Chain Tensor Network: Instability and how to deal with it

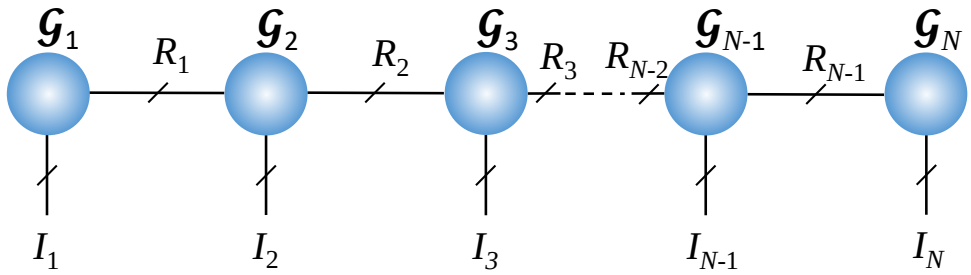
CAIT
Skolkowo Institute of Science and Technology (SKOLTECH), Moscow, Russia

February 28, 2022

Outlines

- Looped Tensor network
- Relation between TC and Tucker, CP decompositions
- Sensitivity
- Algorithms

MPS/Tensor Train - Straight Tensor Network I



Model a tensor $\mathcal{Y}$ of size $I_1 \times I_2 \times \cdots \times I_N$ by N interconnected core tensors $\mathcal{G}_n$ of size $R_n \times I_n \times R_{n+1}$

$$\mathcal{Y} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_{N-1}=1}^{R_{N-1}} \mathcal{G}_1(:, r_1) \circ \mathcal{G}_2(r_1, :, r_2) \circ \cdots \circ \mathcal{G}_N(r_{N-1}, :),$$

or

$$\mathcal{Y} = \mathcal{G}_1 \bullet \mathcal{G}_2 \bullet \cdots \bullet \mathcal{G}_{N-1} \bullet \mathcal{G}_N$$

$(R_1, R_2, \ldots, R_{N-1})$ represents the TT-rank of $\mathcal{Y}$.

MPS/Tensor Train - Straight Tensor Network II

- TT/MPS decomposition can be computed by tool of SVDs Vidal (2003); Oseledets and Tyrtyshnikov (2009); Phan et al. (2020a)
- TT/MPS decomposition is very suited to higher-order tensors.
- Applications: solving a huge system of linear equations or eigenvalue decomposition of large-scale data Holtz et al. (2012); Kressner et al. (2014), PDE, data completion, modelling in system identification, deep learning.

Major problem: Intermediate TT-ranks is often very high

- TT decomposition with bound constraint often exhibit badly unbalaced TT-ranks.
- The ranks grow dramatically with the dimensions of the tensor.

Chain tensor network - Tensor chain I

Figure: Looped tensor networks.

- Khoromskij (2009,2011) introduced the looped Tensor chain as an extension of TT
- Since there are no first and last core tensors, TC is expected to overcome imbalance rank issue in TT decomposition.

Chain tensor network - Tensor chain II

Similar to TT, core tensors $\mathcal{G}_n$ of size $R_{n-1} \times I_n \times R_n$, $R_0 = R_N$, the TC model is written as

$$\mathcal{Y} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_N=1}^{R_N} \mathcal{G}_1(r_N, :, r_1) \circ \mathcal{G}_2(r_1, :, r_2) \circ \cdots \circ \mathcal{G}_N(r_{N-1}, :, r_N),$$

$$y_{i_1 i_2 \cdots i_N} = \mathrm{tr}(\mathbf{G}_1(:, i_1, :)\mathbf{G}_2(:, i_2, :) \cdots \mathbf{G}_N(:, i_N, :))$$

