Numerical Multilinear Algebra in Data Analysis

Lek-Heng Lim
Stanford University

Computer Science Colloquium
Ithaca, NY
February 20, 2007

Collaborators: O. Alter, P. Comon, V. de Silva, I. Dhillon, M. Gu, B. Sturmfels; Thanks: G. Carlsson, G. Golub

A spectroscopic motivation

Spectroscopy: Measurements of spectrum, ie. light absorption or emission of a specimen as a function of energy.

Typical specimen contains 10^{13} to 10^{16} light absorbing entities or chromophores (molecules, amino acids, etc).

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

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

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

Bilinear model aka factor analysis: $A_{m\times n} = E_{m\times r} C_{r\times n}$ rank-revealing factorization or, rather, low-rank approximation

$$\min \| A_{m\times n} - E_{m\times r} C_{r\times n} \|$$

in the presence of noise. Origin of the "spectrum" of matrix.

Text mining

Text mining is the spectroscopy of documents.

Specimens = documents; chromophores = terms.

Absorbance = inverse document frequency:

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

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

Bilinear model

Bilinear model works on 'two-way' data: measurements on object i (genomes, chemical samples, images, webpages, consumers, etc) yield a vector $a_i \in \mathbb{R}^n$ where $n =$ number of features of i.

A collection of m such objects, $A = [a_1, \dots, a_m]$ may be regarded as an m-by-n matrix. E.g. gene × microarray matrices in bioinformatics, terms × documents matrices in text mining, facial images × individuals matrices in computer vision.

Various matrix techniques may be applied to extract useful information: QR, EVD, SVD, NMF, CUR, compressed sensing techniques, etc.

Examples of bilinear models: 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 $k > 2$.

All you need to know about tensors

A set of multiply indexed real numbers $A = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l\times m\times n}$ on which the following algebraic operations are defined:

1. Addition/Scalar Multiplication: for $[\![b_{ijk}]\!] \in \mathbb{R}^{l\times m\times n}$, $\lambda \in \mathbb{R}$,

   $[\![a_{ijk}]\!]+[\![b_{ijk}]\!] := [\![a_{ijk}+b_{ijk}]\!]$ and $\lambda[\![a_{ijk}]\!] := [\![\lambda a_{ijk}]\!] \in \mathbb{R}^{l\times m\times n}$

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

   $$c_{i'j'k'} := \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n} \lambda_{i'i}\mu_{j'j}\nu_{k'k}a_{ijk}.$$

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 $(M, N) \cdot A = MAN^t$.

Aside: mathematician's definition

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

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

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

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

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

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

Aside: physicists' 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.

Change-of-basis theorem for tensors. 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.

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 $\mathbf{u} \in \mathbb{R}^l, \mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$. For order 2, $\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^t$.

In general, any $A \in \mathbb{R}^{l \times m \times n}$ may be written as a sum of decomposable tensors

$$A = \textstyle\sum_{i=1}^{r} \lambda_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i.$$

May be written as a multilinear matrix multiplication:

$$A = (U, V, W) \cdot \Lambda.$$

$U \in \mathbb{R}^{l \times r}, V \in \mathbb{R}^{m \times r}, W \in \mathbb{R}^{n \times r}$ and diagonal $\Lambda \in \mathbb{R}^{r \times r \times r}$.

Tensor ranks

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

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

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

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

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

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

In general, $\text{rank}_{\otimes}(A) \neq r_1(A) \neq r_2(A) \neq r_3(A)$. Original definition due to Frank L. Hitchcock (1927).

Outer product rank is difficult

Matrix: $A \in \mathbb{R}^{m \times n}$, easy to compute $\operatorname{rank}_{\otimes}(A)$.

Theorem (Håstad). Computing $\operatorname{rank}_{\otimes}(A)$ for $A \in \mathbb{R}^{l \times m \times n}$ is an NP-hard problem.

Matrix: $A \in \mathbb{R}^{m \times n} \subset \mathbb{C}^{m \times n}$, $\operatorname{rank}(A)$ is the same whether we regard it as a real matrix or a complex matrix.

Theorem (Bergman). For $A \in \mathbb{R}^{l \times m \times n} \subset \mathbb{C}^{l \times m \times n}$, $\operatorname{rank}_{\otimes}(A)$ is base field dependent.

Matrix: $A \in \mathbb{R}^{m \times n}$, maximal $\operatorname{rank}_{\otimes}(A)$ is $\min\{m, n\}$.

