Numerical Multilinear Algebra II

Lek-Heng Lim

University of California, Berkeley

January 5–7, 2009
Recap: tensor ranks

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

  \[
  \text{rank}(A) = \dim(\text{span}_{\mathbb{R}}\{A_{\bullet 1}, \ldots, A_{\bullet n}\}) \quad \text{(column rank)}
  \]

  \[
  = \dim(\text{span}_{\mathbb{R}}\{A_{1 \bullet}, \ldots, A_{m \bullet}\}) \quad \text{(row rank)}
  \]

  \[
  = \min\{r \mid A = \sum_{i=1}^{r} u_i v_i^T\} \quad \text{(outer product rank)}.
  \]

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

  \[
  r_1(A) = \dim(\text{span}_{\mathbb{R}}\{A_{1 \bullet \bullet}, \ldots, A_{l \bullet \bullet}\})
  \]

  \[
  r_2(A) = \dim(\text{span}_{\mathbb{R}}\{A_{\bullet 1 \bullet}, \ldots, A_{\bullet m \bullet}\})
  \]

  \[
  r_3(A) = \dim(\text{span}_{\mathbb{R}}\{A_{\bullet \bullet 1}, \ldots, A_{\bullet \bullet n}\})
  \]

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

  \[
  \text{rank}_{\otimes}(A) = \min\{r \mid A = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i\}
  \]

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

Matrix EVD and SVD

- Rank revealing decompositions.

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

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

where $\text{rank}(A) = r$, $V \in O(n)$ eigenvectors, $\Lambda$ eigenvalues.

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

$$A = U \Sigma V^\top = \sum_{i=1}^{r} \sigma_i u_i \otimes v_i$$

where $\text{rank}(A) = r$, $U \in O(m)$ left singular vectors, $V \in O(n)$ right singular vectors, $\Sigma$ 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 $A \in S^3(\mathbb{R}^n)$,

$$A = \sum_{i=1}^{r} \lambda_i v_i \otimes v_i \otimes v_i$$

where $\text{rank}_S(A) = r$, $v_i$ unit vector, $\lambda_i \in \mathbb{R}$.

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

$$A = \sum_{i=1}^{r} \sigma_i u_i \otimes v_i \otimes w_i$$

where $\text{rank}_\otimes(A) = r$, $u_i \in \mathbb{R}^l$, $v_i \in \mathbb{R}^m$, $w_i \in \mathbb{R}^n$ 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 $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$,

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

  where $\text{rank} \Box(A) = (r_1, r_2, r_3)$, $U \in \mathbb{R}^{l \times r_1}$, $V \in \mathbb{R}^{m \times r_2}$, $W \in \mathbb{R}^{n \times r_3}$ have orthonormal columns and $C \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

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

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

  where $\text{rank} \Box(A) = (r, r, r)$, $U \in \mathbb{R}^{n \times r}$ has orthonormal columns and $C \in S^3(\mathbb{R}^r)$.

- Ditto for **nonnegative multilinear decomposition**.
Outer product rank is hard to compute

  - 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 \( \text{rank}_{\otimes}(A) \) for \( A \in \mathbb{F}^{l \times m \times n} \) 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?
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 $A \in \mathbb{R}^{l\times m\times n} \subset \mathbb{C}^{l\times m\times n}$, $\operatorname{rank}_\otimes(A)$ is base field dependent.

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

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

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

\[ f(x, y, z) = \sum_{p=1}^{r} \theta_p(x) \varphi_p(y) \psi_p(z) \]

\[ \theta_p(x) = \theta(x, t_p), \quad \varphi_p(y) = \varphi(y, t_p), \quad \psi_p(z) = \psi(z, t_p), \quad r \text{ possibly } \infty. \]

- Discrete

\[ a_{ijk} = \sum_{p=1}^{r} u_{ip} v_{jp} w_{kp} \]

\[ a_{ijk} = f(x_i, y_j, z_k), \quad u_{ip} = \theta_p(x_i), \quad v_{jp} = \varphi_p(y_j), \quad w_{kp} = \psi_p(z_k). \]
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 \( A \in \mathbb{R}^{n \times n} \),
  \[ \exp(x^\top A x) = \exp(z^\top \Lambda z) = \prod_{i=1}^{n} \exp(\lambda_i z_i^2). \]
- Gaussian mixture models
  \[ f(x) = \sum_{j=1}^{m} \alpha_j \exp[(x - \mu_j)^\top A_j (x - \mu_j)], \]
  \( f \) is a sum of separable functions.
Multilinear decomposition: integral kernels

