Numerical Multilinear Algebra II

Lek-Heng Lim

University of California, Berkeley

January 5–7, 2009

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

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

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

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

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

Matrix EVD and SVD

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

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

  where $\operatorname{rank}(A) = r$, $V \in \mathsf{O}(n)$ eigenvectors, Λ eigenvalues.
- **Singular value decomposition** of $A \in \mathbb{R}^{m \times n}$,

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

  where $\operatorname{rank}(\mathcal{A}) = r$, $U \in \mathsf{O}(m)$ left singular vectors, $V \in \mathsf{O}(n)$ right singular vectors, Σ singular values.
- Ditto for **nonnegative matrix decomposition**.

One plausible EVD and SVD for hypermatrices

- Rank revealing decompositions associated with the outer product rank.
- **Symmetric outer product decomposition** of $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$,

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

  where $\operatorname{rank}_{\mathsf{S}}(A) = r$, $\mathbf{v}_i$ unit vector, $\lambda_i \in \mathbb{R}$.
- **Outer product decomposition** of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

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

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

Another plausible EVD and SVD for hypermatrices

- Rank revealing decompositions associated with the multilinear rank.
- **Singular value decomposition** of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

$$\mathcal{A} = (U, V, W) \cdot \mathcal{C}$$

  where $\operatorname{rank}_{\boxplus}(A) = (r_1, r_2, r_3)$, $U \in \mathbb{R}^{l \times r_1}$, $V \in \mathbb{R}^{m \times r_2}$, $W \in \mathbb{R}^{n \times r_3}$ have orthonormal columns and $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.
- **Symmetric eigenvalue decomposition** of $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$,

$$\mathcal{A} = (U, U, U) \cdot \mathcal{C}$$

  where $\operatorname{rank}_{\boxplus}(A) = (r, r, r)$, $U \in \mathbb{R}^{n \times r}$ has orthonormal columns and $\mathcal{C} \in \mathsf{S}^3(\mathbb{R}^r)$.
- Ditto for **nonnegative multilinear decomposition**.

Outer product rank is hard to compute

- Eugene L. Lawler: "The Mystical Power of Twoness."
  - 2-SAT is easy, 3-SAT is hard;
  - 2-dimensional matching is easy, 3-dimensional matching is hard;
  - Order-2 tensor rank is easy, order-3 tensor rank is hard.

Theorem (Håstad)

Computing $\operatorname{rank}_{\otimes}(\mathcal{A})$ *for* $\mathcal{A} \in \mathbb{F}^{l \times m \times n}$ *is NP-hard for* $\mathbb{F} = \mathbb{Q}$ *and NP-complete for* $\mathbb{F} = \mathbb{F}_q$.

- **Open question:** Is tensor rank NP-hard/NP-complete over $\mathbb{F} = \mathbb{R}, \mathbb{C}$ in the sense of BCSS?
  - L. Blum, F. Cucker, M. Shub, S. Smale, *Complexity and real computation*, Springer-Verlag, New York, NY, 1998.

Outer product rank depends on base field

For $A \in \mathbb{R}^{m \times n} \subset \mathbb{C}^{m \times n}$, $\operatorname{rank}_{\mathbb{R}}(A) = \operatorname{rank}_{\mathbb{C}}(A)$. Not true for tensors.

Theorem (Bergman)

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

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

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

- May show that $\operatorname{rank}_{\otimes,\mathbb{R}}(\mathcal{A}) = 3$ and $\operatorname{rank}_{\otimes,\mathbb{C}}(\mathcal{A}) = 2$.
- $\mathbb{R}^{2 \times 2 \times 2}$ has 8 distinct orbits under $\mathrm{GL}_2(\mathbb{R}) \times \mathrm{GL}_2(\mathbb{R}) \times \mathrm{GL}_2(\mathbb{R})$.
- $\mathbb{C}^{2 \times 2 \times 2}$ has 7 distinct orbits under $\mathrm{GL}_2(\mathbb{C}) \times \mathrm{GL}_2(\mathbb{C}) \times \mathrm{GL}_2(\mathbb{C})$.

Outer product decomposition: separation of variables

Approximation by sum or integral of separable functions

- Continuous

$$f(x, y, z) = \int \theta(x, t)\varphi(y, t)\psi(z, t)\, dt.$$

- Semi-discrete

$$f(x, y, z) = \sum\nolimits_{p=1}^{r} \theta_p(x)\varphi_p(y)\psi_p(z)$$

$\theta_p(x) = \theta(x, t_p)$, $\varphi_p(y) = \varphi(y, t_p)$, $\psi_p(z) = \psi(z, t_p)$, r possibly ∞.

- Discrete

$$a_{ijk} = \sum\nolimits_{p=1}^{r} u_{ip} v_{jp} w_{kp}$$

$a_{ijk} = f(x_i, y_j, z_k)$, $u_{ip} = \theta_p(x_i)$, $v_{jp} = \varphi_p(y_j)$, $w_{kp} = \psi_p(z_k)$.

Separation of variables

- Useful for data analysis, machine learning, pattern recognition.
- Gaussians are separable

$$\exp(x^2 + y^2 + z^2) = \exp(x^2)\exp(y^2)\exp(z^2).$$

- More generally for symmetric positive-definite $A \in \mathbb{R}^{n\times n}$,

$$\exp(\mathbf{x}^\top A\mathbf{x}) = \exp(\mathbf{z}^\top \Lambda\mathbf{z}) = \prod\nolimits_{i=1}^{n} \exp(\lambda_i z_i^2).$$

- Gaussian mixture models

$$f(\mathbf{x}) = \sum\nolimits_{j=1}^{m} \alpha_j \exp[(\mathbf{x} - \boldsymbol{\mu}_j)^\top A_j(\mathbf{x} - \boldsymbol{\mu}_j)],$$

  f is a sum of separable functions.

Multilinear decomposition: integral kernels

Approximation by sum or integral kernels

- Continuous

$$f(x,y,z) = \iiint K(x',y',z')\theta(x,x')\varphi(y,y')\psi(z,z')\,dx'dy'dz'.$$

- Semi-discrete

$$f(x,y,z) = \sum\nolimits_{i',j',k'=1}^{p,q,r} c_{i'j'k'}\theta_{i'}(x)\varphi_{j'}(y)\psi_{k'}(z)$$

$c_{i'j'k'} = K(x'_{i'}, y'_{j'}, z'_{k'})$, $\theta_{i'}(x) = \theta(x, x'_{i'})$, $\varphi_{j'}(y) = \varphi(y, y'_{j'})$, $\psi_{k'}(z) = \psi(z, z'_{k'})$, p, q, r possibly ∞.