Shorthand notation for TC

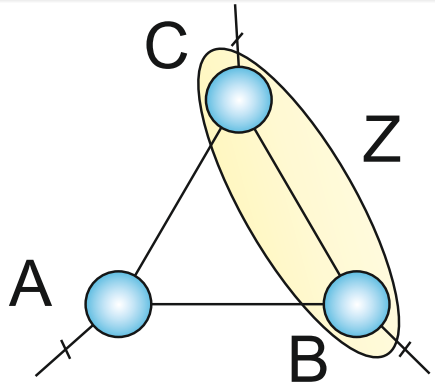
$$\mathcal{Y} = \circledcirc \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N \circledcirc$$

When looped network opens

When all vertical slides of one of the core tensors, e.g., $\mathcal{G}_N$, are identity matrices, the TC model becomes TT model

$$\mathcal{Y}_N = \mathcal{G}_1 \bullet \mathcal{G}_2 \bullet \cdots \bullet \mathcal{G}_{N-1}$$

ALS Algorithm I



- Consider an order-3 TC decomposition

$$\min \quad \|\mathcal{Y} \approx \mathbb{\{}\mathcal{A}, \mathcal{B}, \mathcal{C}\mathbb{\}}\|_F^2$$

- In order to update the core tensor $\mathcal{A}$, we minimize the Frobenius norm Espig et al. (2011, 2012)

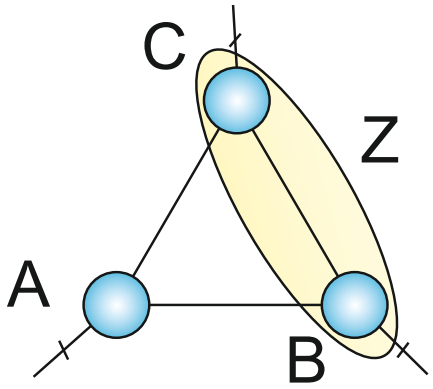
$$\min \quad \|\mathbf{Y}_{(1)} - \mathbf{A}_{(2)}\mathbf{Z}\|_F^2$$

where $\mathbf{Y}_{(1)}$ is mode-1 unfolding of $\mathcal{Y}$, $\mathbf{A}_{(2)}$ is mode-2 unfolding of $\mathcal{A}$, $\mathbf{Z}$ is mode-(1,3) unfolding of $\mathcal{B} \bullet \mathcal{C}$,

- The update rule reads

$$\mathbf{A}_{(2)} = \mathbf{Y}_{(1)}\mathbf{Z}^{\dagger}$$

ALS Algorithm I

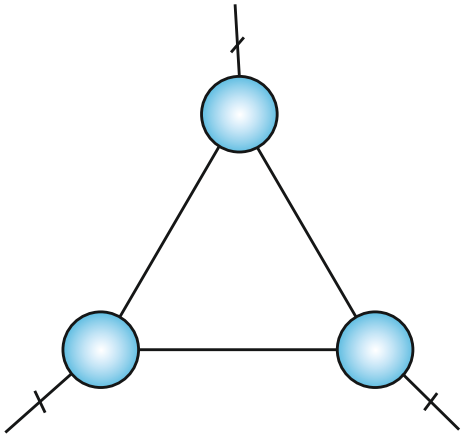


- ALS update rule works efficiently like ALS for other tensor decompositions
- Density-Matrix Renormalization Group (Espig et al. (2011)) and similar ALS algorithms have recently been reinvented or proposed for decomposition of incomplete data

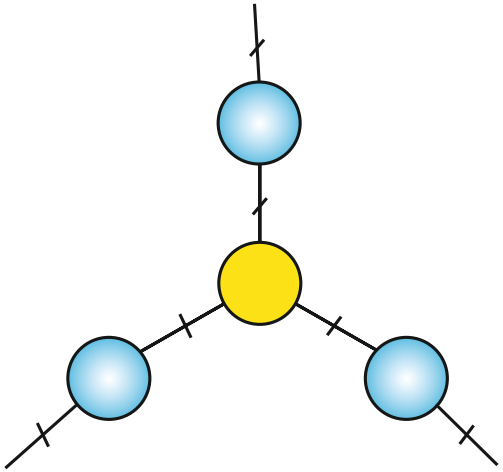
Major problem: Instability

- Unfortunately, loop in TC may lead to severe numerical instability in finding the best TC model, especially when $R_n R_{n+1} > I_n$ (Landsberg (2012) and Handschuh (2015))

