds the fluorescence lifetimes of the two components, $\alpha = 1, 2$,

- $\mathbf{y}_\alpha = (y_{1\alpha}, \ldots, y_{m\alpha})$ holds the excitation spectrum of the αth component, $\alpha = 1, 2$,

- $\mathbf{z}_\alpha = (z_{1\alpha}, \ldots, z_{n\alpha})$ holds the emission spectrum of the αth component, $\alpha = 1, 2$.

Multilinear Decomposition: Computer Vision

Application to facial recognition (TensorFaces) by Vasilescu and Terzopoulos. Facial image database of p male subjects photographed in q poses, r illuminations, s expressions, and stored as a grayscale image with t pixels.

a_{ijklm} = grayscale level of mth pixel of the image of ith person photographed in jth pose, with lth expression, under kth illumination level. Get 5-way data array $A = [\![a_{ijklm}]\!] \in \mathbb{R}^{p\times q\times r\times s\times t}$.

Let multilinear decomposition of A be

$$A = (V, W, X, Y, Z)C,$$

matrices V, W, X, Y, Z chosen to have orthonormal columns.

The column vectors of V, W, X, Y, Z are the 'principal components' or 'parameterizing factors' of the spaces of male subjects, poses, illuminations, expressions, and images respectively. The tensor C governs the interactions between these factors.

Other Models for 3-way Data

- Decomposition into Directional Components (Dedicom):

  $$\sum_{i=1}^{m} \|A_i - Q^t R_i Q\|_F^2$$

  $A_i = A_{i\bullet\bullet} \in \mathbb{R}^{n\times n}$ data. Optimize over $R_i \in \mathbb{R}^{r\times r}$ and $Q \in \mathbb{R}^{r\times n}$ with orthonormal columns.

- Simultaneous Components Analysis (SCA):

  $$\sum_{i=1}^{m} \|A_i - A_i B P_i\|_F^2$$

  $A_i = A_{i\bullet\bullet} \in \mathbb{R}^{m_i\times n}$ $(m_i \geq n)$ data. Optimize over $B \in \mathbb{R}^{n\times r}$ full-rank and $P_i \in \mathbb{R}^{r\times n}$ patterned matrix.

- Individual Difference Scaling (Indscal):

  $$\sum_{i=1}^{m} \|A_i^t A_i - B D_i H D_i^t B^t\|_F^2$$

  $A_i = A_{i\bullet\bullet} \in \mathbb{R}^{m\times n}$ data. Optimize over $B \in \mathbb{R}^{n\times r}$, $D_i \in \mathbb{R}^{r\times r}$ diagonal, and $H \in \mathbb{R}^{r\times r}$ positive definite (actually Indscal refers to the special case $H = I$).

Properties of Matrix Rank

1. Rank of $A \in \mathbb{R}^{m \times n}$ easy to determine (Gaussian Elimination)

2. Optimal rank-r approximation to $A \in \mathbb{R}^{m \times n}$ always exist (Eckart-Young Theorem)

3. Optimal 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. rank$(A) = \min\{m, n\}$

5. rank(A) from a non-orthogonal rank-revealing decomposition (e.g. $A = L_1 D L_2^t$) and rank(A) from an orthogonal rank-revealing decomposition (e.g. $A = Q_1 R Q_2^t$) are equal

6. Let A be a matrix with real entries. Then rank(A) is the same whether we regard A as an element of $\mathbb{R}^{m \times n}$ or as an element of $\mathbb{C}^{m \times n}$

Outer Product Rank vs Multilinear Rank

Every statement on the preceding slide is **false** for the outer product rank of order-k tensors, $k \geq 3$.

Every statement on the preceding slide is **true** for the multilinear rank of order-k tensors, $k \geq 3$.

In the next two slides we will spell these out explicitly for order-3 tensors. The restriction to order-3 tensors is strictly for notational simplicity. All statements generalize to order-k tensors for any $k \geq 3$.