- Discrete

$$a_{ijk} = \sum\nolimits_{i',j',k'=1}^{p,q,r} c_{i'j'k'} u_{ii'} v_{jj'} w_{kk'}$$

$a_{ijk} = f(x_i, y_j, z_k)$, $u_{ii'} = \theta_{i'}(x_i)$, $v_{jj'} = \varphi_{j'}(y_j)$, $w_{kk'} = \psi_{k'}(z_k)$.

Best r-term approximation

$$f \approx \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_r f_r.$$

- **Target function** $f \in \mathcal{H}$ vector space, cone, etc.
- $f_1, \dots, f_r \in \mathscr{D} \subset \mathcal{H}$ **dictionary**.
- $\alpha_1, \dots, \alpha_r \in \mathbb{R}$ or $\mathbb{C}$ (linear), $\mathbb{R}_+$ (convex), $\mathbb{R} \cup \{-\infty\}$ (tropical).
- $\approx$ with respect to $\varphi : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, some measure of 'nearness' between pairs of points (e.g. norms, metric, volumes, expectation, entropy, Brègman divergences, etc), want

$$\operatorname{argmin}\{\varphi(f, \alpha_1 f_1 + \dots \alpha_r f_r) \mid f_i \in \mathscr{D}\}.$$

- For concreteness, $\mathcal{H}$ separable Hilbert space; measure of nearness is a norm, but not necessarily the one induced by its inner product.
- Reference: various papers by A. Cohen, R. DeVore, V. Temlyakov.

Dictionaries

- Number base: $\mathscr{D} = \{10^n \mid n \in \mathbb{Z}\} \subseteq \mathbb{R}$,

$$\tfrac{22}{7} = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 2 \cdot 10^{-3} + \cdots$$

- Spanning set: $\mathscr{D} = \left\{ \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right] \right\} \subseteq \mathbb{R}^2$,

$$\left[\begin{smallmatrix} 2 \\ -3 \end{smallmatrix}\right] = 3\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] - 1\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right].$$

- Taylor: $\mathscr{D} = \{x^n \mid n \in \mathbb{N} \cup \{0\}\} \subseteq C^\omega(\mathbb{R})$,

$$\exp(x) = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{6}x^3 + \cdots$$

- Fourier: $\mathscr{D} = \{\cos(nx), \sin(nx) \mid n \in \mathbb{Z}\} \subseteq L^2(-\pi, \pi)$,

$$\tfrac{1}{2}x = \sin(x) - \tfrac{1}{2}\sin(2x) + \tfrac{1}{3}\sin(3x) - \cdots$$

- $\mathscr{D}$ orthonormal basis, Schauder basis, Hamel basis, Riesz basis, frames, a dense spanning set.

More dictionaries

- Discrete cosine:

$$\mathscr{D} = \left\{ \sqrt{\tfrac{2}{N}} \cos(k + \tfrac{1}{2})(n + \tfrac{1}{2})\tfrac{\pi}{N} \;\middle|\; k \in [N-1] \right\} \subseteq \mathbb{C}^N.$$

- Peter-Weyl:

$$\mathscr{D} = \{\langle \pi(x)\mathbf{e}_i, \mathbf{e}_j \rangle \mid \pi \in \widehat{G}, i, j \in [d_\pi]\} \subseteq L^2(G).$$

- Paley-Wiener:

$$\mathscr{D} = \{\operatorname{sinc}(x - n) \mid n \in \mathbb{Z}\} \subseteq H^2(\mathbb{R}).$$

- Gabor:

$$\mathscr{D} = \{e^{i\alpha nx} e^{-(x - m\beta)^2/2} \mid (m, n) \in \mathbb{Z} \times \mathbb{Z}\} \subseteq L^2(\mathbb{R}).$$

- Wavelet:

$$\mathscr{D} = \{2^{n/2}\psi(2^n x - m) \mid (m, n) \in \mathbb{Z} \times \mathbb{Z}\} \subseteq L^2(\mathbb{R}).$$

- Friends of wavelets: $\mathscr{D} \subseteq L^2(\mathbb{R}^2)$ beamlets, brushlets, curvelets, ridgelets, wedgelets, multiwavelets.

Approximants

Definition

Dictionary $\mathscr{D} \subset \mathcal{H}$. For $r \in \mathbb{N}$, the set of **r-term approximants** is

$$\Sigma_r(\mathscr{D}) := \left\{ \sum_{i=1}^{r} \alpha_i f_i \in \mathcal{H} \;\middle|\; \alpha_i \in \mathbb{C}, f_i \in \mathscr{D} \right\}.$$

Let $f \in \mathcal{H}$. The **error of r-term approximation** is

$$\sigma_n(f) := \inf_{g \in \Sigma_r(\mathscr{D})} \|f - g\|.$$

- Linear combination of two r-term approximants may have more than r non-zero terms.
- $\Sigma_r(\mathscr{D})$ not a subspace of $\mathcal{H}$. Hence **nonlinear approximation**.
- In contrast with usual (linear) approximation, ie.

$$\inf_{g \in \mathrm{span}(\mathscr{D})} \|f - g\|.$$

Small is beautiful

$$f \approx \sum_{i \in \mathscr{I} \subseteq \mathscr{D}} \alpha_i f_i$$

- Want good approximation, ie. $\|f - \sum_{i \in \mathscr{I} \subseteq \mathscr{D}} \alpha_i f_i\|$ small.
- Want sparse/concentrated representation, ie. $|\mathscr{I}|$ small.
- Sparsity depends on choice of $\mathscr{D}$.
  - $\mathscr{D}_{10} = \{10^n \mid n \in \mathbb{Z}\}, \mathscr{D}_3 = \{3^n \mid n \in \mathbb{Z}\} \subseteq \mathbb{R}$,

$$\begin{aligned} \tfrac{1}{3} &= [0.33333\cdots]_{10} = \textstyle\sum_{n=1}^{\infty} 3 \cdot 10^{-n} \\ &= [0.1]_3 = 1 \cdot 3^{-1}. \end{aligned}$$

  - $\mathscr{D}_{\text{fourier}} = \{\cos(nx), \sin(nx) \mid n \in \mathbb{Z}\}$,

$$\tfrac{1}{2}x = \sin(x) - \tfrac{1}{2}\sin(2x) + \tfrac{1}{3}\sin(3x) - \cdots .$$

  - $\mathscr{D}_{\text{taylor}} = \{x^n \mid n \in \mathbb{N} \cup \{0\}\}$,

$$\sin(x) = x - \tfrac{1}{6}x^3 + \tfrac{1}{120}x^5 - \cdots .$$

Bigger is better

- **Union of dictionaries:** allows for efficient (sparse) representation of different features
  - $\mathscr{D} = \mathscr{D}_{\text{fourier}} \cup \mathscr{D}_{\text{wavelets}}$,
  - $\mathscr{D} = \mathscr{D}_{\text{spikes}} \cup \mathscr{D}_{\text{sinusoids}} \cup \mathscr{D}_{\text{splines}}$,
  - $\mathscr{D} = \mathscr{D}_{\text{wavelets}} \cup \mathscr{D}_{\text{curvelets}} \cup \mathscr{D}_{\text{beamlets}} \cup \mathscr{D}_{\text{ridgelets}}$.
- $\mathscr{D}$ **overcomplete** or **redundant** dictionary. Trade off: computational complexity.
- **Rule of thumb:** the larger and more diverse the dictionary, the more efficient/sparser the representation.
- **Observation:** $\mathscr{D}$ above all zero dimensional (at most countably infinite).
- **Question:** What about dictionaries with a continuously varying families of functions?
- **Meta question:** Why should tensor folks care about this?