Maximal outer product rank known only in very special cases: T.D. Howell, "Global properties of tensor rank," *Linear Algebra Appl.*, **22** (1978), pp. 9–23.

Outer product rank is useful

P. Bürgisser, M. Clausen, and M.A. Shokrollahi, *Algebraic complexity theory*, Springer-Verlag, Berlin, 1996.

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

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

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

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

$O(n^{2+\varepsilon})$ algorithm for multiplying two $n\times n$ matrices gives $O(n^{2+\varepsilon})$ algorithm for solving system of n linear equations [Strassen 1969].

Conjecture. $\log_2(\mathrm{rank}_\otimes(T)) \leq 2 + \varepsilon$.

Best known result. $O(n^{2.376})$ [Coppersmith-Winograd 1987; Cohn-Kleinberg-Szegedy-Umans 2005].

Outer product decomposition in spectroscopy

Application to fluorescence spectral analysis 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,

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

yield 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 samples $1, \dots, l$,
- $\mathbf{y}_\alpha = (y_{1\alpha}, \dots, y_{m\alpha})$: excitation spectrum of αth substance,
- $\mathbf{z}_\alpha = (z_{1\alpha}, \dots, z_{n\alpha})$: emission spectrum of αth substance.

Multilinear model: CANDECOMP/PARAFAC

In the noisy case, want

$$\operatorname{argmin}_{\mathbf{x}_\alpha,\mathbf{y}_\alpha,\mathbf{z}_\alpha}\left\|A-\sum\nolimits_{\alpha=1}^{r}\mathbf{x}_\alpha\otimes\mathbf{y}_\alpha\otimes\mathbf{z}_\alpha\right\|.$$

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

$X=[\mathbf{x}_1,\dots,\mathbf{x}_r], Y=[\mathbf{y}_1,\dots,\mathbf{y}_r], Z=[\mathbf{z}_1,\dots,\mathbf{z}_r].$

Theorem (Kruskal). Decomposition is unique up to scaling if

$$\operatorname{spark}(X)+\operatorname{spark}(Y)+\operatorname{spark}(Z)\geq 2r-1.$$

Avoids factor indeterminacy under mild conditions.

Multilinear decomposition in bioinformatics

Application to cell cycle studies by Alter and Omberg. Collection of gene-by-microarray matrices $A_1, \dots, A_l \in \mathbb{R}^{m \times n}$ obtained under varying oxidative stress.

a_{ijk} = expression level of jth gene in kth microarray under ith stress. Get 3-way data array $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.

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 the ‘principal components’ or ‘parameterizing factors’ of the spaces of stress, genes, and microarrays respectively. C governs interactions between these factors.

Noisy case: approximate by discarding small c_{ijk} (Tucker Model).

Fundamental problem of multiway data analysis

Let A be a tensor, symmetric tensor, or nonnegative tensor. Solve

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

where rank may be outer product rank, multilinear rank, symmetric rank (for symmetric tensors), or nonnegative rank (nonnegative tensors).

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

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

or $C \in \mathbb{R}^{r_1\times r_2\times r_3}$ and $L_i \in \mathbb{R}^{d_i\times r_i}$ that minimizes

$$\|A - (L_1, L_2, L_3)\cdot C\|.$$

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

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

Decompositional approach to data analysis

More generally, $\mathbb{F} = \mathbb{C}$, $\mathbb{R}$, $\mathbb{R}_+, \mathbb{R}_{\text{max}}$ (max-plus algebra), $\mathbb{R}[x_1, \dots, x_n]$ (polynomial rings), etc.

Dictionary, $\mathcal{D} \subset \mathbb{F}^N$, not contained in any hyperplane. Let $\mathcal{D}_2 =$ union of bisecants to $\mathcal{D}$, $\mathcal{D}_3 =$ union of trisecants to D, . . . , $\mathcal{D}_r =$ union of r-secants to $\mathcal{D}$.

Define $\mathcal{D}$-rank of $A \in \mathbb{F}^N$ to be $\min\{r \mid A \in \mathcal{D}_r\}$.

If $\varphi : \mathbb{F}^N \times \mathbb{F}^N \to \mathbb{R}$ is some measure of 'nearness' between pairs of points (eg. norms, Bregman divergences, etc), we want to find a best low-rank approximation to A:

$$\operatorname{argmin}\{\varphi(A, B) \mid \mathcal{D}\text{-rank}(B) \leq r\}.$$

Segre variety and secant varieties

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:

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

Tensors that have rank > 1 are elements on the higher secant varieties of $\mathcal{S} = \text{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n)$. Eg. a tensor has rank 2 if it sits on a secant line through two points in $\mathcal{S}$ but not on $\mathcal{S}$, rank 3 if it sits on a secant plane through three points in $\mathcal{S}$ but not on any secant lines, etc.