Properties of Outer Product Rank

1. Computing $\mathrm{rank}_{\otimes}(A)$ for $A \in \mathbb{R}^{l \times m \times n}$ is NP-hard
2. For some $A \in \mathbb{R}^{l \times m \times n}$, $\operatorname{argmin}_{\mathrm{rank}_{\otimes}(B) \leq r} \|A - B\|_F$ does not have a solution
3. When $\operatorname{argmin}_{\mathrm{rank}_{\otimes}(B) \leq r} \|A - B\|_F$ 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 $\mathrm{rank}_{\otimes}(A) = r$ 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 $\mathrm{rank}_{\otimes}(A)$ number of terms
6. Let A be a 3-array with real entries. Then $\mathrm{rank}_{\otimes}(A)$ can take different values depending 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 $A = (X, Y, Z)C$. X, Y, Z may be chosen to have orthonormal columns

6. Let A be a matrix with real entries. Then $\operatorname{rank}_{\boxplus}(A)$ is the same whether we regard A as an element of $\mathbb{R}^{l \times m \times n}$ or as an element of $\mathbb{C}^{l \times m \times n}$

Generalization to Higher Order

- It is straight forward to generalize all statements on the last two slides to order-k tensors for any $k \geq 3$; we give two examples:

- Statement 2 for outer product rank:

  – For some $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, $\operatorname{argmin}_{\mathrm{rank}_{\otimes}(B) \leq r} \|A - B\|_F$ does not have a solution

- Statement 4 for multilinear rank:

  – Pick $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ at random, then A has

$$\mathrm{rank}_{\boxplus}(A) = (\min(d_1, d_2 \cdots d_k), \ldots, \min(d_k, d_1 \cdots d_{k-1}))$$

  with probability 1. The p-th slab rank above is just

$$\min(d_p, d_1 \cdots \widehat{d_p} \cdots d_k)$$

What About ‘Row Rank = Column Rank’

At first glance, this is one property of matrix rank that doesn’t seem to generalize to multilinear rank. Actually, it does in a more subtle way. We use the order-3 case as illustration.

Let $A \in \mathbb{R}^{l\times m\times n}$. Recall that we have defined the p-slab ranks:

$$\begin{aligned}
\operatorname{rank}_1(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{i\bullet\bullet} \mid i = 1,\dots,l\}),\\
\operatorname{rank}_2(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet j\bullet} \mid j = 1,\dots,m\})\\
\operatorname{rank}_3(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet\bullet k} \mid k = 1,\dots,n\}).
\end{aligned}$$

We may also define the (p,q)-slab ranks:

$$\begin{aligned}
\operatorname{rank}_{2,3}(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{\bullet jk} \mid j = 1,\dots,m; k = 1,\dots,n\}),\\
\operatorname{rank}_{1,3}(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{i\bullet k} \mid i = 1,\dots,l; k = 1,\dots,n\}),\\
\operatorname{rank}_{1,2}(A) &= \dim(\operatorname{span}_{\mathbb{R}}\{A_{ij\bullet} \mid i = 1,\dots,l; j = 1,\dots,m\}).
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
\operatorname{rank}_1(A) &= \operatorname{rank}_{2,3}(A),\\
\operatorname{rank}_2(A) &= \operatorname{rank}_{1,3}(A),\\
\operatorname{rank}_3(A) &= \operatorname{rank}_{1,2}(A).
\end{aligned}$$

Higher Level Trilinear Rank

The 1st level trilinear rank for an order-3 tensor is what we simply called trilinear rank earlier:

$$\operatorname{rank}^1_{\boxplus}(A) = (\operatorname{rank}_1(A), \operatorname{rank}_2(A), \operatorname{rank}_3(A))$$

The 2nd level trilinear rank for an order-3 tensor is:

$$\operatorname{rank}^2_{\boxplus}(A) = (\operatorname{rank}_{2,3}(A), \operatorname{rank}_{1,3}(A), \operatorname{rank}_{1,2}(A)).$$

Hence the result at the end of the previous slide may be restated for $A \in \mathbb{R}^{l \times m \times n}$ as simply

$$\operatorname{rank}^1_{\boxplus}(A) = \operatorname{rank}^2_{\boxplus}(A).$$

Note that for $A \in \mathbb{R}^{m \times n} = \mathbb{R}^{1 \times m \times n}$, this reduces to

$$(1, \operatorname{rowrank}(A), \operatorname{colrank}(A)) = (1, \operatorname{colrank}(A), \operatorname{rowrank}(A)),$$

and thus $\operatorname{rowrank}(A) = \operatorname{colrank}(A)$.

