Multilinear algebra in machine learning and signal processing

Pierre Comon and Lek-Heng Lim

ICIAM Minisymposium on Numerical Multilinear Algebra

July 17, 2007

Some metaphysics

- **Question:** What is numerical analysis?
- **One answer:** Numerical analysis is a functor.
- **Better answer:** Numerical analysis is a functor from the category of continuous objects to the category of discrete objects.
- Doug Arnold et. al.: observing functoriality yields better numerical methods (in terms of stability, accuracy, speed).
- **Numerical analysis:**

$$\text{CONTINUOUS} \longrightarrow \text{DISCRETE}$$

- **Machine learning:**

$$\text{DISCRETE} \longrightarrow \text{CONTINUOUS}$$

- **Message:** The continuous counterpart of a discrete model tells us a lot about the discrete model.

Tensors: mathematician's definition

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

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

  where $\alpha \in \mathbb{R}, \mathbf{u} \in U, \mathbf{v} \in V, \mathbf{w} \in W$.
- One condition: $\otimes$ decreed to have the multilinear property

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

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

Tensors: physicist's definition

- "What are tensors?" $\equiv$ "What kind of physical quantities can be represented by tensors?"
- Usual answer: if they satisfy some 'transformation rules' under a change-of-coordinates.

Theorem (Change-of-basis)

Two representations A, A' of $\mathbf{A}$ in different bases are related by

$$(L, M, N) \cdot A = A'$$

with L, M, N respective change-of-basis matrices (non-singular).

- Pitfall: tensor fields (roughly, tensor-valued functions on manifolds) often referred to as tensors — stress tensor, piezoelectric tensor, moment-of-inertia tensor, gravitational field tensor, metric tensor, curvature tensor.

Tensors: computer scientist's definition

- **Data structure:** k-array $A = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$
- **Algebraic structure:**
  1. **Addition/scalar multiplication:** for $[\![b_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$, $\lambda \in \mathbb{R}$,
  $$[\![a_{ijk}]\!] + [\![b_{ijk}]\!] := [\![a_{ijk} + b_{ijk}]\!] \quad \textit{and} \quad \lambda[\![a_{ijk}]\!] := [\![\lambda a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$$
  2. **Multilinear matrix multiplication:** for matrices $L = [\lambda_{i'i}] \in \mathbb{R}^{p \times l}$, $M = [\mu_{j'j}] \in \mathbb{R}^{q \times m}$, $N = [\nu_{k'k}] \in \mathbb{R}^{r \times n}$,
  $$(L, M, N) \cdot A := [\![c_{i'j'k'}]\!] \in \mathbb{R}^{p \times q \times r}$$
  where
  $$c_{i'j'k'} := \sum\nolimits_{i=1}^{l} \sum\nolimits_{j=1}^{m} \sum\nolimits_{k=1}^{n} \lambda_{i'i}\mu_{j'j}\nu_{k'k}a_{ijk}.$$
- Think of A as 3-dimensional array of numbers. $(L, M, N) \cdot A$ as multiplication on '3 sides' by matrices L, M, N.
- Generalizes to arbitrary order k. If $k = 2$, ie. matrix, then $(M, N) \cdot A = MAN^{\mathsf{T}}$.

Continuous data mining

- **Spectroscopy:** measure light absorption/emission of specimen as function of energy.
- Typical **specimen** contains 10^{13} to 10^{16} light absorbing entities or **chromophores** (molecules, amino acids, etc).

Fact (Beer's Law)

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

- Multiple chromophores ($k = 1, \dots, r$) and wavelengths ($i = 1, \dots, m$) and specimens/experimental conditions ($j = 1, \dots, n$),

$$A(\lambda_i, s_j) = \sum\nolimits_{k=1}^{r} \varepsilon_k(\lambda_i) c_k(s_j).$$

- Bilinear model aka **factor analysis**: $A_{m\times n} = E_{m\times r} C_{r\times n}$ rank-revealing factorization or, in the presence of noise, low-rank approximation $\min \| A_{m\times n} - E_{m\times r} C_{r\times n} \|$.

Discrete data mining

- **Text mining** is the spectroscopy of documents.
- Specimens = **documents** (n of these).
- Chromophores = **terms** (m of these).
- Absorbance = inverse document frequency:

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

- Concentration = term frequency: f_{ij}.
- $\sum_j \chi(f_{ij})/n$ = fraction of documents containing t_i.
- $A \in \mathbb{R}^{m\times n}$ term-document matrix. $A = QR = U\Sigma V^{\mathsf{T}}$ rank-revealing factorizations.
- Bilinear models:
  - Gerald Salton et. al.: **vector space model** (QR);
  - Sue Dumais et. al.: **latent sematic indexing** (SVD).
- Art Owen: what do we get when $m, n \to \infty$?