Feature revelation

Get low-rank approximation

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

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

Note that another way to say ‘best low-rank’ is ‘sparsest possible’.

Example. $\mathcal{D} = \{A \mid \operatorname{rank}_{\otimes}(A) \leq 1\}$, $\varphi(A, B) = \|A - B\|_F$ — get usual PARAFAC.

Example. $\mathcal{D} = \{A \mid \operatorname{rank}_{\boxplus}(A) \leq (1, 1, 1)\}$, $\varphi(A, B) = \|A - B\|_F$ — get Tucker Model.

Example. $\mathcal{D} = \{A \mid \operatorname{rank}_{\boxplus}(A) \leq (r_1, r_2, r_3)\}$ (an algebraic set), $\varphi(A, B) = \|A - B\|_F$ — get De Lathauwer model.

Simple lemma

Lemma (de-Silva, L.). 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:

(a) The set $\mathcal{S}_r(d_1, \dots, d_k) := \{A \mid \operatorname{rank}_\otimes(A) \leq r\}$ is not closed.

(b) There exists a sequence A_n, $\operatorname{rank}_\otimes(A_n) \leq r$, $n \in \mathbb{N}$, converging to B with $\operatorname{rank}_\otimes(B) > r$.

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

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

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

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

is not attained (by any A with $\operatorname{rank}_\otimes(A) \leq r$).

Non-existence of best low-rank approximation

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

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

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

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

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

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

[Based on an exercise in D. Knuth, *The art of computer programming*, **2**, 3rd Ed., Addison-Wesley, Reading, MA, 1997.]

Outer product approximations are ill-behaved

Such phenomenon can and will happen for all orders > 2, all norms, and many ranks:

Theorem 1 (de Silva, L). Let $k \geq 3$ and $d_1, \dots, d_k \geq 2$. For any s such that $2 \leq s \leq \min\{d_1, \dots, d_k\} - 1$, there exist $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ with $\operatorname{rank}_{\otimes}(A) = s$ such that A has no best rank-r approximation for some $r < s$. The result is independent of the choice of norms.

For matrices, the quantity $\min\{d_1, d_2\}$ will be the maximal possible rank in $\mathbb{R}^{d_1 \times d_2}$. In general, a tensor in $\mathbb{R}^{d_1 \times \cdots \times d_k}$ can have rank exceeding $\min\{d_1, \dots, d_k\}$.

Tensor rank can jump over an arbitrarily large gap:

Theorem 2 (de Silva, L). Let $k \geq 3$. Given any $s \in \mathbb{N}$, there exists a sequence of order-k tensor A_n such that $\operatorname{rank}_{\otimes}(A_n) \leq r$ and $\lim_{n \to \infty} A_n = A$ with $\operatorname{rank}_{\otimes}(A) = r + s$.

Tensors that fail to have best low-rank approximations are not rare — they occur with non-zero probability:

Theorem 3 (de Silva, L). Let μ be a measure that is positive or infinite on Euclidean open sets in $\mathbb{R}^{d_1 \times \cdots \times d_k}$. There exists some $r \in \mathbb{N}$ such that

$$\mu(\{A \mid A \text{ does not have a best rank-}r\text{ approximation}\}) > 0.$$

All results apply verbatim to approximations of matrices by sums of Kronecker product of matrices, eg. recent work by G. Beylkin, W. Hackbush, B. Khoromskij, E. Tyrtyshnikov, C. Van Loan.

Message

That the best rank-r approximation problem for tensors has no solution poses serious difficulties.

It is 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?

Weak solutions

For a tensor A that has no best rank-r approximation, we will call a $C \in \overline{\{A \mid \operatorname{rank}_{\otimes}(A) \leq r\}}$ attaining

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

a weak solution. In particular, we must have $\operatorname{rank}_{\otimes}(C) > r$.