Higher Level Multilinear Rank

Let $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. For any $\{p_1, \ldots, p_l\} \subset \{1, \ldots, k\}$, $p_1 < \cdots < p_k$, we may define $(p_1, \ldots, p_l)$-slab rank accordingly.

The $\binom{k}{l}$-tuple of $(p_1, \ldots, p_l)$-slab ranks gives the lth level multilinear rank, for $l = 1, \ldots, k-1$.

Theorem (de Silva and L., 2005). The lth level multilinear rank is equal to the $(k-l)$th level multilinear rank, $l = 1, \ldots, k-1$.

Moral

The multilinear rank is the generalization of matrix rank that preserves most of the common properties of matrix rank.

We should stop expecting the outer product rank to resemble matrix rank in any way.

However, the outer product rank is the generalization that is more important in applications.

Furthermore, the mathematical and computational challenges from studies of outer product rank shows that it is a more interesting object than multilinear rank.

The remainder of this talk will be about our efforts in overcoming several of these problems in order to get a satisfactory statistical model based on outer product approximation.

Ultimate Objective

Statistical theory for multiway data analysis

Obstacles:

- Ill-posedness of $\operatorname{argmin}_{\mathrm{rank}_{\otimes}(B)\leq r}\|A - B\|_F$
- Lack of a generic rank, ie. no Candecomp/Parafac model that gives a perfect fit almost everywhere
- Existing algorithm not convergent to globally minima

Ill-posedness of Optimal Low-Rank Approximation

The problem $\operatorname{argmin}_{\operatorname{rank}_{\otimes}(B)\leq r}\|A-B\|_F$ may not have an optimal solution when $r \geq 2$, $k \geq 3$. In fact

Theorem (L. and Golub, 2004). For tensors of any order $k \geq 3$ and with respect to any choice of norm on $\mathbb{R}^{d_1\times\cdots\times d_k}$, there exists an instance $A \in \mathbb{R}^{d_1\times\cdots\times d_k}$ such that A fails to have an optimal rank-r approximation for some $r \geq 2$. On the other hand, an optimal solution always exist for $k = 2$ and $r = 1$.

In the next slide, we give an explicit example.

Example

$\mathbf{x}, \mathbf{y}$ two linearly independent vectors in $\mathbb{R}^2$. Consider the order-3 tensor in $\mathbb{R}^{2\times 2\times 2}$,

$$A := \mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x} + \mathbf{x}\otimes\mathbf{y}\otimes\mathbf{y} + \mathbf{y}\otimes\mathbf{x}\otimes\mathbf{y}.$$

A has rank 3: straight forward.

A has no optimal rank-2 approximation: consider sequence $\{B_n\}_{n=1}^{\infty}$ in $\mathbb{R}^{2\times 2\times 2}$,

$$B_n := \mathbf{x}\otimes\mathbf{x}\otimes(\mathbf{x} - n\mathbf{y}) + \left(\mathbf{x} + \frac{1}{n}\mathbf{y}\right)\otimes\left(\mathbf{x} + \frac{1}{n}\mathbf{y}\right)\otimes n\mathbf{y},$$

Clear that $\operatorname{rank}_{\otimes}(B_n) \le 2$ for all n. By multilinearity of $\otimes$,

$$\begin{aligned} B_n = \mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x} - n\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{y} + n\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{y} \\ + \mathbf{x}\otimes\mathbf{y}\otimes\mathbf{y} + \mathbf{y}\otimes\mathbf{x}\otimes\mathbf{y} + \frac{1}{n}\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y} = A + \frac{1}{n}\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y}. \end{aligned}$$

For any choice of norm on $\mathbb{R}^{2\times 2\times 2}$,

$$\|A - B_n\| = \frac{1}{n}\|\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y}\| \to 0 \qquad \text{as } n \to \infty.$$

Surprising Find

It has always been assumed that an optimal rank-r approximation exist for high-order tensors and there has been continual interests in finding an 'Eckart-Young theorem'-like result for tensors of higher order. The view expressed in the conclusion of the following paper is representative of such efforts:

"An Eckart-Young type of optimal rank-k approximation theorem for tensors continues to elude our investigations but can perhaps eventually be attained by using a different norm or yet other definitions of orthogonality and rank."