TC and Structured Tucker model I



(a) TC-3



(b) Tucker-3

- *Matrix-multiplication tensor* $\mathcal{M}$ obeys the relation

$$\text{vec}(\mathbf{EF}) = \mathcal{M} \times_1 \text{vec}(\mathbf{E})^T \times_2 \text{vec}(\mathbf{F})^T \quad (1)$$

for all matrices **E** and **F** of the size $R_3 \times R_1$ and $R_1 \times R_2$.

TC and Structured Tucker model II

- Applying this relation to order-3 TC tensor, $\mathcal{Y} = \{\!\{\mathcal{A}, \mathcal{B}, \mathcal{C}\}\!\}$ gives

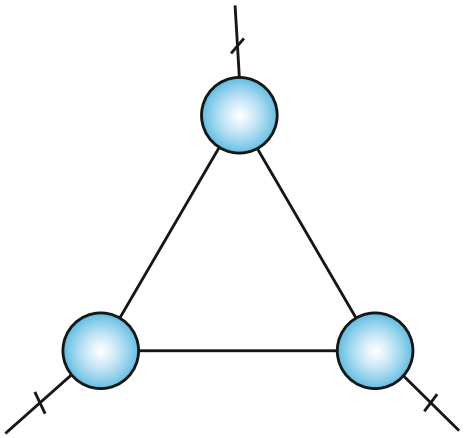
$$\begin{aligned} y_{ijk} &= \mathrm{tr}(\mathbf{A}_i \mathbf{B}_j \mathbf{C}_k) \\ &= \mathrm{vec}(\mathbf{C}_k)^T \mathrm{vec}(\mathbf{A}_i \mathbf{B}_j) \\ &= \mathrm{vec}(\mathbf{C}_k)^T (\mathcal{M} \times_1 \mathrm{vec}(\mathbf{A}_i)^T \times_2 \mathrm{vec}(\mathbf{B}_j)^T) \\ &= \mathcal{M} \times_1 \mathrm{vec}(\mathbf{A}_i)^T \times_2 \mathrm{vec}(\mathbf{B}_j)^T \times_3 \mathrm{vec}(\mathbf{C}_k)^T . \end{aligned}$$

It follows that

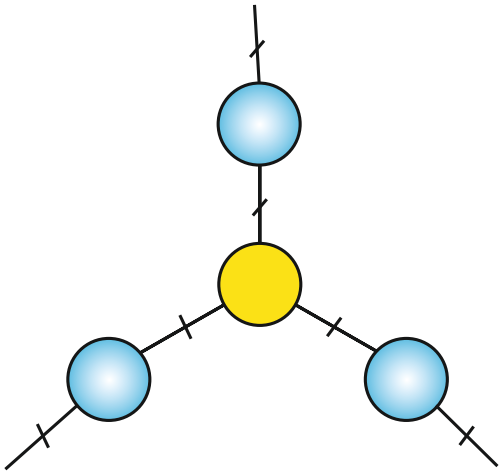
$$\begin{aligned} \mathcal{Y} &= \mathcal{M} \times_1 \mathbf{A}_{(2)} \times_1 \mathbf{B}_{(2)} \times_3 \mathbf{C}_{(2)} \\ &= [\![\mathcal{M}; \mathbf{A}_{(2)}, \mathbf{B}_{(2)}, \mathbf{C}_{(2)}]\!] \end{aligned}$$

where $\mathbf{A}_{(2)}, \mathbf{B}_{(2)}, \mathbf{C}_{(2)}$ are mode-2 unfoldings of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, respectively.

TC and Structured Tucker model III



(c) TC-3



(d) Tucker-3

Order-3 TC is a structured Tucker decomposition with fixed known core tensor.

Similar relation can be established for higher order tensors.