It is perhaps surprising that one may completely parameterize all limit points of order-3 rank-2 tensors:

Theorem 4 (de Silva, L.) Let $d_1, d_2, d_3 \geq 2$. Let $A_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a sequence of tensors with $\operatorname{rank}_{\otimes}(A_n) \leq 2$ and

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

where the limit is taken in any norm topology. If the limiting tensor A has rank higher than 2, then $\operatorname{rank}_{\otimes}(A)$ must be exactly 3 and there exist pairs of linearly independent vectors $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{R}^{d_1}$, $\mathbf{x}_2, \mathbf{y}_2 \in \mathbb{R}^{d_2}$, $\mathbf{x}_3, \mathbf{y}_3 \in \mathbb{R}^{d_3}$ such that

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

In particular, a sequence of order-3 rank-2 tensors cannot ‘jump rank’ by more than 1.

SURVEY OF OTHER RESULTS

- SYMMETRIC TENSORS
- NONNEGATIVE TENSORS
- ALGORITHMS
- MULTILINEAR SPECTRAL THEORY

Symmetric tensors

An order-k cubical tensor $\llbracket a_{i_1 \cdots i_k} \rrbracket \in \mathbb{R}^{n \times \cdots \times n}$ is **symmetric** if

$$a_{i_{\sigma(1)} \cdots i_{\sigma(k)}} = a_{i_1 \cdots i_k}, \qquad i_1, \ldots, i_k \in \{1, \ldots, n\},$$

for all permutations $\sigma \in \mathfrak{S}_k$. $\mathsf{S}^k(\mathbb{R}^n)$ is the set of all order-k symmetric tensors. Write $\mathbf{y}^{\otimes k} := \mathbf{y} \otimes \cdots \otimes \mathbf{y}$ (k times).

Examples. Higher order derivatives of multivariate functions. Moments and cumulants of random vector $\mathbf{x} = (X_1, \ldots, X_n)$:

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

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

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

Symmetric tensors, in the form of cumulants, are of particular importance in Independent Component Analysis (ICA).

Tensors for blind source separation

Problem: $\mathbf{y} = M\mathbf{x} + \mathbf{n}$. Goal: recover $\mathbf{x}$ from $\mathbf{y}$.

Unknown: source vector $\mathbf{x} \in \mathbb{C}^n$, mixing matrix $M \in \mathbb{C}^{m \times n}$, noise $\mathbf{n} \in \mathbb{C}^m$.

Known: observation vector $\mathbf{y} \in \mathbb{C}^m$.

Assumptions: components of $\mathbf{x}$ statistically independent, M full column-rank, $\mathbf{n}$ Gaussian.

Method: use cumulants

$$\kappa_k(\mathbf{y}) = (M, M, \dots, M) \cdot \kappa_k(\mathbf{x}) + \kappa_k(\mathbf{n}).$$

By assumptions, $\kappa_k(\mathbf{n}) = 0$ and $\kappa_k(\mathbf{x})$ is diagonal. So need to diagonalize the symmetric tensor $\kappa_k(\mathbf{y})$.

L. De Lathauwer, B. De Moor, and J. Vandewalle, “An introduction to independent component analysis,” *J. Chemometrics*, **14** (2000), no. 3, pp. 123-149.

Diagonalizing a symmetric tensor

Want to understand properties of symmetric rank, defined for $A \in \mathsf{S}^k(\mathbb{C}^n)$ as

$$\operatorname{rank}_{\mathsf{S}}(A) = \min\left\{r \;\middle|\; A = \textstyle\sum_{i=1}^{r} \alpha_i \mathbf{y}_i^{\otimes k}\right\}.$$

The definition is never vacuous because of the following:

Lemma (Comon, Golub, L, Mourrain). Let $A \in \mathsf{S}^k(\mathbb{C}^n)$. Then there exist $\mathbf{y}_1, \dots, \mathbf{y}_s \in \mathbb{C}^n$ such that

$$A = \textstyle\sum_{i=1}^{s} \alpha_i \mathbf{y}_i^{\otimes k}$$

A best symmetric rank approximation may not exist either:

Example (Comon, Golub, L, Mourrain). Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ be linearly independent. Define for $n \in \mathbb{N}$,

$$A_n := n\left(\mathbf{x} + \frac{1}{n}\mathbf{y}\right)^{\otimes k} - n\mathbf{x}^{\otimes k}$$

and

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

Then $\operatorname{rank}_{\mathsf{S}}(A_n) \leq 2$, $\operatorname{rank}_{\mathsf{S}}(A) = k$, and

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

Nonnegative tensors and nonnegative rank

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

$$\operatorname{rank}_+(A) := \min\Bigl\{r \Bigm| \textstyle\sum_{i=1}^r \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i,\ \mathbf{u}_i, \dots, \mathbf{z}_i \geq 0\Bigr\}$$

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

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

$$\inf\Bigl\{\Bigl\|A - \textstyle\sum_{i=1}^r \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i\Bigr\| \Bigm| \mathbf{u}_i, \dots, \mathbf{z}_i \geq 0\Bigr\}$$

is attained.