Source: T.G. Kolda, "Orthogonal tensor decompositions," *SIAM J. Matrix Anal. Appl.*, **23** (1), 2001 , pp. 243–255.

A simple fact that's often overlooked: in a norm space, the minimum distance of a point A to a non-closed set $\mathcal{S}$ may not be attained by any point in $\mathcal{S}$.

For tensors of order $k \geq 3$, $r \geq 2$, the set

$$\{A \in \mathbb{R}^{d_1 \times \cdots \times d_k} \mid \mathsf{rank}_{\otimes}(A) \leq r\}$$

may not be closed. This is norm independent since all norms are equivalent on finite dimensional spaces.

However we still need to ‘solve’ the problem

$$\mathsf{argmin}_{\mathsf{rank}_{\otimes}(B) \leq r} \|A - B\|_F$$

in order to analyze multiway data.

How can we overcome the ill-posedness?

Quick but Flawed Fix

Current way to force a solution: perturb the problem by small $\varepsilon > 0$ and find approximate solution $\mathbf{x}_i^*(\varepsilon), \mathbf{y}_i^*(\varepsilon) \in \mathbb{R}^{d_i}$ $(i = 1, 2, 3)$ with

$$\begin{aligned}\|A - \mathbf{x}_1^*(\varepsilon) \otimes \mathbf{y}_1^*(\varepsilon) \otimes \mathbf{z}_1^*(\varepsilon) - \mathbf{x}_2^*(\varepsilon) \otimes \mathbf{y}_2^*(\varepsilon) \otimes \mathbf{z}_2^*(\varepsilon)\| \\ = \varepsilon + \inf_{\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{d_i}} \|A - \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 - \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2\|.\end{aligned}$$

Serious numerical problems due to ill-conditioning (a phenomenon often referred to as *degeneracy* or *swamp* in Chemometrics and Psychometrics).

Reason? Rule of thumb in Computational Math:

A well-posed problem near to an ill-posed one is ill-conditioned.

So, even if we may perturb an ill-posed problem slightly to get a well-posed one, the perturbed problem will more often than not be ill-conditioned.

Discriminants

Definition. $f(x_1, \ldots, x_k)$ polynomial, $\deg(f) \leq d$, the discriminant $\Delta(f)$ is a polynomial function in the coefficients of f so that $\Delta(f)$ whenever f has a multiple root (ie. common root of f and ∇f).

Quadratic 1-variable: $f(x) = a + bx + cx^2$, $\Delta(f) = b^2 - 4ac$

Cubic 1-variable: $f(x) = a + bx + cx^2 + dx^3$, $\Delta(f) = b^2c^2 - 4b^3d - 4ac^3 - 27a^2d^2 + 18abcd$

General 1-variable: $f(x) = \prod_{i=1}^{\deg(f)}(x - \lambda_i)$, $\Delta(f) = \prod_{i<j}(\lambda_i - \lambda_j)^2$

Resultant: $\text{Res}(f, g) = \Delta(f(x) + yg(x))$

Determinant: $A \in \mathbb{R}^{n \times n}$, $\det(A) = \Delta(f_A)$ where $f_A(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t A \mathbf{y} = \sum_{i,j} a_{ij} x_i y_j$

Hyperdeterminant

The hyperdeterminant of a tensor $A = [\![a_{ijk}]\!] \in \mathbb{R}^{2\times 2\times 2}$ is defined as

$$\begin{aligned}\Delta(A) := (&a_{000}^2 a_{111}^2 + a_{001}^2 a_{110}^2 + a_{010}^2 a_{101}^2 + a_{011}^2 a_{100}^2)\\ &- 2(a_{000}a_{001}a_{110}a_{111} + a_{000}a_{010}a_{101}a_{111} + a_{000}a_{011}a_{100}a_{111}\\ &+ a_{001}a_{010}a_{101}a_{110} + a_{001}a_{011}a_{110}a_{100} + a_{010}a_{011}a_{101}a_{100})\\ &+ 4(a_{000}a_{011}a_{101}a_{110} + a_{001}a_{010}a_{100}a_{111}).\end{aligned}$$

This formula first appeared in a paper by Cayley published in 1845 but remained obscure until a study by Gelfand, Kapranov, Zelevinsky in 1992.