TC and PARallel FACtor with LINear Depedences I

- CANonical DEcomposition with LINear Constraints Carroll et al. (1970) PARallel FACtor with LINear DepedencesBro et al. (2009)

$$\mathcal{Y} = [\![\mathbf{AU}, \mathbf{BV}, \mathbf{CW}]\!]$$

  where **U**, **U** and **W** are known dependence matrices.
- Note that multiplication tensor, $\mathcal{M}$, can be represented by a Canonical tensor model with binary factor matrices
- For example, $\mathcal{M}$ of size $4 \times 4 \times 4$ has rank-7

$$\mathbf{M}_{(1)} = \left[\begin{array}{cccc|cccc|cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$\mathcal{M} = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

where

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & -1 \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

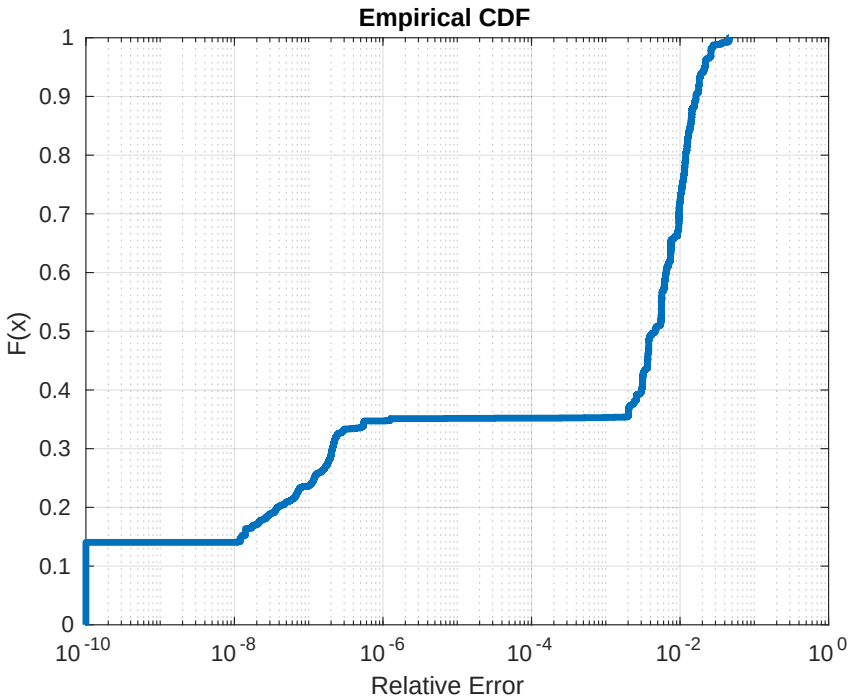
TC is PARALIND

Given the CP decomposition of $\mathcal{M} = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$, TC decomposition of $\mathcal{Y}$ becomes

$$\begin{aligned}
\mathcal{Y} &= \Big\Vert \mathcal{A}, \mathcal{B}, \mathcal{C} \Big\Vert && \text{(TC model)} \\
&= [\![\mathcal{M}; \mathbf{A}_{(2)}, \mathbf{B}_{(2)}, \mathbf{C}_{(2)}]\!] && \text{(Tucker model)} \\
&= [\![\mathbf{A}_{(2)}\mathbf{U}, \mathbf{B}_{(2)}\mathbf{V}, \mathbf{C}_{(2)}\mathbf{W}]\!] && (\textit{PARALIND})
\end{aligned}$$

Fitting a TC/TT model is not much different from seeking a structured Tucker or PARALIND model.