Corollary. The set $\{A \mid \operatorname{rank}_+(A) \leq r\}$ is closed.

NTF as naive Bayes model

Naive Bayes conditional independence assumption: $X_1, \ldots, X_k, H$ finitely supported discrete random variables such that $X_1, \ldots, X_k$ are statistically independent conditional on H, ie.

$$\Pr(X_1 = x_1, \ldots, X_k = x_k \mid H = h) = \textstyle\prod_{i=1}^{k} \Pr(X_i = x_i \mid H = h).$$

and so

$$\Pr(X_1 = x_{j_1}^{(1)}, \ldots, X_k = x_{j_k}^{(k)}) = \textstyle\sum_{i=1}^{r} \Pr(H = h_i) \prod_{\beta=1}^{k} \Pr(X_\beta = x_{j_\beta}^{(\beta)} \mid H = h_i).$$

Let $a_{j_1 \cdots j_k} = \Pr(X_1 = x_{j_1}^{(1)}, \ldots, X_k = x_{j_k}^{(k)})$, etc, get

$$A = \textstyle\sum_{i=1}^{r} \lambda_i \mathbf{u}_i^{(1)} \otimes \cdots \otimes \mathbf{u}_i^{(k)}.$$

$r = \operatorname{rank}_+(A)$ if H has minimal support over all such decompositions.

Algorithms

Even when an optimal solution B_* to $\operatorname{argmin}_{\text{rank}_\otimes(B)\le r}\|A-B\|_F$ exists, B_* is not easy to compute since the objective function is non-convex.

A widely used strategy is a nonlinear Gauss-Seidel algorithm, better known as the Alternating Least Squares algorithm:

Algorithm: ALS for optimal rank-r approximation

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

Convex relaxation

[with Kim-Chuan Toh]

$F(x_{11}, \ldots, z_{nr}) = \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$ is a polynomial.

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

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

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

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

Note rhs is a sum of squares iff $M - \lambda E_{11}$ is positive semi-definite (since $M - \lambda E_{11} = B^t B$). Get convex problem

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

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

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

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 $x^2y^2z^2$ or $uvwxyz$ appear). This can be exploited.

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

So if $f(x_{11}, \ldots, z_{nr}) = \sum_{j=1}^{N} p_j(x_{11}, \ldots, z_{nr})^2$, then only 1 and monomials of the form $x_{i\alpha}y_{j\alpha}z_{k\alpha}$ may occur in $p_1, \ldots, p_N$.

In other words, the complexity is really $rlmn + 1$ instead of $\binom{r(l+m+n)+3}{3}$.

Exploiting semiseparability

[with Ming Gu]

Gauss-Newton Method: $g(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2$. Approximate Hessian using Jacobian: $H_g \approx J_{\mathbf{f}}^t J_{\mathbf{f}}$.

The Hessian of $F(X, Y, Z) = \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$ can be approximated by a semiseparable matrix. This is the case even when X, Y, Z are required to be nonnegative.

Goal: Exploit this in optimization algorithms.

Multilinear spectral theory

Eigenvalues/vectors of symmetric A are critical values/points of Rayleigh quotient, $\mathbf{x}^t A\mathbf{x}/\|\mathbf{x}\|_2^2$. Similar characterization exists for singular values/vectors

For $\mathbf{x} = [x_1, \dots, x_n]^t \in \mathbb{R}^n$, write $\mathbf{x}^p := [x_1^p, \dots, x_n^p]^t$. Define the '$\ell^k$-norm' $\|\mathbf{x}\|_k = (x_1^k + \cdots + x_n^k)^{1/k}$.

Define eigenvalues/vectors of $A \in \mathsf{S}^k(\mathbb{R}^n)$ as critical values/points of the multilinear Rayleigh quotient $A(\mathbf{x}, \dots, \mathbf{x})/\|\mathbf{x}\|_k^k$.

$$A(I_n, \mathbf{x}, \dots, \mathbf{x}) = \lambda \mathbf{x}^{k-1}$$

Note that for a symmetric tensor A,

$$A(I_n, \mathbf{x}, \mathbf{x}, \dots, \mathbf{x}) = A(\mathbf{x}, I_n, \mathbf{x}, \dots, \mathbf{x}) = \cdots = A(\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}, I_n).$$

These equations have also been obtained by L. Qi independently using a different approach.

