Globally convergent algorithms for PARAFAC with semi-definite programming

Lek-Heng Lim

6th ERCIM Workshop on Matrix Computations and Statistics
Copenhagen, Denmark
April 1–3, 2005

Thanks: Vin de Silva, Gunnar Carlsson, Gene Golub,
NSF DMS 01-01364

Acknowledgement

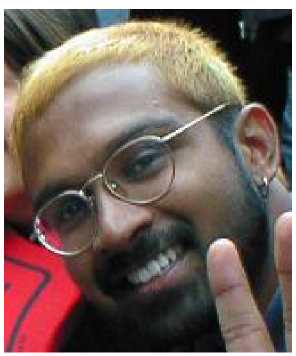
MATLAB Toolbox for Semidefinite Programming — SDPT3

Joint work with:

Kim-Chuan Toh

Department of Mathematics

National University of Singapore



Thanks to:

Vin de Silva

Department of Mathematics

Stanford University

Long term goal

Numerical Multilinear Algebra: Theory, Algorithms and Applications of Tensor Computations

- Develop a collection of standard computational methods for higher order tensors that parallel the methods that have been developed for order-2 tensors, ie. matrices
- Develop the mathematical foundations to facilitate this goal
- Applications

Motivation

Past 50 years, Numerical Linear Algebra played crucial role in:

- the statistical analysis of two-way data,
- the numerical solution of partial differential equations arising from vector fields,
- the numerical solution of second-order optimization methods.

Next step — develop Numerical Multilinear Algebra for:

- the statistical analysis of multi-way data,
- the numerical solution of partial differential equations arising from tensor fields,
- the numerical solution of higher-order optimization methods.

Outer product approximation

A Candecomp/Parafac or outer product model has the following form

$$a_{ijk} = \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} + e_{ijk}$$

where $E = [\![e_{ijk}]\!] \in \mathbb{R}^{l \times m \times n}$ denotes the (unknown) error.

To minimize the error, we want an outer product approximation

$$\text{argmin} \| A - \sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha \|_F$$

where the minimum is taken over all matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_r] \in \mathbb{R}^{l \times r}$, $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_r] \in \mathbb{R}^{m \times r}$, $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_r] \in \mathbb{R}^{n \times r}$.

In short, we want an optimal solution

$$B_\otimes^* = \underset{\text{rank}_\otimes(B) \leq r}{\text{argmin}} \|A - B\|_F.$$

Alternating least squares

Even when an optimal solution $B^*_\otimes$ to $\operatorname{argmin}_{\text{rank}_\otimes(B)\le r}\|A-B\|_F$ exists, $B^*_\otimes$ is not easy to compute since the objective function is non-convex.

A widely used strategy is a nonlinear Gauss-Seidel algorithm, better known as the Alternating Least Squares algorithm:

Algorithm: ALS for optimal rank-r approximation
initialize $X^{(0)} \in \mathbb{R}^{l\times r}, Y^{(0)} \in \mathbb{R}^{m\times r}, Z^{(0)} \in \mathbb{R}^{n\times r}$; initialize $s^{(0)}, \varepsilon > 0, k = 0$; while $\rho^{(k+1)}/\rho^{(k)} > \varepsilon$; $\quad X^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{X}\in\mathbb{R}^{l\times r}}\|T - \sum_{\alpha=1}^{r} \bar{x}_\alpha^{(k+1)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}\|_F^2$; $\quad Y^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Y}\in\mathbb{R}^{m\times r}}\|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes \bar{y}_\alpha^{(k+1)} \otimes z_\alpha^{(k)}\|_F^2$; $\quad Z^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Z}\in\mathbb{R}^{n\times r}}\|T - \sum_{\alpha=1}^{r} x_\alpha^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes \bar{z}_\alpha^{(k+1)}\|_F^2$; $\quad \rho^{(k+1)} \leftarrow \|\sum_{\alpha=1}^{r}[x_\alpha^{(k+1)} \otimes y_\alpha^{(k+1)} \otimes z_\alpha^{(k+1)} - x_\alpha^{(k)} \otimes y_\alpha^{(k)} \otimes z_\alpha^{(k)}]\|_F^2$; $\quad k \leftarrow k+1$;

Word of caution

A sequence $(\theta_k)_{k=1}^{\infty}$ is said to converge if $\lim_{k\to\infty}\theta_k$ exists.

An iterative algorithm for solving a particular problem is said to converge if the sequence of iterates $(\theta_k)_{k=1}^{\infty}$ is convergent *and* $\lim_{k\to\infty}\theta_k$ is the solution to that problem.

The sequence of iterates generate by ALS may be a convergent sequence but the ALS is not convergent as an algorithm for finding the optimal PARAFAC solution.