Approximation by sum or integral kernels

- **Continuous**

\[ f(x, y, z) = \int \int \int K(x', y', z') \theta(x, x') \varphi(y, y') \psi(z, z') \, dx' \, dy' \, dz'. \]

- **Semi-discrete**

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

\[ c_{i'j'k'} = K(x_{i'}, y_{j'}, z_{k'}), \quad \theta_{i'}(x) = \theta(x, x_{i'}), \quad \varphi_{j'}(y) = \varphi(y, y_{j'}), \quad \psi_{k'}(z) = \psi(z, z_{k'}). \]

\[ p, q, r \text{ possibly } \infty. \]

- **Discrete**

\[ a_{ijk} = \sum_{i', j', k'=1}^{p, q, r} c_{i'j'k'} u_{ii'} v_{jj'} w_{kk'} \]

\[ a_{ijk} = f(x_i, y_j, z_k), \quad u_{ii'} = \theta_{i'}(x_i), \quad v_{jj'} = \varphi_{j'}(y_j), \quad w_{kk'} = \psi_{k'}(z_k). \]
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.
- $f_1, \ldots, f_r \in \mathcal{D} \subset \mathcal{H}$ dictionary.
- $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$ 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

\[
\argmin \{ \varphi(f, \alpha_1 f_1 + \ldots + \alpha_r f_r) \mid f_i \in \mathcal{D} \}. 
\]

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

  \[ \frac{22}{7} = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 2 \cdot 10^{-3} + \cdots \]

- Spanning set: $\mathcal{D} = \{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\} \subseteq \mathbb{R}^2$,

  \[ \begin{bmatrix} 2 \\ -3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ -1 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \]

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

  \[ \exp(x) = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots \]

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

  \[ \frac{1}{2}x = \sin(x) - \frac{1}{2} \sin(2x) + \frac{1}{3} \sin(3x) - \cdots \]

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

- Discrete cosine:
  \[ \mathcal{D} = \left\{ \sqrt{\frac{2}{N}} \cos(k + \frac{1}{2})(n + \frac{1}{2}) \frac{\pi}{N} \mid k \in [N - 1] \right\} \subseteq \mathbb{C}^N. \]

- Peter-Weyl:
  \[ \mathcal{D} = \{ \langle \pi(x)e_i, e_j \rangle \mid \pi \in \widehat{G}, i, j \in [d_\pi] \} \subseteq L^2(G). \]

- Paley-Wiener:
  \[ \mathcal{D} = \{ \text{sinc}(x - n) \mid n \in \mathbb{Z} \} \subseteq H^2(\mathbb{R}). \]

- Gabor:
  \[ \mathcal{D} = \left\{ e^{i\alpha nx} e^{-(x-m\beta)^2/2} \mid (m, n) \in \mathbb{Z} \times \mathbb{Z} \right\} \subseteq L^2(\mathbb{R}). \]

- Wavelet:
  \[ \mathcal{D} = \{ 2^{n/2} \psi(2^n x - m) \mid (m, n) \in \mathbb{Z} \times \mathbb{Z} \} \subseteq L^2(\mathbb{R}). \]

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

**Definition**

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

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

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

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

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

$$
\inf_{g \in \text{span}(\mathcal{D})} \| f - g \|.
$$
Small is beautiful

\[ f \approx \sum_{i \in \mathcal{I} \subseteq \mathcal{D}} \alpha_i f_i \]

- Want good approximation, ie. \( \| f - \sum_{i \in \mathcal{I} \subseteq \mathcal{D}} \alpha_i f_i \| \) small.
- Want sparse/concentrated representation, ie. \( |\mathcal{I}| \) small.
- Sparsity depends on choice of \( \mathcal{D} \).

▶ \( \mathcal{D}_{10} = \{10^n | n \in \mathbb{Z}\}, \mathcal{D}_3 = \{3^n | n \in \mathbb{Z}\} \subseteq \mathbb{R} \),

\[
\frac{1}{3} = [0.33333 \cdots]_{10} = \sum_{n=1}^{\infty} 3 \cdot 10^{-n} = [0.1]_3 = 1 \cdot 3^{-1}.
\]

▶ \( \mathcal{D}_{\text{fourier}} = \{\cos(nx), \sin(nx) | n \in \mathbb{Z}\} \),

\[
\frac{1}{2}x = \sin(x) - \frac{1}{2} \sin(2x) + \frac{1}{3} \sin(3x) - \cdots.
\]

▶ \( \mathcal{D}_{\text{taylor}} = \{x^n | n \in \mathbb{N} \cup \{0\}\} \),

\[
\sin(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \cdots.
\]
Bigger is better

- **Union of dictionaries**: allows for efficient (sparse) representation of different features
  - $\mathcal{D} = \mathcal{D}_{\text{fourier}} \cup \mathcal{D}_{\text{wavelets}}$,
  - $\mathcal{D} = \mathcal{D}_{\text{spikes}} \cup \mathcal{D}_{\text{sinusoids}} \cup \mathcal{D}_{\text{splines}}$,
  - $\mathcal{D} = \mathcal{D}_{\text{wavelets}} \cup \mathcal{D}_{\text{curvelets}} \cup \mathcal{D}_{\text{beamlets}} \cup \mathcal{D}_{\text{ridgelets}}$.