Perron-Frobenius theorem for nonnegative tensors

An order-k cubical tensor $A \in \mathsf{T}^k(\mathbb{R}^n)$ is *reducible* if there exist a permutation $\sigma \in \mathfrak{S}_n$ such that the permuted tensor

$$[\![b_{i_1 \cdots i_k}]\!] = [\![a_{\sigma(j_1) \cdots \sigma(j_k)}]\!]$$

has the property that for some $m \in \{1, \dots, n-1\}$, $b_{i_1 \cdots i_k} = 0$ for all $i_1 \in \{1, \dots, n-m\}$ and all $i_2, \dots, i_k \in \{1, \dots, m\}$. We say that A is *irreducible* if it is not reducible. In particular, if $A > 0$, then it is irreducible.

Theorem (L). Let $0 \leq A = [\![a_{j_1 \cdots j_k}]\!] \in \mathsf{T}^k(\mathbb{R}^n)$ be irreducible. Then A has a positive real eigenvalue μ with an eigenvector $\mathbf{x}$ that may be chosen to have all entries non-negative. Furthermore, μ is simple, ie. $\mathbf{x}$ is unique modulo scalar multiplication.

Very basic spectral hypergraph theory

Define the order-3 adjacency tensor A by

$$A_{xyz} = \begin{cases} 1 & \text{if } [x, y, z] \in E, \\ 0 & \text{otherwise.} \end{cases}$$

A is $|V|$-by-$|V|$-by-$|V|$ nonnegative symmetric tensor.

Consider cubic form $A(f, f, f) = \sum_{x,y,z} A_{xyz} f(x) f(y) f(z)$ (f is a vector of dimension $|V|$). Look at eigenvalues/vectors of A, ie. critical values/points of $A(f, f, f)$ constrained to $\sum_x f(x)^3 = 1$.

Lemma (L). Let G be an m-regular 3-hypergraph and A be its adjacency tensor. Then

(a) m is an eigenvalue of A;

(b) if μ is an eigenvalue of A, then $|\mu| \leq m$;

(c) μ has multiplicity 1 if and only if G is connected.

A hypergraph $G = (V, E)$ is said to be k-partite or k-colorable if there exists a partition of the vertices $V = V_1 \cup \cdots \cup V_k$ such that for any k vertices $u, v, \ldots, z$ with $A_{uv\cdots z} \neq 0$, $u, v, \ldots, z$ must each lie in a distinct V_i $(i = 1, \ldots, k)$.

Lemma (L). Let G be a connected m-regular k-partite k-hypergraph on n vertices. Then

(a) If k is odd, then every eigenvalue of G occurs with multiplicity a multiple of k.

(b) If k is even, then the spectrum of G is symmetric (ie. if μ is an eigenvalue, then so is $-\mu$). Furthermore, every eigenvalue of G occurs with multiplicity a multiple of $k/2$. If μ is an eigenvalue of G, then μ and $-\mu$ occurs with the same multiplicity.

WORK IN PROGRESS

- ALGEBRAIC STATISTICS
- SIMULTANEOUS EIGENVECTORS
- LOW-RANK COCLUSTERING

Algebraic statistics in computational biology

[with Bernd Sturmfels]

Problem: Find the polynomial equations that defines the set

$$\{P \in \mathbb{C}^{4\times4\times4} \mid \operatorname{rank}_{\otimes} P \leq 4\}.$$

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 $\mathbb{R}_+$.

Simultaneous eigenvectors in bioinformatics

[with Orly Alter and Bernd Sturmfels]

Background (Alter, Brown, Botstein): $(A_1, A_2) \in \mathbb{R}^{2\times m\times n}$ matrix pencil representing gene × microarray matrix for two organisms. Generalized SVD allows for a simultaneous factorization

$$A_1 = U_1\Sigma_1X^{-1}, \qquad A_2 = U_2\Sigma_2X^{-1}$$

which in turn allows for comparison of decoupled arraylets given by columns of orthogonal matrices U_1 and U_2.

Goal: Extend to $(A_1, A_2, \ldots, A_l) \in \mathbb{R}^{l\times m\times n}$ by finding 'approximate' simultaneous eigenvectors of $(A_1A_1^t, A_2A_2^t, \ldots, A_lA_l^t) \in \mathbb{R}^{l\times m\times m}$.