Recap: hypermatrices are functions on finite sets

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

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

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

Tensor approximations

- General tensor approximation.
  - Target function
    $$f : [l] \times [m] \times [n] \to \mathbb{R}.$$
  - Dictionary of **separable functions**,
    $$\mathscr{D}_{\otimes} = \{g : [l] \times [m] \times [n] \to \mathbb{R} \mid g(i,j,k) = \vartheta(i)\varphi(j)\psi(k)\},$$
    where $\vartheta : [l] \to \mathbb{R}$, $\varphi : [m] \to \mathbb{R}$, $\psi : [n] \to \mathbb{R}$.
- Symmetric tensor approximation.
  - Target function:
    $$f : [n] \times [n] \times [n] \to \mathbb{R}$$
    with $f(i,j,k) = f(j,i,k) = \cdots = f(k,j,i)$.
  - Dictionary of symmetric separable functions:
    $$\mathscr{D}_{\mathsf{S}} = \{g : [n] \times [n] \times [n] \to \mathbb{R} \mid g(i,j,k) = \vartheta(i)\vartheta(j)\vartheta(k)\},$$
    where $\vartheta : [l] \to \mathbb{R}$.

Tensor approximations

- Nonnegative tensor approximation.
  - Target function
$$f : [l] \times [m] \times [n] \to \mathbb{R}_+.$$
  - Dictionary of nonnegative separable functions,
$$\mathscr{D}_+ = \{g : [l] \times [m] \times [n] \to \mathbb{R}_+ \mid g(i, j, k) = \vartheta(i)\varphi(j)\psi(k)\},$$
where $\vartheta : [l] \to \mathbb{R}_+$, $\varphi : [m] \to \mathbb{R}_+$, $\psi : [n] \to \mathbb{R}_+$.

Segre variety and its secant varieties

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

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

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

Same thing different names

- rth secant (quasiprojective) variety of the Segre variety is the set of r term approximants.
- If $\mathscr{D} = \operatorname{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n)$, then

$$\Sigma_r(\mathscr{D}) = \{\mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \operatorname{rank}_\otimes(\mathcal{A}) \le r\}.$$

- Rank revealing matrix decompositions (non-unique: LU, QR, SVD):

$$\mathscr{D} = \{\mathbf{x}\mathbf{y}^\top \mid (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n\} = \{A \in \mathbb{R}^{m \times n} \mid \operatorname{rank}(A) \le 1\}.$$

- Often unique for tensors [Kruskal; 1977], [Sidiroupoulos, Bro; 2000]:
  - $\operatorname{spark}(\mathbf{x}_1, \dots, \mathbf{x}_r)$ = size of minimal linearly dependent subset [Donoho, Elad; 2003].
  - Decomposition $\mathcal{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i$ is unique up to scaling if

$$\operatorname{spark}(\mathbf{u}_1, \dots, \mathbf{u}_r) + \operatorname{spark}(\mathbf{v}_1, \dots, \mathbf{v}_r) + \operatorname{spark}(\mathbf{w}_1, \dots, \mathbf{w}_r) \ge 2r + 5.$$

Dictionaries of positive dimensions

- Neural networks:

$$\mathscr{D} = \{\sigma(\mathbf{w}^\top \mathbf{x} + w_0) \mid (w_0, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^n\}$$

  where $\sigma : \mathbb{R} \to \mathbb{R}$ sigmoid function, eg. $\sigma(x) = [1 + \exp(-x)]^{-1}$.

- Exponential:

$$\mathscr{D} = \{e^{-tx} \mid t \in \mathbb{R}_+\} \qquad \text{or} \qquad \mathscr{D} = \{e^{\tau x} \mid \tau \in \mathbb{C}\}.$$

- Outer product decomposition:

$$\begin{aligned}\mathscr{D} &= \{\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \mid (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n\} \\ &= \{\mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \operatorname{rank}_\otimes(\mathcal{A}) \leq 1\}.\end{aligned}$$

- Symmetric outer product decomposition:

$$\mathscr{D} = \{\mathbf{v} \otimes \mathbf{v} \otimes \mathbf{v} \mid \mathbf{v} \in \mathbb{R}^n\} = \{\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n) \mid \operatorname{rank}_\mathsf{S}(\mathcal{A}) \leq 1\}.$$

- Nonnegative outer product decomposition:

$$\begin{aligned}\mathscr{D} &= \{\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \mid (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{R}^l_+ \times \mathbb{R}^m_+ \times \mathbb{R}^n_+\} \\ &= \{\mathcal{A} \in \mathbb{R}^{l \times m \times n}_+ \mid \operatorname{rank}_+(\mathcal{A}) \leq 1\}.\end{aligned}$$

Recall: fundamental problem of multiway data analysis

- $\mathcal{A}$ hypermatrix, symmetric hypermatrix, or nonnegative hypermatrix. Want
$$\operatorname{argmin}_{\operatorname{rank}(\mathcal{B})\le r}\|\mathcal{A}-\mathcal{B}\|.$$
- $\operatorname{rank}(\mathcal{B})$ may be outer product rank, multilinear rank, symmetric rank (for symmetric hypermatrix), or nonnegative rank (nonnegative hypermatrix).