- **$\mathcal{D}$ 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**: $\mathcal{D}$ 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 \(a = [a_1, \ldots, a_n]^\top \in \mathbb{R}^n\).

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

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,
    \[ \mathcal{D} \otimes = \{ g : [l] \times [m] \times [n] \to \mathbb{R} \mid g(i, j, k) = \vartheta(i)\varphi(j)\psi(k) \}, \]
    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:
    \[ \mathcal{D}_S = \{ g : [n] \times [n] \times [n] \to \mathbb{R} \mid g(i, j, k) = \vartheta(i)\vartheta(j)\vartheta(k) \}, \]
    where \( \vartheta : [l] \to \mathbb{R} \).
Tensor approximations

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

\[
\text{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n) = \{ A \in \mathbb{R}^{l \times m \times n} \mid A = u \otimes v \otimes w \} = \\
\{ A \in \mathbb{R}^{l \times m \times n} \mid a_{i_1i_2i_3}a_{j_1j_2j_3} = a_{k_1k_2k_3}a_{l_1l_2l_3}, \{i_\alpha,j_\alpha\} = \{k_\alpha,l_\alpha\} \}
\]

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

- $r$th secant (quasiprojective) variety of the Segre variety is the set of $r$ term approximants.
- If $\mathcal{D} = \text{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n)$, then
  \[
  \Sigma_r(\mathcal{D}) = \{ \mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \text{rank}_\otimes(\mathcal{A}) \leq r \}.
  \]
- Rank revealing matrix decompositions (non-unique: LU, QR, SVD):
  \[
  \mathcal{D} = \{ \mathbf{x}\mathbf{y}^\top \mid (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n \} = \{ \mathcal{A} \in \mathbb{R}^{m \times n} \mid \text{rank}(\mathcal{A}) \leq 1 \}.
  \]
- Often unique for tensors [Kruskal; 1977], [Sidiroupoulos, Bro; 2000]:
  - $\text{spark}(\mathbf{x}_1, \ldots, \mathbf{x}_r) =$ 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
    \[
    \text{spark}(\mathbf{u}_1, \ldots, \mathbf{u}_r) + \text{spark}(\mathbf{v}_1, \ldots, \mathbf{v}_r) + \text{spark}(\mathbf{w}_1, \ldots, \mathbf{w}_r) \geq 2r + 5.
    \]
Dictionaries of positive dimensions

- Neural networks:
  \[ \mathcal{D} = \{ \sigma(\mathbf{w}^\top \mathbf{x} + w_0) \mid (w_0, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^n \} \]
  where \( \sigma : \mathbb{R} \to \mathbb{R} \) sigmoid function, e.g. \( \sigma(x) = \left[1 + \exp(-x)\right]^{-1} \).

- Exponential:
  \[ \mathcal{D} = \{ e^{-t \mathbf{x}} \mid t \in \mathbb{R}_+ \} \quad \text{or} \quad \mathcal{D} = \{ e^{\tau \mathbf{x}} \mid \tau \in \mathbb{C} \} \]

- Outer product decomposition:
  \[ \mathcal{D} = \{ \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \mid (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n \} \]
  \[ = \{ \mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \text{rank}_\otimes(\mathcal{A}) \leq 1 \} \]

- Symmetric outer product decomposition:
  \[ \mathcal{D} = \{ \mathbf{v} \otimes \mathbf{v} \otimes \mathbf{v} \mid \mathbf{v} \in \mathbb{R}^n \} = \{ \mathcal{A} \in \mathcal{S}^3(\mathbb{R}^n) \mid \text{rank}_\mathcal{S}(\mathcal{A}) \leq 1 \} \]

- Nonnegative outer product decomposition:
  \[ \mathcal{D} = \{ \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \mid (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{R}_+^l \times \mathbb{R}_+^m \times \mathbb{R}_+^n \} \]
  \[ = \{ \mathcal{A} \in \mathbb{R}_+^{l \times m \times n} \mid \text{rank}_+(\mathcal{A}) \leq 1 \} \]
Recall: fundamental problem of multiway data analysis

- A hypermatrix, symmetric hypermatrix, or nonnegative hypermatrix. Want
  \[ \arg\min_{\text{rank}(B) \leq r} \| A - B \|. \]
- rank(B) may be outer product rank, multilinear rank, symmetric rank (for symmetric hypermatrix), or nonnegative rank (nonnegative hypermatrix).