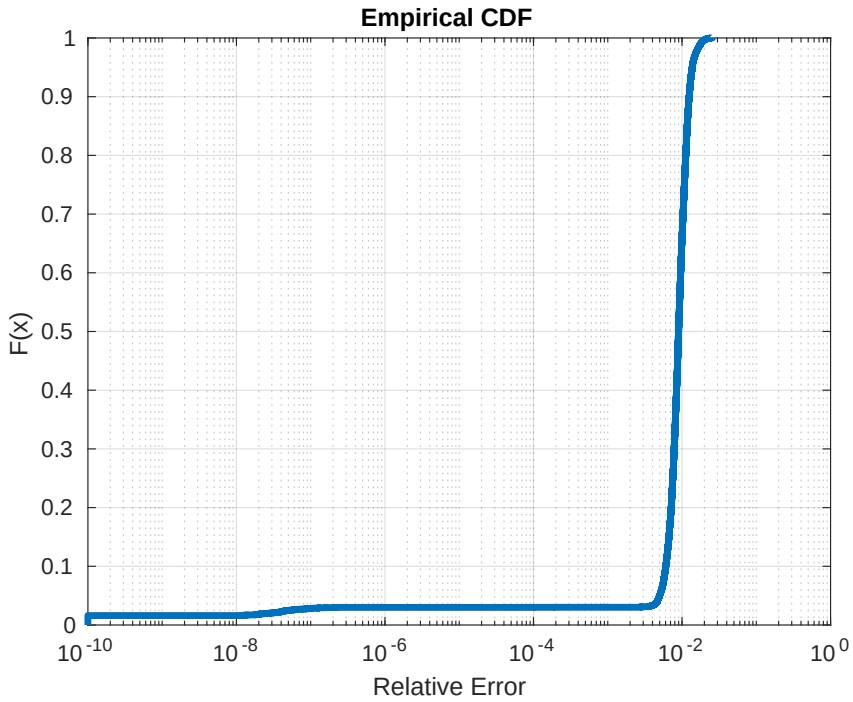
First Example



Decompose synthetic tensor of size $4 \times 4 \times 4$, rank-(2-2-2).

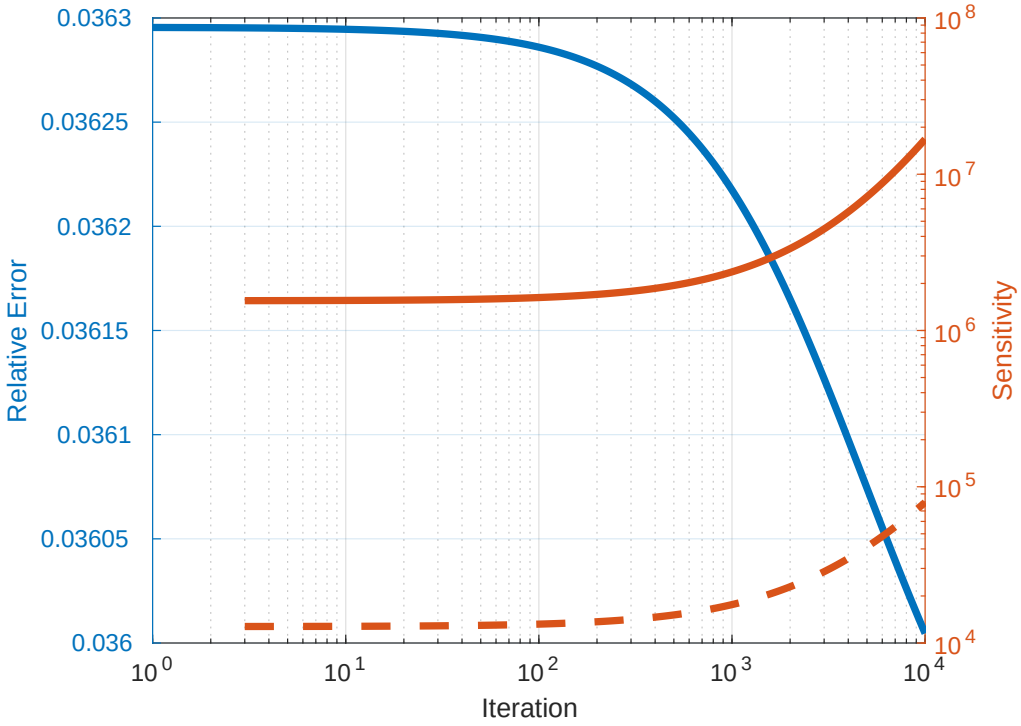
- Tensors are composed from 3 core tensors of size $2 \times 4 \times 2$, generated randomly
- Decomposed using ALS in 5000 iterations with parameters initialized randomly
- Success rate is only 36%.