Example

Given $\mathcal{A}\in\mathbb{R}^{d_1\times d_2\times d_3}$, find $\sigma_i,\mathbf{u}_i,\mathbf{v}_i,\mathbf{w}_i$, $i=1,\dots,r$, that minimizes
$$\|\mathcal{A}-\sigma_1\mathbf{u}_1\otimes\mathbf{v}_1\otimes\mathbf{w}_1-\sigma_2\mathbf{u}_2\otimes\mathbf{v}_2\otimes\mathbf{w}_2-\cdots-\sigma_r\mathbf{u}_r\otimes\mathbf{v}_r\otimes\mathbf{w}_r\|$$
or $\mathcal{C}\in\mathbb{R}^{r_1\times r_2\times r_3}$ and $U\in\mathbb{R}^{d_1\times r_1}, V\in\mathbb{R}^{d_2\times r_2}, W\in\mathbb{R}^{d_3\times r_3}$, that minimizes
$$\|\mathcal{A}-(U,V,W)\cdot\mathcal{C}\|.$$

- May assume $\mathbf{u}_i,\mathbf{v}_i,\mathbf{w}_i$ unit vectors and U,V,W orthonormal columns.

Recall: fundamental problem of multiway data analysis

Example

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

$$\|\mathcal{A} - \lambda_1 \mathbf{u}_1^{\otimes k} - \lambda_2 \mathbf{u}_2^{\otimes k} - \cdots - \lambda_r \mathbf{u}_r^{\otimes k}\|$$

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

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

- May assume $\mathbf{u}_i$ unit vector and U orthonormal columns.

Best low rank approximation of a matrix

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

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

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

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

Theorem (Eckart–Young)

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

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

Then

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

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

Lemma

Let $r \geq 2$ and $k \geq 3$. Given the norm-topology on $\mathbb{R}^{d_1 \times \cdots \times d_k}$, the following statements are equivalent:

1. *The set* $\mathscr{S}_r(d_1, \dots, d_k) := \{\mathcal{A} \mid \operatorname{rank}_\otimes(\mathcal{A}) \leq r\}$ *is not closed.*
2. *There exists a sequence* $\mathcal{A}_n$, $\operatorname{rank}_\otimes(\mathcal{A}_n) \leq r$, $n \in \mathbb{N}$, *converging to* $\mathcal{B}$ *with* $\operatorname{rank}_\otimes(\mathcal{B}) > r$.
3. *There exists* $\mathcal{B}$, $\operatorname{rank}_\otimes(\mathcal{B}) > r$, *that may be approximated arbitrarily closely by hypermatrices of strictly lower rank, i.e.*

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

4. *There exists* $\mathcal{C}$, $\operatorname{rank}_\otimes(\mathcal{C}) > r$, *that does not have a best rank-r approximation, i.e.*

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

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

Non-existence of best low-rank approximation

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

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

- For $n \in \mathbb{N}$,

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

Lemma

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

$$\lim_{n\to\infty} \mathcal{A}_n = \mathcal{A}.$$

- Original result, in a slightly different form, due to:
  - D. Bini, G. Lotti, F. Romani, "Approximate solutions for the bilinear form computational problem," *SIAM J. Comput.*, **9** (1980), no. 4.

Outer product approximations are ill-behaved

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

Theorem

Let $k \geq 3$ and $d_1, \dots, d_k \geq 2$. For any s such that

$$2 \leq s \leq \min\{d_1, \dots, d_k\},$$

there exists $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ with $\operatorname{rank}_{\otimes}(\mathcal{A}) = s$ such that $\mathcal{A}$ 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 hypermatrix in $\mathbb{R}^{d_1 \times \cdots \times d_k}$ can have rank exceeding $\min\{d_1, \dots, d_k\}$.

Outer product approximations are ill-behaved

- Tensor rank can jump over an arbitrarily large gap:

Theorem

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

- Hypermatrices that fail to have best low-rank approximations are not rare. May occur with non-zero probability; sometimes with certainty.

Theorem

Let μ be a measure that is positive or infinite on Euclidean open sets in $\mathbb{R}^{l\times m\times n}$. There exists some $r \in \mathbb{N}$ such that

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

In $\mathbb{R}^{2\times 2\times 2}$, all rank-3 hypermatrices fail to have best rank-2 approximation.

Happens to symmetric tensors ...

- Approximation of a homogeneous polynomial by a sum of powers of linear forms (e.g. Independent Components Analysis).
- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ be linearly independent. Define for $n \in \mathbb{N}$,

$$A_n := n \left[\mathbf{x} + \frac{1}{n} \mathbf{y} \right]^{\otimes p} - n \mathbf{x}^{\otimes p}$$

- Define

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

- Then $\operatorname{rank}_{\mathsf{S}}(\mathcal{A}_n) \leq 2$, $\operatorname{rank}_{\mathsf{S}}(\mathcal{A}) \geq p$, and

$$\lim_{n \to \infty} \mathcal{A}_n = \mathcal{A}.$$

- See [Comon, Golub, L, Mourrain; 08] for details.

. . . and to operators . . .

- Approximation of an operator by a sum of Kronecker product of lower-dimensional operators (e.g. Numerical Operator Calculus).
- For linearly independent operators $P_i, Q_i : V_i \to W_i$, $i = 1, 2, 3$, let $\mathcal{D} : V_1 \otimes V_2 \otimes V_3 \to W_1 \otimes W_2 \otimes W_3$ be

$$\mathcal{D} := P_1 \otimes Q_2 \otimes Q_3 + Q_1 \otimes Q_2 \otimes P_3 + Q_1 \otimes Q_2 \otimes P_3.$$

- If finite-dimensional, then '$\otimes$' may be taken to be Kronecker product of matrices.
- For $n \in \mathbb{N}$,

$$\mathcal{D}_n := n \left[P_1 + \frac{1}{n} Q_1 \right] \otimes \left[P_2 + \frac{1}{n} Q_2 \right] \otimes \left[P_3 + \frac{1}{n} Q_3 \right] - n P_1 \otimes P_2 \otimes P_3.$$

- Then

$$\lim_{n \to \infty} \mathcal{D}_n = \mathcal{D}.$$

. . . and functions too

- Approximation of a multivariate function by a sum of separable functions (e.g. Approximation Theory).
- For linearly independent $\varphi_1, \psi_1 : X \to \mathbb{R}$, $\varphi_2, \psi_2 : Y \to \mathbb{R}$, $\varphi_3, \psi_3 : Z \to \mathbb{R}$, let $f : X \times Y \times Z \to \mathbb{R}$ be

$$f(x, y, z) := \varphi_1(x)\psi_2(y)\psi_3(z) + \psi_1(x)\psi_2(y)\varphi_3(z) + \psi_1(x)\psi_2(y)\varphi_3(z).$$

- For $n \in \mathbb{N}$,

$$\begin{aligned} f_n(x, y, z) :=& \\ & n\left[\varphi_1(x) + \frac{1}{n}\psi_1(x)\right]\left[\varphi_2(y) + \frac{1}{n}\psi_2(y)\right]\left[\varphi_3(z) + \frac{1}{n}\psi_3(z)\right] \\ & - n\varphi_1(x)\varphi_2(y)\varphi_3(z). \end{aligned}$$

- Then

$$\lim_{n\to\infty} f_n = f.$$

Message

- That the best rank-r approximation problem for hypermatrices 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? i.e. what is the 'approximate solution' an approximate of?