Example

Given \( A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \), find \( \sigma_i, u_i, v_i, w_i, i = 1, \ldots, r \), that minimizes

\[ \| A - \sigma_1 u_1 \otimes v_1 \otimes w_1 - \sigma_2 u_2 \otimes v_2 \otimes w_2 - \cdots - \sigma_r u_r \otimes v_r \otimes w_r \| \]

or \( C \in \mathbb{R}^{r_1 \times r_2 \times r_3} \) 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

\[ \| A - (U, V, W) \cdot C \|. \]

- May assume \( u_i, v_i, w_i \) unit vectors and \( U, V, W \) orthonormal columns.
Recall: fundamental problem of multiway data analysis

Example

Given $A \in S^k(\mathbb{C}^n)$, find $u_i$, $i = 1, \ldots, r$, that minimizes

$$\|A - \lambda_1 u_1 \otimes^k - \lambda_2 u_2 \otimes^k - \cdots - \lambda_r u_r \otimes^k\|$$

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

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

- May assume $u_i$ unit vector and $U$ orthonormal columns.
Best low rank approximation of a matrix

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

\[
\text{argmin}_{\text{rank}(B) \leq r} \| A - B \|.
\]

- More precisely, find \( \sigma_i, u_i, v_i, i = 1, \ldots, r \), that minimizes

\[
\| A - \sigma_1 u_1 \otimes v_1 - \sigma_2 u_2 \otimes v_2 - \cdots - \sigma_r u_r \otimes v_r \|.
\]

**Theorem (Eckart–Young)**

Let \( A = U \Sigma V^\top = \sum_{i=1}^{\text{rank}(A)} \sigma_i u_i v_i^\top \) be singular value decomposition. For \( r \leq \text{rank}(A) \), let

\[
A_r := \sum_{i=1}^r \sigma_i u_i v_i^\top.
\]

Then

\[
\| A - A_r \|_F = \min_{\text{rank}(B) \leq r} \| A - B \|_F.
\]

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

\[
\inf\{ \| B - A \| \mid \text{rank}_\otimes(A) \leq r \} = 0.
\]

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

\[
\inf\{ \| C - A \| \mid \text{rank}_\otimes(A) \leq r \}
\]

is not attained (by any \( A \) with \( \text{rank}_\otimes(A) \leq r \)).
Non-existence of best low-rank approximation

- For \( x_i, y_i \in \mathbb{R}^{d_i}, \ i = 1, 2, 3, \)

\[
\mathcal{A} := x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3.
\]

- For \( n \in \mathbb{N}, \)

\[
\mathcal{A}_n := n \left( x_1 + \frac{1}{n} y_1 \right) \otimes \left( x_2 + \frac{1}{n} y_2 \right) \otimes \left( x_3 + \frac{1}{n} y_3 \right) - nx_1 \otimes x_2 \otimes x_3.
\]

**Lemma**

\( \text{rank}_\otimes(\mathcal{A}) = 3 \) iff \( x_i, y_i \) linearly independent, \( i = 1, 2, 3. \) Furthermore, it is clear that \( \text{rank}_\otimes(\mathcal{A}_n) \leq 2 \) and

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

- Original result, in a slightly different form, due to:
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 \( d_1, \ldots, d_k \geq 2 \). For any \( s \) such that

\[
2 \leq s \leq \min\{d_1, \ldots, d_k\},
\]

there exists \( A \in \mathbb{R}^{d_1 \times \cdots \times d_k} \) with \( \text{rank}_{\bigotimes}(A) = s \) 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 hypermatrix in \( \mathbb{R}^{d_1 \times \cdots \times d_k} \) can have rank exceeding \( \min\{d_1, \ldots, d_k\} \).
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 $A_n$ such that $\text{rank} \otimes (A_n) \leq r$ and $\lim_{n \to \infty} A_n = A$ with $\text{rank} \otimes (A) = r + s$.