Bilinear models

- Bilinear models work on 'two-way' data:
  - measurements on object i (genomes, chemical samples, images, webpages, consumers, etc) yield a vector $\mathbf{a}_i \in \mathbb{R}^n$ where $n =$ number of features of i;
  - collection of m such objects, $A = [\mathbf{a}_1, \dots, \mathbf{a}_m]$ may be regarded as an m-by-n matrix, e.g. gene $\times$ microarray matrices in bioinformatics, terms $\times$ documents matrices in text mining, facial images $\times$ individuals matrices in computer vision.
- Various matrix techniques may be applied to extract useful information: QR, EVD, SVD, NMF, CUR, compressed sensing techniques, etc.
- Examples: vector space model, factor analysis, principal component analysis, latent semantic indexing, PageRank, EigenFaces.
- Some problems: **factor indeterminacy** — $A = XY$ rank-revealing factorization not unique; unnatural for k-**way data** when $k > 2$.

Ubiquity of multiway data

- **Batch data:** batch $\times$ time $\times$ variable
- **Time-series analysis:** time $\times$ variable $\times$ lag
- **Computer vision:** people $\times$ view $\times$ illumination $\times$ expression $\times$ pixel
- **Bioinformatics:** gene $\times$ microarray $\times$ oxidative stress
- **Phylogenetics:** codon $\times$ codon $\times$ codon
- **Analytical chemistry:** sample $\times$ elution time $\times$ wavelength
- **Atmospheric science:** location $\times$ variable $\times$ time $\times$ observation
- **Psychometrics:** individual $\times$ variable $\times$ time
- **Sensory analysis:** sample $\times$ attribute $\times$ judge
- **Marketing:** product $\times$ product $\times$ consumer

Outer product

- If $U = \mathbb{R}^l$, $V = \mathbb{R}^m$, $W = \mathbb{R}^n$, $\mathbb{R}^l \otimes \mathbb{R}^m \otimes \mathbb{R}^n$ may be identified with $\mathbb{R}^{l\times m\times n}$ if we define $\otimes$ by
$$\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} = [\![u_i v_j w_k]\!]_{i,j,k=1}^{l,m,n}.$$
- A tensor $A \in \mathbb{R}^{l\times m\times n}$ is said to be decomposable if it can be written in the form
$$A = \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$$
for some $\mathbf{u} \in \mathbb{R}^l, \mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$. For order 2, $\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^\mathsf{T}$.
- In general, any $A \in \mathbb{R}^{l\times m\times n}$ may be written as a sum of decomposable tensors
$$A = \sum\nolimits_{i=1}^{r} \lambda_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i.$$
- May be written as a multilinear matrix multiplication:
$$A = (U, V, W) \cdot \Lambda.$$
$U \in \mathbb{R}^{l\times r}, V \in \mathbb{R}^{m\times r}, W \in \mathbb{R}^{n\times r}$ and diagonal $\Lambda \in \mathbb{R}^{r\times r\times r}$.

Tensor ranks

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

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

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

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

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

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

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

Data analysis for numerical analysts

Idea

rank $\longrightarrow$ *rank revealing decomposition* $\longrightarrow$ *low-rank approximation* $\longrightarrow$ *data analytic model*

Fundamental problem of multiway data analysis

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

Examples

1. **Outer product rank:** $A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i$:
$$\min\|A - \mathbf{u}_1\otimes\mathbf{v}_1\otimes\mathbf{w}_1 - \mathbf{u}_2\otimes\mathbf{v}_2\otimes\mathbf{w}_2 - \cdots - \mathbf{u}_r\otimes\mathbf{v}_r\otimes\mathbf{z}_r\|.$$
2. **Multilinear rank:** $A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $C \in \mathbb{R}^{r_1\times r_2\times r_3}$, $L_i \in \mathbb{R}^{d_i\times r_i}$:
$$\min\|A - (L_1, L_2, L_3)\cdot C\|.$$
3. **Symmetric rank:** $A \in \mathsf{S}^k(\mathbb{C}^n)$, find $\mathbf{u}_i$:
$$\min\|A - \mathbf{u}_1^{\otimes k} - \mathbf{u}_2^{\otimes k} - \cdots - \mathbf{u}_r^{\otimes k}\|.$$
4. **Nonnegative rank:** $0 \leq A \in \mathbb{R}^{d_1\times d_2\times d_3}$, find $\mathbf{u}_i \geq 0, \mathbf{v}_i \geq 0, \mathbf{w}_i \geq 0$.