A result that parallels the matrix case is the following: the system of bilinear equations

$$\begin{aligned}
a_{000}x_0y_0 + a_{010}x_0y_1 + a_{100}x_1y_0 + a_{110}x_1y_1 &= 0,\\
a_{001}x_0y_0 + a_{011}x_0y_1 + a_{101}x_1y_0 + a_{111}x_1y_1 &= 0,\\
a_{000}x_0z_0 + a_{001}x_0z_1 + a_{100}x_1z_0 + a_{101}x_1z_1 &= 0,\\
a_{010}x_0z_0 + a_{011}x_0z_1 + a_{110}x_1z_0 + a_{111}x_1z_1 &= 0,\\
a_{000}y_0z_0 + a_{001}y_0z_1 + a_{010}y_1z_0 + a_{011}y_1z_1 &= 0,\\
a_{100}y_0z_0 + a_{101}y_0z_1 + a_{110}y_1z_0 + a_{111}y_1z_1 &= 0.
\end{aligned}$$

has a non-trivial solution iff $\Delta(A) = 0$.

Theorem (Gelfand, Kapranov, Zelevinsky, 1992). $\mathbb{R}^{(d_1+1)\times\cdots\times(d_k+1)}$ has a non-trivial hyperdeterminant if and only if

$$d_j \leq \sum_{i \neq j} d_i$$

for all $j = 1, \ldots, k$.

For $\mathbb{R}^{m\times n}$, the condition becomes $m \leq n$ and $n \leq m$ — that's why matrix determinants is only defined for square matrices.

Weak solutions to PARAFAC

Theorem (de Silva and L., 2004). Let $l, m, n \geq 2$. Let $A \in \mathbb{R}^{l\times m\times n}$ with $\text{rank}_{\otimes}(A) = 3$. A is the limit of a sequence $B_n \in \mathbb{R}^{l\times m\times n}$ with $\text{rank}_{\otimes}(B_n) \leq 2$ if and only if

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

where $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\{\mathbf{y}_1, \mathbf{y}_2\}$, $\{\mathbf{z}_1, \mathbf{z}_2\}$ are linearly independent sets in $\mathbb{R}^l$, $\mathbb{R}^m$, and $\mathbb{R}^n$ respectively.

With this, we can overcome the ill-posedness of $\text{argmin}_{\text{rank}_{\otimes}(B)\leq r}\|A - B\|_F$ by replacing $\text{rank}_{\otimes}$ with $\text{closedrank}_{\otimes}$, defined by

$$\{A \mid \text{closedrank}_{\otimes}(A) \leq r\} = \overline{\{A \mid \text{rank}_{\otimes}(A) \leq r\}}.$$

For order-3 tensor, it follows from the theorem that

$$\begin{aligned}&\{A \in \mathbb{R}^{l\times m\times n} \mid \text{closedrank}_{\otimes}(A) \leq 2\} = \\ &\quad \{\mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_1 \otimes \mathbf{z}_2 + \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_1 \mid \mathbf{x}_i \in \mathbb{R}^l, \mathbf{y}_i \in \mathbb{R}^m, \mathbf{z}_i \in \mathbb{R}^n\} \\ &\qquad \cup \{\mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2 \mid \mathbf{x}_i \in \mathbb{R}^l, \mathbf{y}_i \in \mathbb{R}^m, \mathbf{z}_i \in \mathbb{R}^n\}\end{aligned}$$

Sketch of Proof

Restrict to the special case $l = m = n = 2$. There is a technical lemma that allows us to generalize to arbitrary l, m, n.

1. Regard $A \in \mathbb{R}^{2\times 2\times 2}$ as two slabs of 2×2 matrices:

$$A = [A_0 \mid A_1] = \left[\begin{array}{cc|cc} a_{000} & a_{001} & a_{100} & a_{101} \\ a_{010} & a_{011} & a_{110} & a_{111} \end{array}\right] \in \mathbb{R}^{2\times 2\times 2}.$$

2. One can check that

$$\det(\lambda_0 A_0 + \lambda_1 A_1) = \lambda_0^2 \det(A_0) + \lambda_0\lambda_1 \frac{\det(A_0 + A_1) - \det(A_0 - A_1)}{2} + \lambda_1^2 \det(A_1).$$

3. Define Δ to be the discriminant of this quadratic polynomial:

$$\Delta(A) = \left[\frac{\det(A_0 + A_1) - \det(A_0 - A_1)}{2}\right]^2 - 4\det(A_0)\det(A_1).$$

Easy to check that $\Delta(A)$ is exactly the hyperdeterminant of A defined earlier.