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

**Theorem**

Let $\mu$ 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

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

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 $x, y \in \mathbb{R}^m$ be linearly independent. Define for $n \in \mathbb{N}$,
  \[
  A_n := n \left[ x + \frac{1}{n} y \right] \otimes^p - nx \otimes^p
  \]

- Define
  \[ A := x \otimes y \otimes \cdots \otimes y + y \otimes x \otimes \cdots \otimes y + \cdots + y \otimes y \otimes \cdots \otimes x. \]

- Then $\text{rank}_S(A_n) \leq 2$, $\text{rank}_S(A) \geq p$, and
  \[ \lim_{n \to \infty} A_n = A. \]

- See [Comon, Golub, L, Mourrain; 08] for details.
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 $n \in \mathbb{N}$,

$$\mathcal{D}_n := n \left[ P_1 + \frac{1}{n} Q_1 \right] \otimes \left[ P_2 + \frac{1}{n} Q_2 \right] \otimes \left[ P_3 + \frac{1}{n} Q_3 \right] -nP_1 \otimes P_2 \otimes P_3.$$

Then

$$\lim_{n \to \infty} \mathcal{D}_n = \mathcal{D}.$$
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 $n \in \mathbb{N}$,

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

Then

$$\lim_{n \to \infty} f_n = f.$$
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 $C \in \{ \mathcal{A} | \text{rank}_{\otimes}(\mathcal{A}) \leq r \}$ attaining

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

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

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 $A_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a sequence of hypermatrices with $\text{rank}_{\otimes}(A_n) \leq 2$ and

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

where the limit is taken in any norm topology. If the limiting hypermatrix $A$ has rank higher than 2, then $\text{rank}_{\otimes}(A)$ must be exactly 3 and there exist pairs of linearly independent vectors $x_1, y_1 \in \mathbb{R}^{d_1}$, $x_2, y_2 \in \mathbb{R}^{d_2}$, $x_3, y_3 \in \mathbb{R}^{d_3}$ such that

$$A = x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3.$$

- 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 \( b \in \mathbb{R}^n \). Suppose we want to solve system of linear equations \( Ax = b \).

- \( \mathcal{M} = \{ A \in \mathbb{R}^{n \times n} \mid \det(A) = 0 \} \) is the manifold of ill-posed problems.

- \( A \in \mathcal{M} \) iff \( Ax = 0 \) 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

\[
\| A^{-1} \|_2^{-1} = \sigma_n = \min_{x_i,y_i, i} \| A - x_1 \otimes y_1 - \cdots - x_{n-1} \otimes y_{n-1} \|_2.
\]
Conditioning of linear systems

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

  $$S_{\text{forward}}(\varepsilon) = \{ x' \in \mathbb{R}^n \mid Ax = b, \quad \|x' - x\|_2 \leq \varepsilon \}$$
  $$= \{ x' \in \mathbb{R}^n \mid \sum_{i=1}^n |x'_i - x_i|^2 \leq \varepsilon^2 \},$$

  $$S_{\text{backward}}(\varepsilon) = \{ x' \in \mathbb{R}^n \mid Ax' = b', \quad \|b' - b\|_2 \leq \varepsilon \}$$
  $$= \{ x' \in \mathbb{R}^n \mid x' - x = V(y' - y), \quad \sum_{i=1}^n \sigma_i^2 |y'_i - y_i|^2 \leq \varepsilon^2 \}.$$  

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 $b_0, b_1, b_2 \in \mathbb{R}^2$.

\begin{align*}
    a_{000}x_0y_0 + a_{010}x_0y_1 + a_{100}x_1y_0 + a_{110}x_1y_1 &= b_0, \\
    a_{001}x_0y_0 + a_{011}x_0y_1 + a_{101}x_1y_0 + a_{111}x_1y_1 &= b_1, \\
    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_2, \\
    a_{100}y_0z_0 + a_{101}y_0z_1 + a_{110}y_1z_0 + a_{111}y_1z_1 &= b_{21}.
\end{align*}

- When does this have a solution?
- What is the corresponding manifold of ill-posed problems?
- When does the homogeneous system, ie. $b_0 = b_1 = b_2 = 0$, have a non-trivial solution, ie. $x \neq 0, y \neq 0, z \neq 0$?
Hyperdeterminant

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

$$\mathcal{M} := \{ A \in \mathbb{C}^{(d_1+1)\times\cdots\times(d_k+1)} \mid \nabla A(x_1, \ldots, x_k) = 0 \}$$

for non-zero $x_1, \ldots, x_k$.