Pitfall: An algorithm that monotonically decreases the objective function must converge to the infimum/minimum of the function. (Not necessary, eg. $f_k = f(\theta_k) = 2 + \frac{1}{k}$ and $f^* = \inf_D f = 1$.

Some history

f polynomial in variables $\mathbf{x} = (x_1, \ldots, x_N)$. Suppose $f : \mathbb{R}^N \to \mathbb{R}$ non-negative valued, ie. $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^N$.

Question: Can we write f as a sum of squares of polynomials, ie. $p_1, \ldots, p_M$ such that

$$f(\mathbf{x}) = \sum_{j=1}^{M} p_j(\mathbf{x})^2 \quad ?$$

Answer (Hilbert): Not in general, eg. $f(w, x, y, z) = w^4 + x^2y^2 + y^2z^2 + z^2x^2 - 4xyzw$.

Hilbert's 17th Problem: Can we write f as a sum of squares of rational functions, ie. $p_1, \ldots, p_M$ and $q_1, \ldots, q_M$ such that

$$f(\mathbf{x}) = \sum_{j=1}^{M} \left(\frac{p_j(\mathbf{x})}{q_j(\mathbf{x})} \right)^2 \quad ?$$

Answer (Artin): Yes!

SDP-based algorithm

Observation 1:

$$
\begin{aligned}
F(x_{11}, \dots, z_{nr}) &= \|A - \textstyle\sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2 \\
&= \textstyle\sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \textstyle\sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} \right)^2
\end{aligned}
$$

is a polynomial of total degree 6 (resp. $2k$ for order k-tensors) in variables $x_{11}, \dots, z_{nr}$.

Recent breakthroughs in multivariate polynomial optimization [Lasserre 2001], [Parrilo 2003] [Parrilo-Sturmfels 2003] show that the non-convex problem

$$
\operatorname{argmin} F(x_{11}, \dots, z_{nr})
$$

may be relaxed to a convex problem (thus global optima is guranteed) which can in turn be solved using SDP.

How it works

Observation 2: If $F - \lambda$ can be expressed as a sum of squares of polynomials

$$F(x_{11}, \ldots, z_{nr}) - \lambda = \sum_{i=1}^{n} P_i(x_{11}, \ldots, z_{nr})^2,$$

then λ is a global lower bound for F, ie.

$$F(x_{11}, \ldots, z_{nr}) \geq \lambda$$

for all $x_{11}, \ldots, z_{nr} \in \mathbb{R}$.

Simple strategy: Find the largest λ^* such that $F - \lambda^*$ is a sum of squares. Then λ^* is often $\min F(x_{11}, \ldots, z_{nr})$.

Write $\mathbf{v} = (1, x_{11}, \ldots, z_{nr}, \ldots, x_{l1} y_{m1} z_{n1}, \ldots, z_{nr}^6)^t$, the D-tuple of monomials of total degree ≤ 6, where

$$D := \binom{r(l+m+n)+3}{3}.$$

Write $F(x_{11}, \ldots, z_{nr}) = \boldsymbol{\alpha}^t \mathbf{v}$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}^D$ are the coefficients of the respective monomials.

Since deg(F) is even, F may also be written as

$$F(x_{11}, \ldots, z_{nr}) = \mathbf{v}^t M \mathbf{v}$$

for some $M \in \mathbb{R}^{D \times D}$. So

$$F(x_{11}, \ldots, z_{nr}) - \lambda = \mathbf{v}^t (M - \lambda E_{11}) \mathbf{v}$$

where $E_{11} = \mathbf{e}_1 \mathbf{e}_1^t \in \mathbb{R}^{D \times D}$.

Observation 3: The rhs is a sum of squares iff $M - \lambda E_{11}$ is positive semi-definite (since $M - \lambda E_{11} = B^t B$).

Hence we have

$$\begin{array}{rl} \text{minimize} & -\lambda \\ \text{subjected to} & \mathbf{v}^t (S + \lambda E_{11}) \mathbf{v} = F, \\ & S \succeq 0. \end{array}$$

This is an SDP problem

$$\begin{array}{rll} \text{minimize} & 0 \circ S - \lambda & \\ \text{subjected to} & S \circ B_1 + \lambda = \alpha_1, & \\ & S \circ B_k = \alpha_k, & k = 2, \ldots, D \\ & S \succeq 0, & \lambda \in \mathbb{R}. \end{array}$$

This problem can be solved in polynomial time. Like all SDP-based algorithms, the SPD duality produces a certificate that tells us whether we have arrived at a globally optimal solution.

The *duality gap*, ie. difference between the values of the primal and dual objective functions, is 0 at a global minima.