Weak solutions

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

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

  a **weak solution**. In particular, we must have $\operatorname{rank}_{\otimes}(\mathcal{C}) > r$.
- It is perhaps surprising that one may completely parameterize all limit points of order-3 rank-2 hypermatrices.

Weak solutions

Theorem

Let $d_1, d_2, d_3 \geq 2$*. Let* $\mathcal{A}_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *be a sequence of hypermatrices with* $\operatorname{rank}_{\otimes}(\mathcal{A}_n) \leq 2$ *and*

$$\lim_{n \to \infty} \mathcal{A}_n = \mathcal{A},$$

where the limit is taken in any norm topology. If the limiting hypermatrix $\mathcal{A}$ *has rank higher than* 2*, then* $\operatorname{rank}_{\otimes}(\mathcal{A})$ *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*

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

- In particular, a sequence of order-3 rank-2 hypermatrices cannot 'jump rank' by more than 1.

Conditioning of linear systems

- Let $A \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$. Suppose we want to solve system of linear equations $A\mathbf{x} = \mathbf{b}$.
- $\mathscr{M} = \{A \in \mathbb{R}^{n\times n} \mid \det(A) = 0\}$ is the manifold of ill-posed problems.
- $A \in \mathscr{M}$ iff $A\mathbf{x} = \mathbf{0}$ has nontrivial solutions.
- Note that $\det(A)$ is a poor measure of conditioning.
- Conditioning is the inverse distance to ill-posedness [Demmel; 1987] (also Dedieu, Shub, Smale), ie.

$$\frac{1}{\|A^{-1}\|_2}.$$

- Normalizing by $\|A\|_2$ yields **condition number**

$$\frac{1}{\|A\|_2\|A^{-1}\|_2} = \frac{1}{\kappa_2(A)}.$$

- Note that

$$\|A^{-1}\|_2^{-1} = \sigma_n = \min_{\mathbf{x}_i,\mathbf{y}_i} \|A - \mathbf{x}_1 \otimes \mathbf{y}_1 - \cdots - \mathbf{x}_{n-1} \otimes \mathbf{y}_{n-1}\|_2.$$

Conditioning of linear systems

- Important for error analysis [Wilkinson, 1961].
- Let $A = U\Sigma V^\top$ and define

$$\begin{aligned}
S_{\text{forward}}(\varepsilon) &= \{\mathbf{x}' \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b}, \quad \|\mathbf{x}' - \mathbf{x}\|_2 \le \varepsilon\} \\
&= \{\mathbf{x}' \in \mathbb{R}^n \mid \textstyle\sum_{i=1}^n |x_i' - x_i|^2 \le \varepsilon^2\}, \\
S_{\text{backward}}(\varepsilon) &= \{\mathbf{x}' \in \mathbb{R}^n \mid A\mathbf{x}' = \mathbf{b}', \quad \|\mathbf{b}' - \mathbf{b}\|_2 \le \varepsilon\} \\
&= \{\mathbf{x}' \in \mathbb{R}^n \mid \mathbf{x}' - \mathbf{x} = V(\mathbf{y}' - \mathbf{y}), \\
&\qquad\qquad \textstyle\sum_{i=1}^n \sigma_i^2 |y_i' - y_i|^2 \le \varepsilon^2\}.
\end{aligned}$$

Then

$$S_{\text{backward}}(\varepsilon) \subseteq S_{\text{forward}}(\sigma_n^{-1}\varepsilon), \quad S_{\text{forward}}(\varepsilon) \subseteq S_{\text{backward}}(\sigma_1\varepsilon).$$

- Determined by $\sigma_1 = \|A\|_2$ and $\sigma_n^{-1} = \|A^{-1}\|_2$.
- Rule of thumb: $\log_{10} \kappa_2(A) \approx$ loss in number of digits of precision.

What about multilinear systems?

Look at the simplest case. Take $A = [\![a_{ijk}]\!] \in \mathbb{R}^{2\times 2\times 2}$ and $\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^2$.

$$
\begin{aligned}
a_{000}x_0y_0 + a_{010}x_0y_1 + a_{100}x_1y_0 + a_{110}x_1y_1 &= b_{00},\\
a_{001}x_0y_0 + a_{011}x_0y_1 + a_{101}x_1y_0 + a_{111}x_1y_1 &= b_{01},\\
a_{000}x_0z_0 + a_{001}x_0z_1 + a_{100}x_1z_0 + a_{101}x_1z_1 &= b_{10},\\
a_{010}x_0z_0 + a_{011}x_0z_1 + a_{110}x_1z_0 + a_{111}x_1z_1 &= b_{11},\\
a_{000}y_0z_0 + a_{001}y_0z_1 + a_{010}y_1z_0 + a_{011}y_1z_1 &= b_{20},\\
a_{100}y_0z_0 + a_{101}y_0z_1 + a_{110}y_1z_0 + a_{111}y_1z_1 &= b_{21}.
\end{aligned}
$$

- When does this have a solution?
- What is the corresponding manifold of ill-posed problems?
- When does the homogeneous system, ie. $\mathbf{b}_0 = \mathbf{b}_1 = \mathbf{b}_2 = \mathbf{0}$, have a non-trivial solution, ie. $\mathbf{x} \neq \mathbf{0}, \mathbf{y} \neq \mathbf{0}, \mathbf{z} \neq \mathbf{0}$?

Hyperdeterminant

- Work in $\mathbb{C}^{(d_1+1)\times\cdots\times(d_k+1)}$ for the time being ($d_i \geq 1$). Consider

$$\mathscr{M} := \{\mathcal{A} \in \mathbb{C}^{(d_1+1)\times\cdots\times(d_k+1)} \mid \nabla\mathcal{A}(\mathbf{x}_1,\ldots,\mathbf{x}_k) = \mathbf{0} \text{ for non-zero } \mathbf{x}_1,\ldots,\mathbf{x}_k\}.$$

Theorem (Gelfand, Kapranov, Zelevinsky)

$\mathscr{M}$ is a hypersurface iff for all $j = 1,\ldots,k$,

$$d_j \leq \sum\nolimits_{i\neq j} d_i.$$

- The **hyperdeterminant** $\operatorname{Det}(\mathcal{A})$ is the equation of the hypersurface, i.e. a multivariate polynomial in the entries of $\mathcal{A}$ such that

$$\mathscr{M} = \{\mathcal{A} \in \mathbb{C}^{(d_1+1)\times\cdots\times(d_k+1)} \mid \operatorname{Det}(\mathcal{A}) = 0\}.$$

- $\operatorname{Det}(\mathcal{A})$ may be chosen to have integer coefficients.
- For $\mathbb{C}^{m\times n}$, condition becomes $m \leq n$ and $n \leq m$, i.e. square matrices.

