Multilinear Algebra in Data Analysis:

tensors, symmetric tensors, nonnegative tensors

Lek-Heng Lim
Stanford University

Workshop on Algorithms for Modern Massive Datasets
Stanford, CA
June 21–24, 2006

Thanks: G. Carlsson, L. De Lathauwer, J.M. Landsberg, M. Mahoney, L. Qi, B. Sturmfels; Collaborators: P. Comon, V. de Silva, P. Drineas, G. Golub

References

`http://www-sccm.stanford.edu/nf-publications-tech.html`

[**CGLM2**] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain, "Symmetric tensors and symmetric tensor rank," *SCCM Tech. Rep.*, **06-02**, 2006.

[**CGLM1**] P. Comon, B. Mourrain, L.-H. Lim, and G.H. Golub, "Genericity and rank deficiency of high order symmetric tensors," *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **31** (2006), no. 3, pp. 125–128.

[**dSL**] V. de Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SCCM Tech. Rep.*, **06-06** (2006).

[**GL**] G. Golub and L.-H. Lim, "Nonnegative decomposition and approximation of nonnegative matrices and tensors," *SCCM Tech. Rep.*, **06-01** (2006), forthcoming.

[**L**] L.-H. Lim, "Singular values and eigenvalues of tensors: a variational approach," *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, **1** (2005), pp. 129–132.

What is not a tensor, I

- **What is a vector?**
  - Mathematician: An element of a vector space.
  - Physicist: "What kind of physical quantities can be represented by vectors?"
    Answer: Once a basis is chosen, an n-dimensional vector is something that is represented by n real numbers only if those real numbers transform themselves as expected (ie. according to the change-of-basis theorem) when one changes the basis

- **What is a tensor?**
  - Mathematician: An element of a tensor product of vector spaces.
  - Physicist: "What kind of physical quantities can be represented by tensors?" **Answer:** Slide 7.

What is not a tensor, II

By a tensor, physicists and geometers often mean a tensor field (roughly, these are tensor-valued functions on manifolds):

- stress tensor
- moment-of-intertia tensor
- gravitational field tensor
- metric tensor
- curvature tensor

Tensor product of vector spaces

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

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

Tensors and multiway arrays

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

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

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

2. Multilinear Matrix Multiplication: for matrices $L = [\lambda_{i'i}] \in \mathbb{R}^{p \times l}, M = [\mu_{j'j}] \in \mathbb{R}^{q \times m}, N = [\nu_{k'k}] \in \mathbb{R}^{r \times n}$,

   $$(L, M, N) \cdot A := [\![c_{i'j'k'}]\!] \in \mathbb{R}^{p \times q \times r}$$

   where

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

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).

Henceforth, we will not distinguish between an order-k tensor and a k-way array that represents it (with respect to some implicit choice of basis).

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:

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

$\boxed{\text{Tensor ranks}}$

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

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

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

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

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

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

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

Credit

Both notions of tensor rank (also the corresponding decomposition) due to Frank L. Hitchcock in 1927. Multilinear rank is a special case (uniplex) of his more general multiplex rank.

F.L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *J. Math. Phys.*, **6** (1927), no. 1, pp. 164–189.

F.L. Hitchcock, "Multiple invariants and generalized rank of a p-way matrix or tensor," *J. Math. Phys.*, **7** (1927), no. 1, pp. 39–79.

Often wrongly attributed. Predates CANDECOMP/PARAFAC or Tucker by 40 years.

Outer product rank

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.

Example. $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ linearly independent and let $\mathbf{z} = \mathbf{x} + i\mathbf{y}$.

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

$\operatorname{rank}_{\otimes}(A)$ is 3 over $\mathbb{R}$ and is 2 over $\mathbb{C}$.

Fundamental problem of multiway data analysis

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

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

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

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

Harmonic analytic approach to data analysis

More generally, $\mathbb{F} = \mathbb{C}$, $\mathbb{R}$, $\mathbb{R}_+, \mathbb{R}_{\text{max}}$ (max-plus algebra), $\mathbb{R}[x_1, \ldots, x_n]$ (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:

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

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 CANDECOMP/PARAFAC.

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 decomposition.

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 $\mathcal{S}_r(d_1, \ldots, d_k) := \{A \mid \operatorname{rank}_\otimes(A) \leq r\}$ 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

D. Bini, M. Capovani, F. Romani, and G. Lotti, "$O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication," *Inform. Process. Lett.*, **8** (1979), no. 5, pp. 234–235.