Reducing the complexity

Complexity: For rank-r approximations to order-k tensors $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$,

$$D = \binom{r(d_1 + \cdots + d_k) + k}{k}$$

is large even for moderate d_i, r and k.

Sparsity to the rescue: The polynomials that we are interested in are always sparse (eg. for $k = 3$, only terms of the form xyz or $x^2y^2z^2$ or $uvwxyz$ appear). This can be exploited.

Newton polytope

Newton polytope of a polynomial f is the convex hull of the powers of the monomials in f.

Example. The Newton polytope of the polynomial $f(x,y) = 3.67x^4y^10 + -2.03x^3y^3 + 5.74x^3 - 20.1y^2 - 7.23$ is the convex hull of the points $(4,10), (3,3), (3,0), (2,0), (0,0)$ in $\mathbb{R}^2$.

Example. The Newton polytope of the polynomial $f(x,y,z) = 1.7x^4y^6z^2 + 7.4x^3z^5 - 3.0y^4 + 0.1yz^2$ is the convex hull of the points $(4,6,2), (3,0,5), (0,4,0), (0,1,2)$ in $\mathbb{R}^3$.

Theorem (Reznick). If $f(\mathrm{x}) = \sum_{i=1}^{m} p_i(\mathrm{x})^2$, then the powers of the monomials in p_i must lie in $\frac{1}{2}$Newton(f).

PARAFAC polynomial

The Newton polytope for a polynomial of the form

$$f(x_{11},\ldots,z_{nr}) = -\lambda + \sum_{i,j,k=1}^{l,m,n}\left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha}y_{j\alpha}z_{k\alpha}\right)^2$$

is spanned by 1 and monomials of the form $x_{i\alpha}^2 y_{j\alpha}^2 z_{k\alpha}^2$ (ie. monomials of the form $x_{i\alpha}y_{j\alpha}z_{k\alpha}$ and $x_{i\alpha}y_{j\alpha}z_{k\alpha}x_{i\beta}y_{j\beta}z_{k\beta}$ may all be dropped).

So if $f(x_{11},\ldots,z_{nr}) = \sum_{j=1}^{N} p_j(x_{11},\ldots,z_{nr})^2$, then only 1 and monomials of the form $x_{i\alpha}y_{j\alpha}z_{k\alpha}$ may occur in $p_1,\ldots,p_N$.

In other words, we have reduced the size of the problem from $\binom{r(l+m+n)+3}{3}$ to $rlmn+1$.

Global convergence issues

If polynomials of the form

$$-\lambda + \sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} \right)^2$$

can *always* be written as a sum of polynomials (we don't know), then the SDP algorithm for optimal low-rank tensor approximation will *always* converge globally.

Numerical experiments performed by Parrilo on general polynomials yield $\lambda^* = \min F$ in all cases.

Ill-posedness of PARAFAC: existence

Well known to practitioners in multiway data analysis, the problem $\operatorname{argmin}_{\mathrm{rank}_{\otimes}(B)\le r}\|A - B\|_F$ may not have an optimal solution when $r \ge 2$, $k \ge 3$. In fact

Theorem (L. and Golub, 2004). For tensors of any order $k \ge 3$ and with respect to any choice of norm on $\mathbb{R}^{d_1\times\cdots\times d_k}$, there exists an instance $A \in \mathbb{R}^{d_1\times\cdots\times d_k}$ such that A fails to have an optimal rank-r approximation for some $r \ge 2$. On the other hand, an optimal solution always exist for $k = 2$ and $r = 1$.

In the next slide, we give an explicit example.

Example

$\mathbf{x}, \mathbf{y}$ two linearly independent vectors in $\mathbb{R}^2$. Consider the order-3 tensor in $\mathbb{R}^{2\times 2\times 2}$,

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

A has rank 3: straight forward.

A has no optimal rank-2 approximation: consider sequence $\{B_n\}_{n=1}^{\infty}$ in $\mathbb{R}^{2\times 2\times 2}$,

$$B_n := \mathbf{x}\otimes\mathbf{x}\otimes(\mathbf{x}-n\mathbf{y}) + \left(\mathbf{x}+\frac{1}{n}\mathbf{y}\right)\otimes\left(\mathbf{x}+\frac{1}{n}\mathbf{y}\right)\otimes n\mathbf{y},$$

Clear that $\operatorname{rank}_{\otimes}(B_n) \leq 2$ for all n. By multilinearity of $\otimes$,

$$\begin{aligned} B_n = {} & \mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x} - n\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{y} + n\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{y} \\ & + \mathbf{x}\otimes\mathbf{y}\otimes\mathbf{y} + \mathbf{y}\otimes\mathbf{x}\otimes\mathbf{y} + \frac{1}{n}\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y} = A + \frac{1}{n}\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y}. \end{aligned}$$