$2 \times 2 \times 2$ hyperdeterminant

Hyperdeterminant of $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{2\times 2\times 2}$ [Cayley; 1845] is

$$\operatorname{Det}_{2,2,2}(\mathcal{A}) = \frac{1}{4}\left[\det\left(\begin{bmatrix} a_{000} & a_{010} \\ a_{001} & a_{011}\end{bmatrix} + \begin{bmatrix} a_{100} & a_{110} \\ a_{101} & a_{111}\end{bmatrix}\right) - \det\left(\begin{bmatrix} a_{000} & a_{010} \\ a_{001} & a_{011}\end{bmatrix} - \begin{bmatrix} a_{100} & a_{110} \\ a_{101} & a_{111}\end{bmatrix}\right)\right]^2 - 4\det\begin{bmatrix} a_{000} & a_{010} \\ a_{001} & a_{011}\end{bmatrix}\det\begin{bmatrix} a_{100} & a_{110} \\ a_{101} & a_{111}\end{bmatrix}.$$

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 $\operatorname{Det}_{2,2,2}(\mathcal{A}) = 0$.

$2 \times 2 \times 3$ hyperdeterminant

Hyperdeterminant of $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{2\times 2\times 3}$ is

$$\operatorname{Det}_{2,2,3}(\mathcal{A}) = \det\begin{bmatrix} a_{000} & a_{001} & a_{002} \\ a_{100} & a_{101} & a_{102} \\ a_{010} & a_{011} & a_{012} \end{bmatrix} \det\begin{bmatrix} a_{100} & a_{101} & a_{102} \\ a_{010} & a_{011} & a_{012} \\ a_{110} & a_{111} & a_{112} \end{bmatrix} - \det\begin{bmatrix} a_{000} & a_{001} & a_{002} \\ a_{100} & a_{101} & a_{102} \\ a_{110} & a_{111} & a_{112} \end{bmatrix} \det\begin{bmatrix} a_{000} & a_{001} & a_{002} \\ a_{010} & a_{011} & a_{012} \\ a_{110} & a_{111} & a_{112} \end{bmatrix}$$

Again, the following is true:

$$\begin{aligned}
a_{000}x_0y_0 + a_{010}x_0y_1 + a_{100}x_1y_0 + a_{110}x_1y_1 &= 0,\\
a_{001}x_0y_0 + a_{011}x_0y_1 + a_{101}x_1y_0 + a_{111}x_1y_1 &= 0,\\
a_{002}x_0y_0 + a_{012}x_0y_1 + a_{102}x_1y_0 + a_{112}x_1y_1 &= 0,\\
a_{000}x_0z_0 + a_{001}x_0z_1 + a_{002}x_0z_2 + a_{100}x_1z_0 + a_{101}x_1z_1 + a_{102}x_1z_2 &= 0,\\
a_{010}x_0z_0 + a_{011}x_0z_1 + a_{012}x_0z_2 + a_{110}x_1z_0 + a_{111}x_1z_1 + a_{112}x_1z_2 &= 0,\\
a_{000}y_0z_0 + a_{001}y_0z_1 + a_{002}y_0z_2 + a_{010}y_1z_0 + a_{011}y_1z_1 + a_{012}y_1z_2 &= 0,\\
a_{100}y_0z_0 + a_{101}y_0z_1 + a_{102}y_0z_2 + a_{110}y_1z_0 + a_{111}y_1z_1 + a_{112}y_1z_2 &= 0,
\end{aligned}$$

has a non-trivial solution iff $\operatorname{Det}_{2,2,3}(\mathcal{A}) = 0$.

Cayley hyperdeterminant and tensor rank

- The Cayley hyperdeterminant $\operatorname{Det}_{2,2,2}$ may be extended to any $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ with $\operatorname{rank}_{\otimes}(\mathcal{A}) \le 2$.

Theorem

Let $d_1, d_2, d_3 \ge 2$. $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *is a weak solution, i.e.*

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

iff $\operatorname{Det}_{2,2,2}(\mathcal{A}) = 0$.

Theorem (Kruskal)

Let $\mathcal{A} \in \mathbb{R}^{2\times 2\times 2}$. *Then* $\operatorname{rank}_{\otimes}(\mathcal{A}) = 2$ *if* $\operatorname{Det}_{2,2,2}(\mathcal{A}) > 0$ *and* $\operatorname{rank}_{\otimes}(\mathcal{A}) = 3$ *if* $\operatorname{Det}_{2,2,2}(\mathcal{A}) < 0$.

Condition number of a multilinear system

- Like the matrix determinant, the value of the hyperdeterminant is a poor measure of conditioning. Need to compute distance to $\mathscr{M}$.

Theorem

Let $A \in \mathbb{R}^{2\times 2\times 2}$*.* $\operatorname{Det}_{2,2,2}(A) = 0$ *iff*

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

for some $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2$*,* $i = 1, 2, 3$*.*

- Conditioning of the problem can be obtained from

$$\min_{\mathbf{x},\mathbf{y}} \|A - \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{y} - \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{x} - \mathbf{y} \otimes \mathbf{x} \otimes \mathbf{x}\|.$$

- $\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{y} + \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{x} + \mathbf{y} \otimes \mathbf{x} \otimes \mathbf{x}$ has outer product rank 3 generically (in fact, iff $\mathbf{x}, \mathbf{y}$ are linearly independent).
- Surprising: the manifold of ill-posed problem has full rank almost everywhere!