4. For $L = (L_1, L_2, L_3) \in \mathsf{GL}_2(\mathbb{R}) \times \mathsf{GL}_2(\mathbb{R}) \times \mathsf{GL}_2(\mathbb{R})$, we can show that

$$\Delta(LA) = \det(L_1)^2 \det(L_2)^2 \det(L_3)^2 \Delta(A).$$

Thus the sign of $\Delta(A)$ is invariant under multiplication by non-singular matrices on three sides. In particular, Gaussian elimination applied to the three sides of A does not change the sign of Δ.

5. Can show that $\text{rank}_{\otimes}(A) \geq 3 \Rightarrow \Delta(A) \leq 0$ and $\text{rank}_{\otimes}(A) \leq 2 \Rightarrow \Delta(A) \geq 0$.

6. If A is a limit point of rank-2 tensors, then $\Delta(A) = 0$ and by Gaussian elimination, A can be transformed into

$$\left[\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right] = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2.$$

Generic Rank

Zariski topology on $\mathbb{C}^{d_1 \times \cdots \times d_k} = \mathbb{C}^d$, ie. topology generated by closed sets of the form $\mathbb{V}(f) = \{\mathbf{z} \in \mathbb{C}^d \mid f(\mathbf{z}) = 0\}$.

A property P is said to be generic in $\mathbb{C}^d$ if the set of elements for which P doesn't hold is contained in a union of closed sets, each of dimension not more than $n-1$. In particular, the set of elements where P doesn't hold has zero volume in $\mathbb{C}^d$ and the set of elements where P holds has the same volume as $\mathbb{C}^d$.

It follows that if a generic rank $\bar{r}$ exists for $\mathbb{C}^{d_1 \times \cdots \times d_k} \cong \mathbb{C}^d$ $(d = d_1 \cdots d_k)$, then the set of elements $A \in \mathbb{C}^{d_1 \times \cdots \times d_k}$ satisfying the property $\operatorname{rank}_{\otimes}(A) = \bar{r}$ must have the same volume as $\mathbb{C}^d$ and thus is dense

Caution: The term generic rank is used in a different way in the Computational Complexity literature (its existence is trivial).

Rank Depends on Base Field

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ where $n \geq 2$. Write $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbb{C}^n$. Then

$$\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} - \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{y} + \mathbf{y} \otimes \mathbf{x} \otimes \mathbf{y} + \mathbf{y} \otimes \mathbf{y} \otimes \mathbf{x} \\ = \frac{1}{2}(\mathbf{z} \otimes \bar{\mathbf{z}} \otimes \bar{\mathbf{z}} + \bar{\mathbf{z}} \otimes \mathbf{z} \otimes \mathbf{z}).$$

lhs has rank 3 in $\mathbb{R}^{n \times n \times n}$ while rhs has rank 2 in $\mathbb{C}^{n \times n \times n}$.

Recall in $\mathbb{R}^{2\times2\times2}$, $\Delta(A) > 0 \Rightarrow \operatorname{rank}_{\otimes}(A) = 2$, $\Delta(A) < 0 \Rightarrow \operatorname{rank}_{\otimes}(A) = 3$. It follows that both the set of rank-2 tensors and the set of rank-3 tensors have positive volumes in $\mathbb{R}^{2\times2\times2}$. So there is no 'generic' rank.

On the other hand, we may show that the generic rank in $\mathbb{C}^{2\times2\times2}$ is 2.

Generic Rank for Complex Tensors

A locally closed sets is one that can be written as an intersection of an open set and a closed set. A constructible set is one that can be written as a finite union of locally closed set.

Theorem (L., de Silva, Comon, 2005). For every $r \in \mathbb{N}$, the set of tensors of rank r is a constructible set.