**Theorem (Gelfand, Kapranov, Zelevinsky)**

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

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

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

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

- $\text{Det}(A)$ 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

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

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

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

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

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

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

$$- \det \begin{bmatrix} a_{000} & a_{001} & a_{002} \\ a_{100} & a_{101} & a_{102} \\ a_{110} & a_{111} & a_{112} \end{bmatrix} \det \begin{bmatrix} a_{000} & a_{001} & a_{002} \\ a_{010} & a_{011} & a_{012} \\ a_{110} & a_{111} & a_{112} \end{bmatrix}$$

Again, the following is true:

$$a_{000}x_0y_0 + a_{010}x_0y_1 + a_{100}x_1y_0 + a_{110}x_1y_1 = 0,$$
$$a_{001}x_0y_0 + a_{011}x_0y_1 + a_{101}x_1y_0 + a_{111}x_1y_1 = 0,$$
$$a_{002}x_0y_0 + a_{012}x_0y_1 + a_{102}x_1y_0 + a_{112}x_1y_1 = 0,$$

$$a_{000}x_0z_0 + a_{001}x_0z_1 + a_{002}x_0z_2 + a_{100}x_1z_0 + a_{101}x_1z_1 + a_{102}x_1z_2 = 0,$$
$$a_{010}x_0z_0 + a_{011}x_0z_1 + a_{012}x_0z_2 + a_{110}x_1z_0 + a_{111}x_1z_1 + a_{112}x_1z_2 = 0,$$
$$a_{000}y_0z_0 + a_{001}y_0z_1 + a_{002}y_0z_2 + a_{010}y_1z_0 + a_{011}y_1z_1 + a_{012}y_1z_2 = 0,$$
$$a_{100}y_0z_0 + a_{101}y_0z_1 + a_{102}y_0z_2 + a_{110}y_1z_0 + a_{111}y_1z_1 + a_{112}y_1z_2 = 0,$$

has a non-trivial solution iff $\text{Det}_{2,2,3}(\mathcal{A}) = 0$. 
Cayley hyperdeterminant and tensor rank

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

### Theorem

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

$$A = x_1 \otimes x_2 \otimes y_3 + x_1 \otimes y_2 \otimes x_3 + y_1 \otimes x_2 \otimes x_3,$$

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

### Theorem (Kruskal)

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

**Theorem**

Let $A \in \mathbb{R}^{2\times2\times2}$. $\text{Det}_{2,2,2}(A) = 0$ iff

$$A = x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x$$

for some $x_i, y_i \in \mathbb{R}^2, i = 1, 2, 3$.

- Conditioning of the problem can be obtained from

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

- $x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x$ has outer product rank 3 generically (in fact, iff $x, y$ are linearly independent).

- Surprising: the manifold of ill-posed problem has full rank almost everywhere!
Nonnegative matrix factorization

- **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:
  \[ A = XY^\top = \sum_{i=1}^{r} x_i \otimes y_i. \]
- **Noisy situation**: approximate $A$ by a sum of outer-products of nonnegative vectors
  \[ \min_{X \geq 0, Y \geq 0} \| A - XY^\top \|_F = \min_{x_i \geq 0, y_i \geq 0} \| A - \sum_{i=1}^{r} x_i \otimes y_i \|_F. \]
Generalizing to hypermatrices

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

  $$A = \sum_{p=1}^{r} x_p \otimes y_p \otimes z_p$$

  where $x_p \in \mathbb{R}_+^l$, $y_p \in \mathbb{R}_+^m$, $z_p \in \mathbb{R}_+^n$.

- Clear that such a decomposition exists for any $A \geq 0$.

- **Nonnegative outer-product rank**: minimal $r$ for which such a decomposition is possible.

- Best nonnegative outer-product rank-$r$ approximation:

  $$\arg\min \|A - \sum_{p=1}^{r} x_p \otimes y_p \otimes z_p\|_F \mid x_p, y_p, z_p \geq 0.$$
Nonnegativity helps

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

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

\[H \circ X \leftarrow \leftarrow Y \leftarrow \leftarrow Z\]

Theorem (L-Comon)

The set \( \{ \mathcal{A} \in \mathbb{R}^{l \times m \times n}_+ \mid \text{rank}_+(\mathcal{A}) \leq r \} \) 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
  
  \[ a_{ijk} = \text{fluorescence emission intensity at wavelength } \lambda_{j}^{\text{em}} \text{ of } i\text{th sample excited with light at wavelength } \lambda_{k}^{\text{ex}}. \]
- Get 3-way data \( A = [a_{ijk}] \in \mathbb{R}^{l \times m \times n}. \)
- Get outer product decomposition of \( A \)
  \[
  A = x_{1} \otimes y_{1} \otimes z_{1} + \cdots + x_{r} \otimes y_{r} \otimes z_{r}.
  \]
- Get the true chemical factors responsible for the data.
  