Nonnegative matrix factorization

- D.D. Lee and H.S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, **401** (1999), pp. 788–791.
- **Main idea behind NMF** (everything else is fluff): the way dictionary functions combine to build 'target objects' is an exclusively additive process and should not involve any cancellations between the dictionary functions.
- **NMF in a nutshell**: given nonnegative matrix A, decompose it into a sum of outer-products of nonnegative vectors:

$$A = XY^{\top} = \sum\nolimits_{i=1}^{r} \mathbf{x}_i \otimes \mathbf{y}_i.$$

- **Noisy situation**: approximate A by a sum of outer-products of nonnegative vectors

$$\min_{X \geq 0, Y \geq 0} \|A - XY^{\top}\|_F = \min_{\mathbf{x}_i \geq 0, \mathbf{y}_i \geq 0} \left\| A - \sum\nolimits_{i=1}^{r} \mathbf{x}_i \otimes \mathbf{y}_i \right\|_F .$$

Generalizing to hypermatrices

- **Nonnegative outer-product decomposition** for hypermatrix $\mathcal{A} \geq 0$ is

$$\mathcal{A} = \sum\nolimits_{p=1}^{r} \mathbf{x}_p \otimes \mathbf{y}_p \otimes \mathbf{z}_p$$

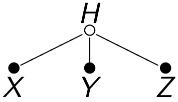
  where $\mathbf{x}_p \in \mathbb{R}_+^l, \mathbf{y}_p \in \mathbb{R}_+^m, \mathbf{z}_p \in \mathbb{R}_+^n$.
- Clear that such a decomposition exists for any $\mathcal{A} \geq 0$.
- **Nonnegative outer-product rank**: minimal r for which such a decomposition is possible.
- Best nonnegative outer-product rank-r approximation:

$$\operatorname{argmin}\Big\{\Big\|\mathcal{A} - \sum\nolimits_{p=1}^{r} \mathbf{x}_p \otimes \mathbf{y}_p \otimes \mathbf{z}_p\Big\|_F \;\Big|\; \mathbf{x}_p, \mathbf{y}_p, \mathbf{z}_p \geq 0\Big\}.$$

Nonnegativity helps

Approximation of joint probability distributions by conditional probability distributions under the Naïve Bayes Hypothesis:

$$\Pr(x, y, z) = \sum_h \Pr(h) \Pr(x \mid h) \Pr(y \mid h) \Pr(z \mid h)$$



Theorem (L-Comon)

The set $\{\mathcal{A} \in \mathbb{R}_+^{l \times m \times n} \mid \operatorname{rank}_+(\mathcal{A}) \leq r\}$ *is closed.*

- *Extends to arbitrary order.*
- *Independent of norms and even Brègman divergences.*
- *Holds more generally over* $C_1 \otimes \cdots \otimes C_p$ *where* $C_1, \ldots, C_p$ *are line-free cones.*

Recap: outer product decomposition in spectroscopy

- Application to fluorescence spectral analysis by [Bro; 1997].
- Specimens with a number of pure substances in different concentration
  - a_{ijk} = fluorescence emission intensity at wavelength λ_j^{em} of ith sample excited with light at wavelength λ_k^{ex}.
  - Get 3-way data $\mathcal{A} = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
  - Get outer product decomposition of $\mathcal{A}$

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

- Get the true chemical factors responsible for the data.
  - r: number of pure substances in the mixtures,
  - $\mathbf{x}_p = (x_{1p}, \dots, x_{lp})$: relative concentrations of pth substance in specimens $1, \dots, l$,
  - $\mathbf{y}_p = (y_{1p}, \dots, y_{mp})$: excitation spectrum of pth substance,
  - $\mathbf{z}_p = (z_{1p}, \dots, z_{np})$: emission spectrum of pth substance.
- Noisy case: find best rank-r approximation (CANDECOMP/PARAFAC).

Symmetric hypermatrices for blind source separation

Problem

Given $\mathbf{y} = M\mathbf{x} + \mathbf{n}$*. 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* $\mathbf{y} \in \mathbb{C}^m$*. Goal: recover* $\mathbf{x}$ *from* $\mathbf{y}$*.*

- Assumptions:
  1. components of $\mathbf{x}$ statistically independent,
  2. M full column-rank,
  3. $\mathbf{n}$ 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 hypermatrix $\kappa_k(\mathbf{y})$.

Diagonalizing a symmetric hypermatrix

- A best symmetric rank approximation may not exist either:

Example

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be linearly independent. Define for $n \in \mathbb{N}$,

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

and

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

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

$$\lim_{n\to\infty} \mathcal{A}_n = \mathcal{A}.$$

Variational approach to eigenvalues/vectors

- $A \in \mathbb{R}^{m \times n}$ symmetric.
- Eigenvalues and eigenvectors are critical values and critical points of

$$\mathbf{x}^\top A\mathbf{x}/\|\mathbf{x}\|_2^2.$$

- Equivalently, critical values/points of $\mathbf{x}^\top A\mathbf{x}$ constrained to unit sphere.
- Lagrangian:

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

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

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

Variational approach to singular values/vectors

- $A \in \mathbb{R}^{m \times n}$.
- Singular values and singular vectors are critical values and critical points of

$$\mathbf{x}^\top A\mathbf{y}/\|\mathbf{x}\|_2\|\mathbf{y}\|_2.$$

- Lagrangian:

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

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

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

- Writing $\mathbf{u}_c = \mathbf{x}_c/\|\mathbf{x}_c\|_2$ and $\mathbf{v}_c = \mathbf{y}_c/\|\mathbf{y}_c\|_2$ yields familiar

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

Eigenvalues/vectors of a tensor

- Extends to hypermatrices.
- For $\mathbf{x} = [x_1, \dots, x_n]^\top \in \mathbb{R}^n$, write $\mathbf{x}^p := [x_1^p, \dots, x_n^p]^\top$.
- Define the 'ℓ^k-norm' $\|\mathbf{x}\|_k = (x_1^k + \cdots + x_n^k)^{1/k}$.
- Define eigenvalues/vectors of $\mathcal{A} \in \mathsf{S}^k(\mathbb{R}^n)$ as critical values/points of the multilinear Rayleigh quotient

$$\mathcal{A}(\mathbf{x}, \dots, \mathbf{x}) / \|\mathbf{x}\|_k^k.$$

- Lagrangian

$$L(\mathbf{x}, \lambda) := \mathcal{A}(\mathbf{x}, \dots, \mathbf{x}) - \lambda(\|\mathbf{x}\|_k^k - 1).$$

- At a critical point

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

Eigenvalues/vectors of a tensor

- If $\mathcal{A}$ is symmetric,

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

- Also obtained by Liqun Qi independently:
  - L. Qi, "Eigenvalues of a real supersymmetric tensor," *J. Symbolic Comput.*, **40** (2005), no. 6.
  - L, "Singular values and eigenvalues of tensors: a variational approach," *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, **1** (2005).
- For unsymmetric hypermatrices — get different eigenpairs for different modes (unsymmetric matrix have different left/right eigenvectors).
- Falls outside Classical Invariant Theory — not invariant under $Q \in \mathrm{O}(n)$, ie. $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$.
- Invariant under $Q \in \mathrm{GL}(n)$ with $\|Q\mathbf{x}\|_k = \|\mathbf{x}\|_k$.

