Optimal solutions to non-negative PARAFAC/multilinear NMF always exist

Lek-Heng Lim
Stanford University

Workshop on Tensor Decompositions and Applications
CIRM, Luminy, France
August 29–September 2, 2005

Thanks: Pierre Comon, Gene Golub, NSF DMS 01-01364

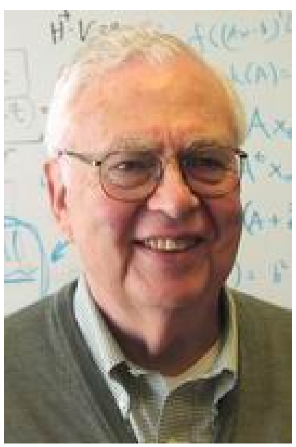
Acknowledgements

Rasmus Bro
Chemometrics Group
Royal Veterinary and Agricultural University

Richard Harshman
Department of Psychology
University of Western Ontario

Pierre Comon
Laboratoire I3S
Université de Nice Sophia Antipolis

Lieven de Lathauwer
Equipes Traitement des Images et du Signal
Ecole Nationale Supérieure d'Electronique et de ses Applications



Gene Golub
Department of Computer Science
Stanford University

Origin

Date: Fri, 10 Jun 2005 00:11:38 -0400
From: Richard A. Harshman <harshman@uwo.ca>
To: 'Lek-Heng Lim' <lekheng@stanford.edu>
Cc: "'Rasmus Bro (KVL)'" <rb@kvl.dk>,
'Lieven De Lathauwer' <Lieven.DeLathauwer@esat.kuleuven.be>
Subject: degeneracy, nonexistence of solutions and/or of problems?

Hi Rasmus, Lek-Heng, and Lieven,

Here is an interesting exchange between Rasmus and I that happened earlier today, but now seems relevant to our latest discussion. There are questions for Rasmus as well.

Please see below about Rasmus's lack of experiences with the "nonexistence of a solution" problem (also known as degenerate solutions) -- then an interesting apparent explanation and then an issue regarding positivity [non-negativity] constraints and degeneracies.

> -----Original Message-----
> From: Rasmus Bro (KVL) [mailto:rb@kvl.dk]
> Sent: Thursday, June 09, 2005 4:23 PM
> To: harshman@uwo.ca
> Subject: RE: hello, holiday greetings, asking about exchanges with
> Jos tenBerge
>
> Hi Richard
>
> Once I get degeneracies, they mostly reflect some problems

> in my specification of the problem. Once that problem
> is solved, the degeneracy usually disappear.
> > Do you ever enforce *positivity*[nonnegativity] on all modes,
> > and if so when? On some modes?
> I actually mostly do nonnegativity on all modes because it makes
> physical sense. Well, I always start without them, and then
> I actually mostly do nonnegativity on all modes because it makes
> physical sense. Well, I always start without them, and then
> if I run into problems, I start imposing nonnegativity. But I
> seldom think too much about what he problems is. Sometimes it's
> degeneracy, sometimes just 'ugly-looking' models.

[-----   Richard A. Harshman:]

So doesn't this absolutely prevent degenerate solutions from arising
when you do that? It's no wonder you don't have a problem with
degeneracies.

Perhaps Lek-Heng could work out what the characteristics of the
modified problem do about the existence of an exact minimum rather
than an infimum.

> Best
> Rasmus

Best regards,
--richard

Example: Analytical Chemistry

R. Bro, *Multi-way analysis in the food industry: models, algorithms, and applications*, Ph.D. thesis, Universiteit van Amsterdam, 1998.

$A = [\![a_{j_1 \cdots j_k}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ non-negative, denoted $A \geq 0$, if all $a_{j_1 \cdots j_k} \geq 0$.

a_{ijk} = fluorescence emission intensity at wavelength λ_j^{em} of ith sample excited with light at wavelength λ_k^{ex}. Get 3-way data $A = [\![a_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$.

Decomposing A into a sum of outer products,

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

yield the true chemical factors responsible for the data (in practice: *approximation* instead of *decomposition* because of the presence of noise).