  \( r: \) number of pure substances in the mixtures,
- \( x_{p} = (x_{1p}, \ldots, x_{lp}): \) relative concentrations of \( p\)th substance in specimens \( 1, \ldots, l, \)
- \( y_{p} = (y_{1p}, \ldots, y_{mp}): \) excitation spectrum of \( p\)th substance,
- \( z_{p} = (z_{1p}, \ldots, z_{np}): \) emission spectrum of \( p\)th substance.
- Noisy case: find best rank-\( r \) approximation (CANDDECOMP/PARAFAC).
Symmetric hypermatrices for blind source separation

Problem

Given \( y = Mx + n \). Unknown: source vector \( x \in \mathbb{C}^n \), mixing matrix \( M \in \mathbb{C}^{m \times n} \), noise \( n \in \mathbb{C}^m \). Known: observation vector \( y \in \mathbb{C}^m \). Goal: recover \( x \) from \( y \).

- Assumptions:
  1. components of \( x \) statistically independent,
  2. \( M \) full column-rank,
  3. \( n \) Gaussian.

- Method: use cumulants

\[
\kappa_k(y) = (M, M, \ldots, M) \cdot \kappa_k(x) + \kappa_k(n).
\]

By assumptions, \( \kappa_k(n) = 0 \) and \( \kappa_k(x) \) is diagonal. So need to diagonalize the symmetric hypermatrix \( \kappa_k(y) \).
Diagonalizing a symmetric hypermatrix

A best symmetric rank approximation may not exist either:

Example

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

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

and

$$
A := x \otimes y \otimes \cdots \otimes y + y \otimes x \otimes \cdots \otimes y + \cdots + y \otimes y \otimes \cdots \otimes x.
$$

Then $\text{rank}_S(A_n) \leq 2$, $\text{rank}_S(A) = k$, and

$$
\lim_{n \to \infty} A_n = A.
$$
Variational approach to eigenvalues/vectors

- $A \in \mathbb{R}^{m \times n}$ symmetric.
- Eigenvalues and eigenvectors are critical values and critical points of
  \[ \frac{x^\top Ax}{\|x\|_2^2}. \]
- Equivalently, critical values/points of $x^\top Ax$ constrained to unit sphere.
- Lagrangian:
  \[ L(x, \lambda) = x^\top Ax - \lambda(\|x\|_2^2 - 1). \]
- Vanishing of $\nabla L$ at critical $(x_c, \lambda_c) \in \mathbb{R}^n \times \mathbb{R}$ yields familiar
  \[ Ax_c = \lambda_c x_c. \]
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
  $$x^\top A y / \|x\|_2 \|y\|_2.$$ 
- Lagrangian:
  $$L(x, y, \sigma) = x^\top A y - \sigma(\|x\|_2 \|y\|_2 - 1).$$
- At critical $(x_c, y_c, \sigma_c) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$,
  $$A y_c / \|y_c\|_2 = \sigma_c x_c / \|x_c\|_2, \quad A^\top x_c / \|x_c\|_2 = \sigma_c y_c / \|y_c\|_2.$$
- Writing $u_c = x_c / \|x_c\|_2$ and $v_c = y_c / \|y_c\|_2$ yields familiar
  $$A v_c = \sigma_c u_c, \quad A^\top u_c = \sigma_c v_c.$$
Eigenvalues/vectors of a tensor

- Extends to hypermatrices.
- For $\mathbf{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, write $\mathbf{x}^p := [x_1^p, \ldots, x_n^p]^\top$.
- Define the ‘$\ell^k$-norm’ $\|\mathbf{x}\|_k = (x_1^k + \cdots + x_n^k)^{1/k}$.
- Define eigenvalues/vectors of $A \in S^k(\mathbb{R}^n)$ as critical values/points of the multilinear Rayleigh quotient
  \[
  A(\mathbf{x}, \ldots, \mathbf{x})/\|\mathbf{x}\|_k^k.
  \]
- Lagrangian
  \[
  L(\mathbf{x}, \lambda) := A(\mathbf{x}, \ldots, \mathbf{x}) - \lambda(\|\mathbf{x}\|_k^k - 1).
  \]
- At a critical point
  \[
  A(l_n, \mathbf{x}, \ldots, \mathbf{x}) = \lambda x^{k-1}.
  \]
Eigenvalues/vectors of a tensor

- If $A$ is symmetric,

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

- Also obtained by Liqun Qi independently:

- For unsymmetric hypermatrices — get different eigenpairs for different modes (unsymmetric matrix have different left/right eigenvectors).

- Falls outside Classical Invariant Theory — not invariant under $Q \in O(n)$, ie. $\|Qx\|_2 = \|x\|_2$.

- Invariant under $Q \in GL(n)$ with $\|Qx\|_k = \|x\|_k$. 