Let $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$ be linearly independent. Define

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

and, for $\varepsilon > 0$,

$$\begin{aligned} B_\varepsilon := (\mathbf{y}+\varepsilon\mathbf{x}) \otimes (\mathbf{y}+\varepsilon\mathbf{w}) \otimes \varepsilon^{-1}\mathbf{z} + (\mathbf{z}+\varepsilon\mathbf{x}) \otimes \varepsilon^{-1}\mathbf{x} \otimes (\mathbf{x}+\varepsilon\mathbf{y}) \\ - \varepsilon^{-1}\mathbf{y} \otimes \mathbf{y} \otimes (\mathbf{x}+\mathbf{z}+\varepsilon\mathbf{w}) - \varepsilon^{-1}\mathbf{z} \otimes (\mathbf{x}+\mathbf{y}+\varepsilon\mathbf{z}) \otimes \mathbf{x} \\ + \varepsilon^{-1}(\mathbf{y}+\mathbf{z}) \otimes (\mathbf{y}+\varepsilon\mathbf{z}) \otimes (\mathbf{x}+\varepsilon\mathbf{w}). \end{aligned}$$

Then $\operatorname{rank}_\otimes(B_\varepsilon) \leq 5$, $\operatorname{rank}_\otimes(A) = 6$ and $\|B_\varepsilon - A\| \to 0$ as $\varepsilon \to 0$.

A has no optimal approximation by tensors of rank ≤ 5.

Simpler example

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

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

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

Furthermore

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 $\min\{d_1, \dots, d_k\}$.

$\boxed{\textbf{Furthermore}}$

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 $\operatorname{rank}_{\otimes}(A) = r + s$.

$\boxed{\textbf{Furthermore}}$

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

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

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.

Symmetric tensors

Write $\mathsf{T}^k(\mathbb{R}^n) = \mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n = \mathbb{R}^{n\times\cdots\times n}$, the set of all order-$k$ dimension-n cubical tensors.

An order-k cubical tensor $[\![a_{i_1\cdots i_k}]\!] \in \mathsf{T}^k(\mathbb{R}^n)$ is called **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 $\sigma \in \mathfrak{S}_k$.

These are order-k generalization of symmetric matrices. They are often mistakenly called 'supersymmetric tensors'.

Write $\mathsf{S}^k(\mathbb{R}^n)$ for the set of all order-k symmetric tensors. Write

$$\mathbf{y}^{\otimes k} := \overbrace{\mathbf{y}\otimes\cdots\otimes\mathbf{y}}^{k \text{ copies}}$$

Examples. higher order derivatives of smooth functions, moments and cumulants of random vectors.

Cumulants

$X_1, \ldots, X_n$ random variables. Moments and cumulants of $\mathbf{X} = (X_1, \ldots, X_n)$ are

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

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

For $n = 1$, $\kappa_k(X)$ for $k = 1, 2, 3, 4$ are the expectation, variance, skewness, and kurtosis of the random variable X respectively.

Symmetric tensors, in the form of cumulants, are of particular importance in Independent Component Analysis. Good read:

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

Symmetric $\otimes$ decomposition and symmetric rank

Let $A \in \mathsf{S}^k(\mathbb{R}^n)$. Define the symmetric rank of A as

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

The definition is never vacuous because of the following:

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

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

Question: given $A \in \mathsf{S}^k(\mathbb{R}^n)$, is $\operatorname{rank}_{\mathsf{S}}(A) = \operatorname{rank}_{\otimes}(A)$?
Partial answer: yes in many instances (cf. [**CGLM2**]).

Non-existence of best low-symmetric-rank approximation

Example (Comon, Golub, L, Mourrain). Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ 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

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

ie. symmetric rank can jump over an arbitrarily large gap too.

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.

NMD as an ℓ^1-SVD

$A \in \mathbb{R}^{m\times n}$. The SVD of A is, in particular, an expression

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

$r = \operatorname{rank}(A)$ is the minimal number where such a decomposition is possible,

$$\|\boldsymbol{\sigma}\|_2 = \left(\sum_{i=1}^{r} |\sigma_i|^2\right)^{1/2} = \|A\|_F, \quad \text{and} \quad \|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1,$$

for $i = 1, \ldots, r$.

Lemma (Golub, L). Let $0 \leq A \in \mathbb{R}^{m\times n}$, there exist $\mathbf{u}_i, \mathbf{v}_i \geq 0$ such that

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

$r = \operatorname{rank}_+(A)$ is the minimal number where such a decomposition is possible,

$$\|\boldsymbol{\lambda}\|_1 = \sum_{i=1}^{r} |\lambda_i| = \|A\|_G, \quad \text{and} \quad \|\mathbf{u}_i\|_1 = \|\mathbf{v}_i\|_1 = 1,$$

for $i = 1, \ldots, r$. The G-norm of A,

$$\|A\|_G = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|,$$

is the ℓ^1-equivalent of the F-norm

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}.$$

The NMD, viewed in the light of an ℓ^1-SVD, will be called an ℓ^1-NMD.