- r: number of pure substances in the mixtures,
- $\mathbf{x}_\alpha = (x_{1\alpha}, \dots, x_{l\alpha})$: relative concentrations of αth substance in samples $1, \dots, l$,
- $\mathbf{y}_\alpha = (y_{1\alpha}, \dots, y_{m\alpha})$: emission spectrum of αth substance,
- $\mathbf{z}_\alpha = (z_{1\alpha}, \dots, z_{n\alpha})$: excitation spectrum of αth substance.

$\mathbf{x}_\alpha, \mathbf{y}_\alpha, \mathbf{z}_\alpha \geq 0$ — concentration and intensity cannot be negative.

Non-negative Matrix Factorization

D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, **401** (1999), pp. 788–791.

Central idea behind NMF (everything else is fluff): the way 'basis functions' combine to build 'target objects' is an exclusively additive process and should not involve any cancellations between the basis functions.

NMF in a nutshell: given non-negative matrix A, decompose it into a sum of outer-products of non-negative vectors:

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

Noisy situation: approximate A by a sum of outer-products of non-negative vectors

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

Generalizing NMF to Higher Order

Non-negative outer-product decomposition for $A \geq 0$ is

$$A = \sum_{p=1}^{r} \mathbf{x}_p^1 \otimes \cdots \otimes \mathbf{x}_p^k$$

where $\mathbf{x}_p^i \in \mathbb{R}_+^{d_i} := \{\mathbf{x} \in \mathbb{R}^{d_i} \mid x \geq 0\}$. Clear that such a decomposition exists for any $A \geq 0$.

Non-negative outer-product rank: minimal r for which such a decomposition is possible.

Optimal non-negative outer-product rank-r approximation:

$$\operatorname{argmin}\left\{\left\|A - \sum_{p=1}^{r} \mathbf{x}_p^1 \otimes \cdots \otimes \mathbf{x}_p^k\right\|_F \;\middle|\; \mathbf{x}_p^i \in \mathbb{R}_+^{d_i}\right\}. \qquad (\dagger)$$

In other words,

Multilinear NMF = Non-negative PARAFAC

Immediate Question

Since a general tensor can fail to have an optimal low-rank approximation (ie. A is degenerate or (†) is ill-posed), the first question that one should ask in a multilinear generalization of a bilinear model is whether the generalized problem still has a solution.

We will show that it does.

Degeneracy/Ill-posedness

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

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

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

and, for $\varepsilon > 0$,

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

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

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

Existence of Global Minimizer

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

$$\inf\Bigl\{\Bigl\|A - \textstyle\sum_{p=1}^{r} \mathbf{x}_p^1 \otimes \cdots \otimes \mathbf{x}_p^k\Bigr\|_F \;\Bigm|\; \mathbf{x}_p^i \in \mathbb{R}_+^{d_i}\Bigr\}$$

is attained.

Idea of proof: If a continuous real-valued function has a non-empty compact level set, then it has to attain its infimum. We will show that for a suitably redefined non-negative PARAFAC objective function, *all* its level sets are compact.

Removing Trivial Divergence

Naive choice of objective: $g : (\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_k})^r \to \mathbb{R}$,

$$g(\mathbf{x}_1^1, \ldots, \mathbf{x}_1^k, \ldots, \mathbf{x}_r^1, \ldots, \mathbf{x}_r^k) := \left\| A - \sum_{p=1}^r \mathbf{x}_p^1 \otimes \cdots \otimes \mathbf{x}_p^k \right\|_F^2.$$

Need to show g attains infimum on $(\mathbb{R}_+^{d_1} \times \cdots \times \mathbb{R}_+^{d_k})^r$.

Doesn't work because of an additional degree of freedom — $\mathbf{x}^1, \ldots, \mathbf{x}^k$ may be scaled by non-zero positive scalars that product to 1,

$$\alpha_1 \mathbf{x}^1 \otimes \cdots \otimes \alpha_k \mathbf{x}^k = \mathbf{x}^1 \otimes \cdots \otimes \mathbf{x}^k, \qquad \alpha_1 \cdot \cdots \cdot \alpha_k = 1,$$

e.g. $(n\mathbf{x}) \otimes \mathbf{y} \otimes (\mathbf{z}/n)$ can have a diverging loading vector even while the outer-product remains fixed.