Feature revelation

- More generally, $\mathcal{D}$ = dictionary. Minimal r with

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

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

Examples

1. **parafac:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_{\otimes}(A) \leq 1\}$.
2. **Tucker:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_{\boxplus}(A) \leq (1, 1, 1)\}$.
3. **De Lathauwer:** $\mathcal{D} = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_{\boxplus}(A) \leq (r_1, r_2, r_3)\}$.
4. **ICA:** $\mathcal{D} = \{A \in \mathsf{S}^k(\mathbb{C}^n) \mid \operatorname{rank}_{\mathsf{S}}(A) \leq 1\}$.
5. **NTF:** $\mathcal{D} = \{A \in \mathbb{R}_+^{d_1 \times d_2 \times d_3} \mid \operatorname{rank}_+(A) \leq 1\}$.

Outer product decomposition in spectroscopy

- Application to fluorescence spectral analysis by Bro.
- 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 $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
  - Get outer product decomposition of A

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

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

Multilinear decomposition in bioinformatics

- Application to cell cycle studies by Alter and Omberg.
- Collection of gene-by-microarray matrices $A_1, \ldots, A_l \in \mathbb{R}^{m \times n}$ obtained under varying oxidative stress.
    - a_{ijk} = expression level of jth gene in kth microarray under ith stress.
    - Get 3-way data array $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.
    - Get multilinear decomposition of A

$$A = (X, Y, Z) \cdot C,$$

    to get orthogonal matrices X, Y, Z and core tensor C by applying SVD to various 'flattenings' of A.

- Column vectors of X, Y, Z are 'principal components' or 'parameterizing factors' of the spaces of stress, genes, and microarrays; C governs interactions between these factors.
- Noisy case: approximate by discarding small c_{ijk} (Tucker Model).

Bad news: outer product approximations are ill-behaved

- D. Bini, M. Capovani, F. Romani, and G. Lotti, "$O(n^{2.7799})$ complexity for $n \times n$ approximate matrix multiplication," *Inform. Process. Lett.*, **8** (1979), no. 5, pp. 234–235.
- Let $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$ be linearly independent. Define

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

- For $\varepsilon > 0$, define

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

- Then $\operatorname{rank}_\otimes(B_\varepsilon) \le 5$, $\operatorname{rank}_\otimes(A) = 6$ and $\|B_\varepsilon - A\| \to 0$ as $\varepsilon \to 0$.
- A has no optimal approximation by tensors of rank ≤ 5.

Worse news: ill-posedness is common

Theorem (de Silva and Lim)

1. *Tensors failing to have a best rank-r approximation exist for*
   1. *all* **orders** $k > 2$,
   2. *all* **norms** *and Brègman divergences,*
   3. *all* **ranks** $r = 2, \dots, \min\{d_1, \dots, d_k\}$.
2. *Tensors that fail to have best low-rank approximations occur with* **non-zero probability** *and sometimes with certainty — all* $2 \times 2 \times 2$ *tensors of rank* 3 *fail to have a best rank-2 approximation.*
3. *Tensor rank can* **jump arbitrarily large gaps**. *There exists sequence of rank-r tensor converging to a limiting tensor of rank* $r + s$.

Message

- That the best rank-r approximation problem for tensors has no solution poses serious difficulties.
- Incorrect to think that if we just want an 'approximate solution', then this doesn't matter.
- If there is no solution in the first place, then what is it that are we trying to approximate? ie. what is the 'approximate solution' an approximate of?
- Problems near an ill-posed problem are generally **ill-conditioned**.

CP degeneracy

- **CP degeneracy:** the phenomenon that individual rank-1 terms in PARAFAC solutions sometime diverges to infinity but in a way that the sum remains finite.
- **Example:** minimize $\|A - \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} - \mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z}\|$ via, say, alternating least squares,

$$\|\mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k\| \quad \text{and} \quad \|\mathbf{x}_k \otimes \mathbf{y}_k \otimes \mathbf{z}_k\| \to \infty$$

  but not

$$\|\mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k + \mathbf{x}_k \otimes \mathbf{y}_k \otimes \mathbf{z}_k\|.$$

- If a sequence of rank-r tensors converges to a limiting tensor of rank $> r$, then all rank-1 terms must become unbounded [de Silva and L].
- In other words, rank jumping always imply CP degeneracy.