Singular values/vectors of a tensor

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

$$L(x, y, z, \sigma) = A(x, y, z) - \sigma(||x||||y||||z|| - 1)$$

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

$$A(l_l, y/||y||, z/||z||) = \sigma x/||x||,$$
$$A(x/||x||, l_m, z/||z||) = \sigma y/||y||,$$
$$A(x/||x||, y/||y||, l_n) = \sigma z/||z||.$$

- Normalize to get

$$A(l_l, v, w) = \sigma u, \quad A(u, l_m, w) = \sigma v, \quad A(u, v, l_n) = \sigma w.$$
Immediate properties

- Largest singular value is the norm of the multilinear functional associated with $A$ induced by the $p$-norm, i.e.
  \[ \sigma_{\text{max}}(A) = \|A\|_{p,\ldots,p}. \]

- For $d_1, \ldots, d_k$ such that
  \[ d_i - 1 \leq \sum_{j \neq i} (d_j - 1) \quad \text{for all } i = 1, \ldots, k, \]
  and $\text{Det}_{d_1,\ldots,d_k}$ 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
  \[ \text{Det}_{d_1,\ldots,d_k}(A) = 0. \]

- Pseudospectrum of square matrix $A \in \mathbb{C}^{n \times n}$,
  \[ \sigma_{\varepsilon}(A) = \{ \lambda \in \mathbb{C} \mid \| (A - \lambda I)^{-1} \|_2 > \varepsilon^{-1} \} = \{ \lambda \in \mathbb{C} \mid \sigma_{\text{min}}(A - \lambda I) < \varepsilon \}. \]

- Plausible generalizations to cubical hypermatrix $A \in \mathbb{C}^{n \times \cdots \times n}$,
  \[ \sigma_{\varepsilon}^{\Sigma}(A) = \{ \lambda \in \mathbb{C} \mid \sigma_{\text{min}}(A - \lambda I) < \varepsilon \} \]
  \[ \sigma_{\varepsilon}^{\Delta}(A) = \{ \lambda \in \mathbb{C} \mid \inf_{\text{Det}_n,\ldots,n(B)=0} \| A - \lambda I - B \|_F < \varepsilon^{-1} \}. \]
Perron-Frobenius theorem for hypermatrices

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

\[
\begin{bmatrix} b_{i_1 \ldots i_k} \end{bmatrix} = \begin{bmatrix} a_{\sigma(j_1) \ldots \sigma(j_k)} \end{bmatrix}
\]

has the property that for some \(m \in \{1, \ldots, n - 1\}\), \(b_{i_1 \ldots i_k} = 0\) for all \(i_1 \in \{1, \ldots, n - m\}\) and all \(i_2, \ldots, i_k \in \{1, \ldots, m\}\).

- We say that \(A\) is **irreducible** if it is not reducible. In particular, if \(A > 0\), then it is irreducible.

**Theorem (L)**

Let \(0 \leq A = \begin{bmatrix} a_{j_1 \ldots j_k} \end{bmatrix} \in T^k(\mathbb{R}^n)\) be irreducible. Then \(A\) has

1. a positive real eigenvalue \(\lambda\) with an eigenvector \(x\);
2. \(x\) may be chosen to have all entries non-negative;
3. if \(\mu\) is an eigenvalue of \(A\), then \(|\mu| \leq \lambda\).

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.
Spectral hypergraph theory

- Define the order-3 **adjacency hypermatrix** $A = [a_{ijk}]$ by
  \[
  a_{xyz} = \begin{cases} 
  1 & \text{if } [x, y, z] \in E, \\
  0 & \text{otherwise}.
  \end{cases}
  \]

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

- Consider cubic form
  \[
  A(f, f, f) = \sum_{x, y, z} a_{xyz} f(x)f(y)f(z),
  \]
  where $f \in \mathbb{R}^V$.

- Eigenvalues (resp. eigenvectors) of $A$ are the critical values (resp. critical points) of $A(f, f, f)$ constrained to the $f \in \ell^3(V)$, ie.
  \[
  \sum_{x \in V} f(x)^3 = 1.
  \]
We have the following.

### Lemma (L)

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

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

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 \( u, v, \ldots, z \) 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

1. If \( k \equiv 1 \mod 4 \), then every eigenvalue of \( G \) occurs with multiplicity a multiple of \( k \).
2. If \( k \equiv 3 \mod 4 \), then the spectrum of \( G \) is symmetric, ie. if \( \lambda \) is an eigenvalue, then so is \(-\lambda\).
3. Furthermore, every eigenvalue of \( G \) occurs with multiplicity a multiple of \( k/2 \), ie. if \( \lambda \) is an eigenvalue of \( G \), then \( \lambda \) and \(-\lambda \) occurs with the same multiplicity.
To do

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