Modified Objective Function

Define $f : \mathbb{R}^r \times (\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_k})^r \to \mathbb{R}$ by

$$f(X) := \left\| A - \sum_{p=1}^{r} \lambda_p \mathbf{u}_p^1 \otimes \cdots \otimes \mathbf{u}_p^k \right\|_F^2$$

where $X = (\lambda_1, \ldots, \lambda_r; \mathbf{u}_1^1, \ldots, \mathbf{u}_1^k, \ldots, \mathbf{u}_r^1, \ldots, \mathbf{u}_r^k)$. Let

$$\mathcal{D} := \mathbb{R}_+^r \times (\mathbb{S}_+^{d_1-1} \times \cdots \times \mathbb{S}_+^{d_k-1})^r,$$

$\mathbb{S}_+^{n-1} := \{\mathbf{x} \in \mathbb{R}_+^n \mid \|\mathbf{x}\| = 1\}$.

Global minimizer of f on $\mathcal{D}$, $(\lambda_1, \ldots, \lambda_r; \mathbf{u}_1^1, \ldots, \mathbf{u}_1^k, \ldots, \mathbf{u}_r^1, \ldots, \mathbf{u}_r^k) \in \mathcal{D}$, gives required global minimizer (albeit in a non-unique way), e.g. may take $\mathbf{x}_1^i = \lambda_i \mathbf{u}_1^i$ and $\mathbf{x}_p^i = \mathbf{u}_p^i$ for $p \geq 2$.

Compactness of Level Sets

Note $\mathcal{D}$ is closed but unbounded. Will show that the level set of f restricted to $\mathcal{D}$,

$$\mathcal{E}_\alpha = \{X \in \mathcal{D} \mid f(X) \leq \alpha\}$$

is compact for all α. $\mathcal{E}_\alpha = \mathcal{D} \cap f^{-1}(-\infty, \alpha]$ closed since f is polynomial, thus continuous.

Now to show $\mathcal{E}_\alpha$ bounded. Suppose there is a sequence $\{X_n\}_{n=1}^{\infty} \subset \mathcal{D}$ with $\|X_n\| \to \infty$ but $f(X_n) \leq \alpha$ for all n. Clearly, $\|X_n\| \to \infty$ implies that $\lambda_q^{(n)} \to \infty$ for at least one $q \in \{1, \ldots, r\}$.

By Cauchy-Schwartz,

$$f(X) \geq \left(\|A\|_F - \left\|\sum_{p=1}^{r} \lambda_p \mathbf{u}_p^1 \otimes \cdots \otimes \mathbf{u}_p^k\right\|_F\right)^2.$$

Taking $X \geq 0$ into account,

$$\begin{aligned}
\left\|\sum_{p=1}^{r} \lambda_p \mathbf{u}_p^1 \otimes \cdots \otimes \mathbf{u}_p^k\right\|_F^2 &= \sum_{i_1,\ldots,i_k=1}^{d_1,\ldots,d_k} \left(\sum_{p=1}^{r} \lambda_p u_{pi_1}^1 \cdots u_{pi_k}^k\right)^2 \\
&\geq \sum_{i_1,\ldots,i_k=1}^{d_1,\ldots,d_k} (\lambda_q u_{qi_1}^1 \cdots u_{qi_k}^k)^2 \\
&= \lambda_q^2 \sum_{i_1,\ldots,i_k=1}^{d_1,\ldots,d_k} (u_{qi_1}^1 \cdots u_{qi_k}^k)^2 \\
&= \lambda_q^2 \|\mathbf{u}_q^1 \otimes \cdots \otimes \mathbf{u}_q^k\|_F^2 \\
&= \lambda_q^2
\end{aligned}$$

since $\|\mathbf{u}_q^1\| = \cdots = \|\mathbf{u}_q^k\| = 1$.

Hence, as $\lambda_q^{(n)} \to \infty$, $f(X_n) \to \infty$. This contradicts $f(X_n) \leq \alpha$ for all n.

Uniqueness?

Uniqueness with conditions on Kruskal dimension (hard to check):

N. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemometrics*, **14** (2000), pp. 229–239.

J.M.F. ten Berge and N. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrika*, **67** (2002), no. 3, pp. 399–409.

Uniqueness with conditions on simplicial cones (bilinear only):

D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," *Adv. Neural Information Processing Systems*, **16** (2003), pp. 1141–1148.