Singular values/vectors of a tensor

- Likewise for singular values/vectors of $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$.
- Lagrangian is

$$L(\mathbf{x}, \mathbf{y}, \mathbf{z}, \sigma) = \mathcal{A}(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \sigma(\|\mathbf{x}\|\|\mathbf{y}\|\|\mathbf{z}\| - 1)$$

  where $\sigma \in \mathbb{R}$ is the Lagrange multiplier.
- At a critical point,

$$\begin{aligned} \mathcal{A}\left(I_l, \mathbf{y}/\|\mathbf{y}\|, \mathbf{z}/\|\mathbf{z}\|\right) &= \sigma \mathbf{x}/\|\mathbf{x}\|, \\ \mathcal{A}\left(\mathbf{x}/\|\mathbf{x}\|, I_m, \mathbf{z}/\|\mathbf{z}\|\right) &= \sigma \mathbf{y}/\|\mathbf{y}\|, \\ \mathcal{A}\left(\mathbf{x}/\|\mathbf{x}\|, \mathbf{y}/\|\mathbf{y}\|, I_n\right) &= \sigma \mathbf{z}/\|\mathbf{z}\|. \end{aligned}$$

- Normalize to get

$$\mathcal{A}(I_l, \mathbf{v}, \mathbf{w}) = \sigma \mathbf{u}, \quad \mathcal{A}(\mathbf{u}, I_m, \mathbf{w}) = \sigma \mathbf{v}, \quad \mathcal{A}(\mathbf{u}, \mathbf{v}, I_n) = \sigma \mathbf{w}.$$

Immediate properties

- Largest singular value is the norm of the multilinear functional associated with A induced by the p-norm, i.e.

$$\sigma_{\max}(A) = \|A\|_{p,\dots,p}.$$

- For $d_1, \dots, d_k$ such that

$$d_i - 1 \le \sum_{j \ne i} (d_j - 1) \quad \text{for all } i = 1, \dots, k,$$

  and $\mathrm{Det}_{d_1,\dots,d_k}$ the hyperdeterminant in $\mathbb{R}^{d_1 \times \cdots \times d_k}$. 0 is a singular value of $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ if and only if

$$\mathrm{Det}_{d_1,\dots,d_k}(A) = 0.$$

- Pseudospectrum of square matrix $A \in \mathbb{C}^{n \times n}$,

$$\sigma_\varepsilon(A) = \{\lambda \in \mathbb{C} \mid \|(A - \lambda I)^{-1}\|_2 > \varepsilon^{-1}\} = \{\lambda \in \mathbb{C} \mid \sigma_{\min}(A - \lambda I) < \varepsilon\}.$$

- Plausible generalizations to cubical hypermatrix $\mathcal{A} \in \mathbb{C}^{n \times \cdots \times n}$,

$$\sigma_\varepsilon^{\Sigma}(\mathcal{A}) = \{\lambda \in \mathbb{C} \mid \sigma_{\min}(\mathcal{A} - \lambda \mathcal{I}) < \varepsilon\}$$

$$\sigma_\varepsilon^{\Delta}(\mathcal{A}) = \{\lambda \in \mathbb{C} \mid \inf_{\mathrm{Det}_{n,\dots,n}(\mathcal{B})=0} \|\mathcal{A} - \lambda \mathcal{I} - \mathcal{B}\|_F < \varepsilon^{-1}\}.$$

Perron-Frobenius theorem for hypermatrices

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

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

  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 $\mathcal{A}$ is **irreducible** if it is not reducible. In particular, if $\mathcal{A} > 0$, then it is irreducible.

Theorem (L)

Let $0 \le \mathcal{A} = [\![a_{j_1 \cdots j_k}]\!] \in \mathsf{T}^k(\mathbb{R}^n)$ *be irreducible. Then* $\mathcal{A}$ *has*

1. *a positive real eigenvalue* λ *with an eigenvector* $\mathbf{x}$*;*
2. $\mathbf{x}$ *may be chosen to have all entries non-negative;*
3. *if* μ *is an eigenvalue of* $\mathcal{A}$*, then* $|\mu| \le \lambda$*.*

Result extended by K.-C. Chang, K. Pearson, and T. Zhang.

Hypergraphs

- $G = (V, E)$ is 3-**hypergraph**.
  - V is the finite set of **vertices**.
  - E is the subset of **hyperedges**, ie. 3-element subsets of V.
- Write elements of E as $[x, y, z]$ $(x, y, z \in V)$.
- G is **undirected**, so $[x, y, z] = [y, z, x] = \cdots = [z, y, x]$.
- Hyperedge is said to **degenerate** if 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.
- G is r-**uniform** if every edge contains exactly r vertices.
- Good reference: D. Knuth, *The art of computer programming*, **4**, pre-fascicle 0a, 2008.

Spectral hypergraph theory

- Define the order-3 **adjacency hypermatrix** $\mathcal{A} = [\![a_{ijk}]\!]$ by

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

- $\mathcal{A} \in \mathbb{R}^{|V|\times|V|\times|V|}$ nonnegative symmetric hypermatrix.
- Consider cubic form

$$\mathcal{A}(f, f, f) = \sum_{x,y,z} a_{xyz} f(x) f(y) f(z),$$

where $f \in \mathbb{R}^V$.

- Eigenvalues (resp. eigenvectors) of A are the critical values (resp. critical points) of $\mathcal{A}(f, f, f)$ constrained to the $f \in \ell^3(V)$, ie.

$$\sum_{x\in V} f(x)^3 = 1.$$

Spectral hypergraph theory

We have the following.

Lemma (L)

Let G be an m-regular 3-hypergraph. $\mathcal{A}$ its adjacency hypermatrix. Then

1. *m is an eigenvalue of $\mathcal{A}$;*
2. *if λ is an eigenvalue of $\mathcal{A}$, then $|\lambda| \leq m$;*
3. *λ has multiplicity 1 if and only if G is connected.*

Related work: J. Friedman, A. Wigderson, "On the second eigenvalue of hypergraphs," *Combinatorica*, **15** (1995), no. 1.

Spectral hypergraph theory

- 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 $u, v, \dots, z$ with $a_{uv\cdots z} \neq 0$, $u, v, \dots, z$ must each lie in a distinct V_i ($i = 1, \dots, k$).

Lemma (L)

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

1. *If $k \equiv 1 \bmod 4$, then every eigenvalue of G occurs with multiplicity a multiple of k.*
2. *If $k \equiv 3 \bmod 4$, then the spectrum of G is symmetric, ie. if λ is an eigenvalue, then so is $-\lambda$.*
3. *Furthermore, every eigenvalue of G occurs with multiplicity a multiple of $k/2$, ie. if λ is an eigenvalue of G, then λ and $-\lambda$ occurs with the same multiplicity.*

To do

- Cases $k \equiv 0, 2 \bmod 4$
- Cheeger type isoperimetric inequalities
- Expander hypergraphs
- Algorithms for eigenvalues/vectors of a hypermatrix