Corollary 1. There exist $p_1, \dots, p_N \in \mathbb{C}[X_1, \dots, X_d]$ such that for any $r \in \{0, \dots, r_{\text{max}}\}$,

$$\{A \in \mathbb{C}^{d_1 \times \cdots \times d_k} \mid \text{rank}_{\otimes}(A) = r\} = \mathcal{Y}(I_{r,1}) \cup \cdots \cup \mathcal{Y}(I_{r,m_r})$$

where for $I = \{i_1, \dots, i_s\} \subseteq \{1, \dots, N\}$ and $I^c = \{j_1, \dots, j_{N-s}\} = \{1, \dots, N\} \setminus I$, we write

$$\begin{aligned}\mathcal{Y}(I) &:= \mathbb{V}(p_{i_1} \cdots p_{i_s}) \cap \mathbb{V}(p_{j_1}, \dots, p_{j_{N-s}})^c \\ &= \mathbb{V}(p_{i_1}) \cap \cdots \cap \mathbb{V}(p_{i_s}) \cap \mathbb{V}(p_{j_1})^c \cap \cdots \cap \mathbb{V}(p_{j_{N-s}})^c.\end{aligned}$$

Corollary 2. A generic outer-product rank exist for $\mathbb{C}^{d_1 \times \cdots \times d_k}$.

Sketch of Proof

Work over complex projective space $\mathbb{P}^n$ for simplicity.

1. $S_1 := \{A \in \mathbb{P}^d \mid \operatorname{rank}_\otimes(A) = 1\}$ is exactly image of the Segre map
$$\mathbb{P}^{d_1} \times \cdots \times \mathbb{P}^{d_k} \to \mathbb{P}^d, \qquad (\mathbf{x}_1, \dots, \mathbf{x}_k) \mapsto \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_k$$
where $d = (d_1 + 1) \cdots (d_k + 1) - 1$.

2. For $r \geq 2$, $S_r := \{A \in \mathbb{P}^d \mid \operatorname{rank}_\otimes(A) \leq r\}$ is the union of the secant r-planes to S_1, ie.
$$S_r = \{B \in \overline{A_1 \cdots A_r} \mid A_1, \dots, A_r \in S_1\}.$$
Note that this is not a Zariski-closed set.

3. S_r can be parameterized in the following way:
$$\varphi : S_1 \times_s \cdots \times_s S_1 \times_s \mathbb{P}^r \to \mathbb{P}^d$$

where there are r copies of S_1 and $\times_s$ denotes Segre product. The image of φ is S_r.

4. φ is a morphism of finite-type and the Chevalley's Constructibility Theorem, which in its full generality says that the image of a morphism of finite-type between two Noetherian schemes is constructible, shows that S_r is constructible.

5. Let $X_r := \{A \in \mathbb{P}^d \mid \text{rank}_{\otimes}(A) = r\}$. Then $X_r = S_r \setminus S_{r-1}$ is also a constructible set.

6. Alternatively, may show directly without invoking Chevalley's Constructibility Theorem that S_r is a quasi-projective variety (locally closed set) and thus X_r is constructible.

7. Since $\mathbb{P}^d$ is a disjoint union $X_1 \sqcup \cdots \sqcup X_{r_{\text{max}}}$, exactly one X_g, $g \in \{2, \ldots, r_{\text{max}}\}$, must contain an open set. Thus g is the required generic rank.