For any choice of norm on $\mathbb{R}^{2\times 2\times 2}$,

$$\|A - B_n\| = \frac{1}{n}\|\mathbf{y}\otimes\mathbf{y}\otimes\mathbf{y}\| \to 0 \qquad \text{as } n\to\infty.$$

Quick but flawed fix

Current way to force a solution: perturb the problem by small $\varepsilon > 0$ and find approximate solution $\mathbf{x}_i^*(\varepsilon), \mathbf{y}_i^*(\varepsilon) \in \mathbb{R}^{d_i}$ $(i = 1, 2, 3)$ with

$$\begin{aligned}\|A - \mathbf{x}_1^*(\varepsilon) \otimes \mathbf{y}_1^*(\varepsilon) \otimes \mathbf{z}_1^*(\varepsilon) - \mathbf{x}_2^*(\varepsilon) \otimes \mathbf{y}_2^*(\varepsilon) \otimes \mathbf{z}_2^*(\varepsilon)\| \\ = \varepsilon + \inf_{\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{d_i}} \|A - \mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 - \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2\|.\end{aligned}$$

Serious numerical problems due to ill-conditioning (a phenomenon often referred to as *degeneracy* or *swamp* in Chemometrics and Psychometrics).

Reason? Rule of thumb in Computational Math:

A well-posed problem near to an ill-posed one is ill-conditioned.

So, even if we may perturb an ill-posed problem slightly to get a well-posed one, the perturbed problem will more often than not be ill-conditioned.

Weak solutions to PARAFAC

Theorem (de Silva and L., 2004). Let $l, m, n \geq 2$. Let $A \in \mathbb{R}^{l\times m\times n}$ with $\operatorname{rank}_{\otimes}(A) = 3$. A is the limit of a sequence $B_n \in \mathbb{R}^{l\times m\times n}$ with $\operatorname{rank}_{\otimes}(B_n) \leq 2$ if and only if

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

where $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\{\mathbf{y}_1, \mathbf{y}_2\}$, $\{\mathbf{z}_1, \mathbf{z}_2\}$ are linearly independent sets in $\mathbb{R}^l$, $\mathbb{R}^m$, and $\mathbb{R}^n$ respectively.

With this, we can overcome the ill-posedness of $\operatorname{argmin}_{\operatorname{rank}_{\otimes}(B)\leq r}\|A - B\|_F$ by replacing $\operatorname{rank}_{\otimes}$ with $\operatorname{closedrank}_{\otimes}$, defined by

$$\{A \mid \operatorname{closedrank}_{\otimes}(A) \leq r\} = \overline{\{A \mid \operatorname{rank}_{\otimes}(A) \leq r\}}.$$

For order-3 tensor, it follows from the theorem that

$$\begin{aligned}
&\{A \in \mathbb{R}^{l\times m\times n} \mid \operatorname{closedrank}_{\otimes}(A) \leq 2\} = \\
&\qquad \{\mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_1 \otimes \mathbf{z}_2 + \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_1 \mid \mathbf{x}_i \in \mathbb{R}^l, \mathbf{y}_i \in \mathbb{R}^m, \mathbf{z}_i \in \mathbb{R}^n\} \\
&\qquad\qquad \cup \{\mathbf{x}_1 \otimes \mathbf{y}_1 \otimes \mathbf{z}_1 + \mathbf{x}_2 \otimes \mathbf{y}_2 \otimes \mathbf{z}_2 \mid \mathbf{x}_i \in \mathbb{R}^l, \mathbf{y}_i \in \mathbb{R}^m, \mathbf{z}_i \in \mathbb{R}^n\}
\end{aligned}$$

Ill-posedness of PARAFAC: uniqueness

Note that in PARAFAC:

$$\text{argmin}\|A - \textstyle\sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F,$$

we are really interested in minimizer $X^* = [\mathbf{x}_1^*, \ldots, \mathbf{x}_r^*] \in \mathbb{R}^{l \times r}$, $Y^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_r^*] \in \mathbb{R}^{m \times r}$, $Z^* = [\mathbf{z}_1^*, \ldots, \mathbf{z}_r^*] \in \mathbb{R}^{n \times r}$ rather than the minimum value.

If X^*, Y^*, Z^* is a minimizer, then so is $X^* D_1, Y^* D_2, Z^* D_3$ for any diagonal $D_1, D_2, D_3 \in \mathbb{R}^{r \times r}$ with $D_1 D_2 D_3 = I$.

In fact, the SDP method will not work if there is an infinite number of possible minimizers.

Right now, we impose constraints (eg. requiring $\|\mathbf{y}_\alpha\| = \|\mathbf{z}_\alpha\| = 1$) to get uniqueness up to signs but every additional constraint increases the complexity of the problem.