Some good news: separation rank avoids this problem

- G. Beylkin and M.J. Mohlenkamp, "Numerical operator calculus in higher dimensions," *Proc. Natl. Acad. Sci.*, **99** (2002), no. 16, pp. 10246–10251.
- Given ε, find *small* $r(\varepsilon) \in \mathbb{N}$ so that

$$\|A - \mathbf{u}_1 \otimes \mathbf{v}_1 \otimes \mathbf{w}_1 - \mathbf{u}_2 \otimes \mathbf{v}_2 \otimes \mathbf{w}_2 - \cdots - \mathbf{u}_{r(\epsilon)} \otimes \mathbf{v}_{r(\epsilon)} \otimes \mathbf{z}_{r(\epsilon)}\| < \varepsilon.$$

- Great for compressing A.
- However, data analytic models sometime require a fixed, predetermined r.

More good news: weak solutions may be characterized

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

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

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

Theorem (de Silva and L)

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

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

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

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

Even more good news: nonnegative tensors are better behaved

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

$$\operatorname{rank}_+(A) := \min\left\{ r \mid \sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i,\ \mathbf{u}_i, \dots, \mathbf{z}_i \ge 0 \right\}$$

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

Theorem (Golub and L)

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

$$\inf\left\{ \left\| A - \sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{v}_i \otimes \cdots \otimes \mathbf{z}_i \right\| \,\middle|\, \mathbf{u}_i, \dots, \mathbf{z}_i \ge 0 \right\}$$

is always attained.

Corollary

Nonnegative tensor approximation always have solutions.

Continuous and semi-discrete PARAFAC

Khoromskij, Tyrtyshnikov: approximation by sum of separable functions

- Continuous PARAFAC

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

- Semi-discrete PARAFAC

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

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

- Discrete PARAFAC

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

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

Continuous and semi-discrete Tucker models

- Continuous Tucker model

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

- Semi-discrete Tucker model

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

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

- Discrete Tucker model

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

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

What continuous tells us about the discrete

Noisy case — approximation instead of exact decomposition. In both

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

and

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

we almost always want the functions θ, φ, ψ to come from some restricted subspaces of $\mathbb{R}^{\mathbb{R}}$ — eg. $L^p(\mathbb{R})$, $C^k(\mathbb{R})$, $C_0^k(\mathbb{R})$, etc.; or take some special forms — eg. splines, wavelets, Chebyshev polynomials, etc.

What continuous tells us about the discrete

View discrete models

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

and

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

as discretization of continuous counterparts.

Conditions on θ, φ, ψ tells us how to pick $\mathbf{u}, \mathbf{v}, \mathbf{w}$.

Example: probability densities

- X, Y, Z random variables, $f(x, y, z) = \Pr(X = x, Y = y, Z = z)$
- X, Y, Z conditionally independent upon some hidden H
- Semi-discrete PARAFAC — Naïve Bayes Model, Nonnegative Tensor Decomposition (Lee & Seung, Paatero), Probabilistic Latent Sematic Indexing (Hoffman)

$$\Pr(X = x, Y = y, Z = z) = \sum_{h=1}^{r} \Pr(H = h) \Pr(X = x \mid H = h) \Pr(Y = y \mid H = h) \Pr(Z = z \mid H = h)$$

Example: probability densities

- X, Y, Z random variables, $f(x, y, z) = \Pr(X = x, Y = y, Z = z)$
- X, Y, Z conditionally independent hidden X', Y', Z' (not necessarily independent)
- Semi-discrete Tucker — Information Theoretic Co-clustering (Dhillon et. al.) Nonnegative Tucker (Mørup et. al.)

$$\Pr(X = x, Y = y, Z = z) = \sum_{x',y',z'=1}^{p,q,r} \Pr(X' = x', Y' = y', Z' = z') \Pr(X = x \mid X' = x') \Pr(Y = y \mid Y' = y') \Pr(Z = z \mid Z' = z')$$

Coming Attractions

- Brett Bader and Tammy Kolda's minisympoisum on Thursday, 11:15–13:15 & 15:45–17:45, CAB G 51
- Speakers: Brett Bader, Morten Mørup, Lars Eldén, Evrim Acar, Lieven De Lathauwer, Derry FitzGerald, Giorgio Tomasi, Tammy Kolda

- Berkant Savas's talk on Thursday, 11:15, KO2 F 172
- Given $A \in \mathbb{R}^{l \times m \times n}$, want $\operatorname{rank}_{\boxplus}(B) = (r_1, r_2, r_3)$ with

$$\min \|A - B\|_F = \min \|A - (X, Y, Z) \cdot C\|_F$$

  $C \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, $X \in \mathbb{R}^{l \times r_1}$, $Y \in \mathbb{R}^{m \times r_2}$. Quasi-Newton method on a product of Grassmannians.

- Ming Gu's talk on Thursday, 16:15, KOL F 101
- The Hessian of $F(X, Y, Z) = \|A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$ can be approximated by a semiseparable matrix.