Alternating Least Squares

Even when an optimal solution $A_{\otimes}^{*}$ to $\text{argmin}_{\text{rank}_{\otimes}(B)\le r}\|A-B\|_F$ exists, $A_{\otimes}^{*}$ is not easy to compute since the objective function is non-convex. Since 1979, multiway data analysis rely primarily on the following nonlinear Gauss-Seidel 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 $\rho^{(k+1)}/\rho^{(k)} > \varepsilon$;
        $X^{(k+1)} \leftarrow \text{argmin}_{\bar{X}\in\mathbb{R}^{l\times r}}\|T - \sum_{\alpha=1}^{r} \bar{x}_\alpha^{(k+1)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}\|_F^2$;
        $Y^{(k+1)} \leftarrow \text{argmin}_{\bar{Y}\in\mathbb{R}^{m\times r}}\|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes \bar{y}_\alpha^{(k+1)} \otimes z_\alpha^{(k)}\|_F^2$;
        $Z^{(k+1)} \leftarrow \text{argmin}_{\bar{Z}\in\mathbb{R}^{n\times r}}\|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes \bar{z}_\alpha^{(k+1)}\|_F^2$;
        $\rho^{(k+1)} \leftarrow \|\sum_{\alpha=1}^{r}[x_\alpha^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes z_\alpha^{(k+1)} - x_\alpha^{(k)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}]\|_F^2$;
        $k \leftarrow k + 1$;
```

Problem: Not globally convergent even when there is an optimal solution. When it converges, there's no guarantee that the solution will be a global minima.

SDP-Based Algorithm

Observation 1:

$$\begin{aligned} F(x_{11},\dots,z_{nr}) &= \|A - \textstyle\sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2 \\ &= \textstyle\sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha}\right)^2 \end{aligned}$$

is a polynomial of total degree 6 (resp. $2k$ for order k-tensors) in variables $x_{11},\dots,z_{nr}$.

Recent breakthroughs in multivariate polynomial optimization [Lasserre 2001], [Parrilo 2003] [Parrilo-Sturmfels 2003] show that the non-convex problem

$$\operatorname{argmin} F(x_{11},\dots,z_{nr})$$

may be relaxed to a convex problem (thus global optima is guranteed) which can in turn be solved using SDP.

How It Works

Observation 2: If $F - \lambda$ can be expressed as a sum of squares of polynomials

$$F(x_{11}, \ldots, z_{nr}) - \lambda = \sum_{i=1}^{n} P_i(x_{11}, \ldots, z_{nr})^2,$$

then λ is a global lower bound for F, ie.

$$F(x_{11}, \ldots, z_{nr}) \geq \lambda$$

for all $x_{11}, \ldots, z_{nr} \in \mathbb{R}$.

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

Write $\mathbf{v} = (1, x_{11}, \ldots, z_{nr}, \ldots, x_{l1}y_{m1}z_{n1}, \ldots, z_{nr}^6)^t$, the N-tuple of monomials of total degree ≤ 6, where

$$N = \binom{r(l+m+n)+3}{3}.$$

Write $F(x_{11}, \ldots, z_{nr}) = \boldsymbol{\alpha}^t \mathbf{v}$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ are the coefficients of the respective monomials.

Since deg(F) is even, F may also be written as

$$F(x_{11}, \ldots, z_{nr}) = \mathbf{v}^t M \mathbf{v}$$

for some $M \in \mathbb{R}^{N \times N}$. So

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

where $E_{11} = \mathbf{e}_1 \mathbf{e}_1^t \in \mathbb{R}^{N \times N}$.

Observation 3: The rhs is a sum of squares iff $M - \lambda E_{11}$ is positive semi-definite (since $M - \lambda E_{11} = B^t B$).

Hence we have

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

This is an SDP problem

$$\begin{array}{rll} \text{minimize} & 0 \circ S - \lambda & \\ \text{subjected to} & S \circ B_1 + \lambda = \alpha_1, & \\ & S \circ B_k = \alpha_k, & k = 2, \ldots, N \\ & S \succeq 0, & \lambda \in \mathbb{R}. \end{array}$$

This problem can be solved in polynomial time. Like all SDP-based algorithms, the SPD duality produces a certificate that tells us whether we have arrived at a globally optimal solution.

The Candecomp/Parafac model is used as an example but the algorithm, like ALS, applies to other models (Tucker, Dedicom, SCA, Indscal) as well.

Global Convergence Issues

Hilbert 17th Problem (Artin 1927). Any multivariate polynomial function $F : \mathbb{R}^N \to \mathbb{R}$ that has $F(\mathbf{x}) \geq 0$ for every $\mathbf{x} \in \mathbb{R}^N$ is a sum of squares of rational functions.

Cannot replace rational functions by polynomial functions in general (eg. $w^4 + x^2y^2 + y^2z^2 + z^2x^2 - 4xyzw$).

However, *if* those of the form

$$\mu + \sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} \right)^2$$

can *always* be written as a sum of polynomials (we don't know), then the SDP algorithm for optimal low-rank tensor approximation will *always* converge globally.

Numerical experiments performed by Parrilo on more general polynomials yield $\lambda^* = \min F$ in all cases.

Catch: For rank-r approximations to order-k tensors $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$,

$$N = \binom{r(d_1 + \cdots + d_k) + k}{k}$$

is large even for moderate d_i, r and k.

Sparsity to the rescue? The polynomials that we are interested in are always sparse (eg. for $k = 3$, only terms of the form xyz or $uvwxyz$ appear).