ℓ^1-nonnegative tensor decomposition

The SVD of a matrix does not generalize to tensors in any obvious way. The ℓ^1-NMD, however, generalizes to nonnegative tensors easily.

Lemma (Golub, L). Let $0 \leq A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. Then there exist $\mathbf{u}_i, \mathbf{v}_i, \ldots, \mathbf{z}_i \geq 0$ such that

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

$r = \operatorname{rank}_+(A)$ is the minimal number where such a decomposition is possible,

$$\|\boldsymbol{\lambda}\|_1 = \|A\|_G, \quad \text{and} \quad \|\mathbf{u}_i\|_1 = \|\mathbf{v}_i\|_1 = \cdots = \|\mathbf{z}_i\|_1 = 1$$

for $i = 1, \ldots, r$. Here

$$\|A\|_G := \textstyle\sum_{i_1,\ldots,i_k=1}^{n} |a_{i_1 \cdots i_k}|.$$

Naive Bayes model

Let $X_1, X_2, \ldots, X_k, H$ be finitely supported discrete random variables be such that

$$X_1, X_2, \ldots, X_k \text{ are statistically independent conditional on } H$$

or, in notation, $(X_1 \perp X_2 \perp \cdots \perp X_k) \parallel H$. In other words, the probability densities satisfy

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \mid H = h) = \prod_{i=1}^{k} \Pr(X_i = x_i \mid H = h).$$

This is called the Naive Bayes conditional independence assumption.

ℓ^1-NTD and Naive Bayes model

For $\beta = 1, \dots, k$, let support of X_β be $\{x_1^{(\beta)}, \dots, x_{d_\beta}^{(\beta)}\}$ and support of H be $\{h_1, \dots, h_r\}$. Marginal probability density is then

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

Let $a_{j_1 \cdots j_k} = \Pr(X_1 = x_{j_1}^{(1)}, \dots, X_k = x_{j_k}^{(k)})$, $u_{i,j_\beta}^{(\beta)} = \Pr(X_\beta = x_{j_\beta}^{(\beta)} \mid H = h_i)$, $\lambda_i = \Pr(H = h_i)$. We get

$$a_{j_1 \cdots j_k} = \sum_{p=1}^{r} \lambda_p \prod_{\beta=1}^{k} u_{p,j_\beta}^{(\beta)}.$$

Set $A = [\![a_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, $\mathbf{u}_i^{(\beta)} = [u_{i,1}^{(\beta)}, \dots, u_{i,d_\beta}^{(\beta)}]^\mathsf{T} \in \mathbb{R}^{d_\beta}$, $\beta = 1, \dots, k$, to get

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

Note that the quantities $A, \boldsymbol{\lambda}, \mathbf{u}_i^{(\beta)}$, being probability densities values, must satisfy

$$\|\boldsymbol{\lambda}\|_1 = \|A\|_G = \|\mathbf{u}_i\|_1 = \|\mathbf{v}_i\|_1 = \cdots = \|\mathbf{z}_i\|_1 = 1.$$

By earlier lemma, this is always possible for any non-negative tensor, provided that we first normalize A by $\|A\|_G$.

ℓ^1-NTD as a graphical model/Bayesian network

Corollary (Golub, L). Given $0 \leq A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$, there exist $X_1, X_2, \ldots, X_k, H$ finitely supported discrete random variables in a Naive Bayes model, $(X_1 \perp X_2 \perp \cdots \perp X_k) \parallel H$, such that its marginal-conditional decomposition is precisely the NTD of $A/\|A\|_G$. Furthermore. the support of H is minimal over all such admissible models.

Remark. This is prompted by a more high-brow algebraic geometric approach relating the Naive Bayes model with secant varieties of Segre variety in projective spaces:

L.D. Garcia, M. Stillman and B. Sturmfels, "Algebraic geometry of Bayesian networks," *J. Symbolic Comp.*, **39** (2005), no. 3–4, pp. 331–355.

Variational approach to eigen/singular values/vectors

A symmetric matrix. Eigenvalues/vectors are critical values/points of Rayleigh quotient, $\mathbf{x}^\mathsf{T} A\mathbf{x}/\|\mathbf{x}\|_2^2$, or equivalently, the critical values/points of quadratic form $\mathbf{x}^\mathsf{T} A\mathbf{x}$ constrained to vectors with unit l^2-norm, $\{\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$. Associated Lagrangian,

$$L(\mathbf{x}, \lambda) = \mathbf{x}^\mathsf{T} A\mathbf{x} - \lambda(\|\mathbf{x}\|_2^2 - 1).$$

Vanishing of ∇L at a critical point $(\mathbf{x}_c, \lambda_c) \in \mathbb{R}^n \times \mathbb{R}$ yields the familiar

$$A\mathbf{x}_c = \lambda_c \mathbf{x}_c.$$

$A \in \mathbb{R}^{m\times n}$. Singular values/vectors may likewise be obtained with $\mathbf{x}^\mathsf{T} A\mathbf{y}/\|\mathbf{x}\|_2\|\mathbf{y}\|_2$ playing the role of the Rayleigh quotient. Associated Lagrangian function now

$$L(\mathbf{x}, \mathbf{y}, \sigma) = \mathbf{x}^\mathsf{T} A\mathbf{y} - \sigma(\|\mathbf{x}\|_2\|\mathbf{y}\|_2 - 1).$$

At a critical point $(\mathbf{x}_c, \mathbf{y}_c, \sigma_c) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$,

$$A\mathbf{y}_c/\|\mathbf{y}_c\|_2 = \sigma_c \mathbf{x}_c/\|\mathbf{x}_c\|_2, \quad A^\mathsf{T}\mathbf{x}_c/\|\mathbf{x}_c\|_2 = \sigma_c \mathbf{y}_c/\|\mathbf{y}_c\|_2.$$

Write $\mathbf{u}_c = \mathbf{x}_c/\|\mathbf{x}_c\|_2$ and $\mathbf{v}_c = \mathbf{y}_c/\|\mathbf{y}_c\|_2$ to get the familiar

$$A\mathbf{v}_c = \sigma_c \mathbf{u}_c, \quad A^\mathsf{T}\mathbf{u}_c = \sigma_c \mathbf{v}_c.$$

Multilinear spectral theory

May extend the variational approach to tensors to obtain a theory of eigen/singular values/vectors for tensors (cf. [**L**] for details).

For $\mathbf{x} = [x_1, \dots, x_n]^\mathsf{T} \in \mathbb{R}^n$, write

$$\mathbf{x}^p := [x_1^p, \dots, x_n^p]^\mathsf{T}.$$

We also define the 'ℓ^k-norm'

$$\|\mathbf{x}\|_k = (x_1^k + \cdots + x_n^k)^{1/k}.$$

Define ℓ^2- and ℓ^k-eigenvalues/vectors of $A \in \mathsf{S}^k(\mathbb{R}^n)$ as the critical values/points of the multilinear Rayleigh quotient $A(\mathbf{x}, \dots, \mathbf{x})/\|\mathbf{x}\|_p^k$. Differentiating the Lagrangian

$$L(\mathbf{x}_1, \dots, \mathbf{x}_k, \sigma) := A(\mathbf{x}_1, \dots, \mathbf{x}_k) - \sigma(\|\mathbf{x}_1\|_{p_1} \cdots \|\mathbf{x}_k\|_{p_k} - 1).$$

yields

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

and

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

respectively. Note that for a symmetric tensor A,

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

This doesn't hold for nonsymmetric cubical tensors $A \in \mathsf{S}^k(\mathbb{R}^n)$ and we get different eigenpair for different modes (this is to be expected: even for matrices, a nonsymmetric matrix will have different left/right eigenvectors).

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 l^k-eigenvalue μ with an l^k-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.

Hypergraphs

For notational simplicity, the following is stated for a 3-hypergraph but it generalizes to k-hypergraphs for any k.

$G = (V, E)$ be a 3-hypergraph. V is the finite set of vertices and E is the subset of hyperedges, ie. 3-element subsets of V. We write the elements of E as $[x, y, z]$ $(x, y, z \in V)$.

G is undirected, so $[x, y, z] = [y, z, x] = \cdots = [z, y, x]$. A hyperedge is said to degenerate if it is of the form $[x, x, y]$ or $[x, x, x]$ (hyperloop at x). We do not exclude degenerate hyperedges.

G is m-regular if every $v \in V$ is adjacent to exactly m hyperedges. We can 'regularize' a non-regular hypergraph by adding hyperloops.

Adjacency tensor of a hypergraph

Define the order-3 adjacency tensor A by

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

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

Consider cubic form $A(f, f, f) = \sum_{x,y,z} A_{xyz} f(x) f(y) f(z)$ (note that f is a vector of dimension $|V|$).

Call critical values and critical points of $A(f, f, f)$ constrained to the set $\sum_x f(x)^3 = 1$ (like the ℓ^3-norm except we do not take absolute value) the ℓ^3-eigenvalues and ℓ^3-eigenvectors of A respectively.

Very basic spectral hypergraph theory I

As in the case of spectral graph theory, combinatorial/topological properties of a k-hypergraph may be deduced from ℓ^k-eigenvalues of its adjacency tensor (henceforth, in the context of a k-hypergraph, an eigenvalue will always mean an ℓ^k-eigenvalue).

Straightforward generalization of a basic result in spectral graph theory:

Theorem (Drineas, 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.

Very basic spectral hypergraph theory II

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 (Drineas, 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.