Second Example (Much harder) I



- Noise free synthetic tensor of size $7 \times 7 \times 7$ with rank-(3-3-3), randomly generated.
- Tensors are decomposed using ALS within 5000 iterations with parameters initialized randomly
- Success rate is much worse, less than 3%.

Second Example (Much harder) II

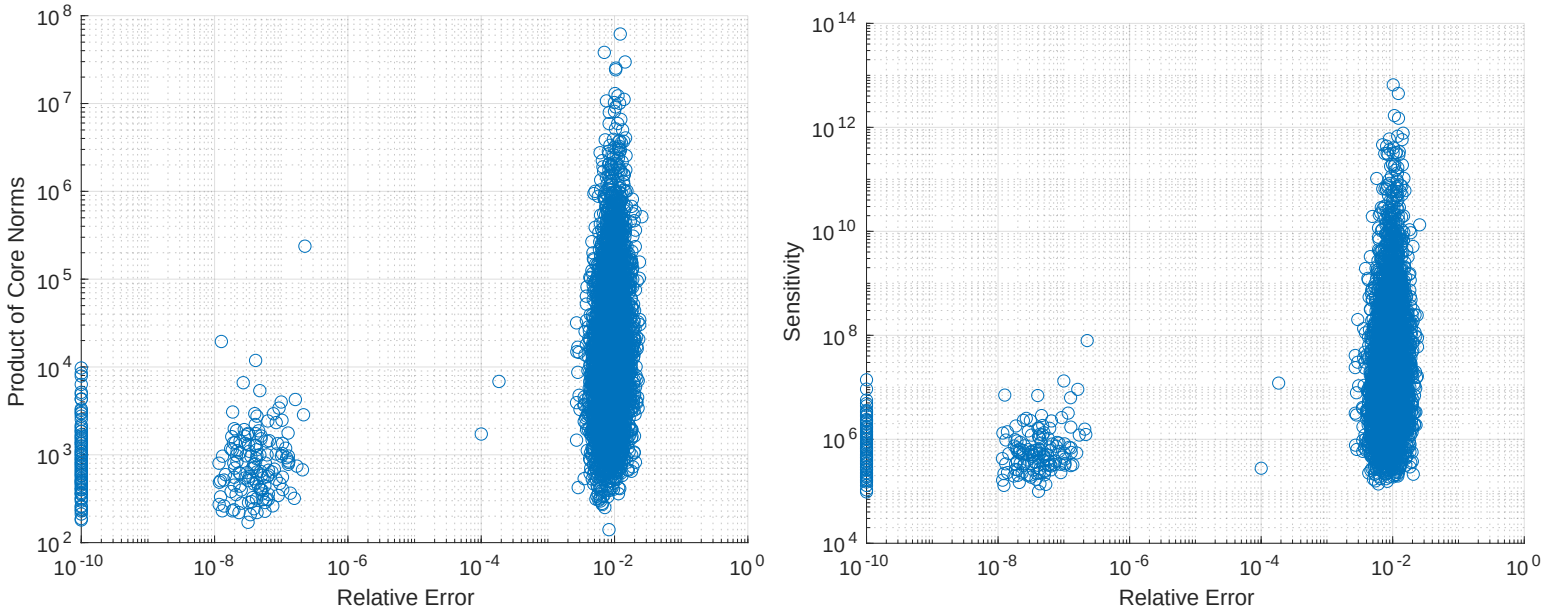


Why ALS and many other algorithms for TC fail

First observation: norms of core tensors quickly increase after several thousand iterations

Approximation gets stuck in false local minima.

Second Example (Much harder) III



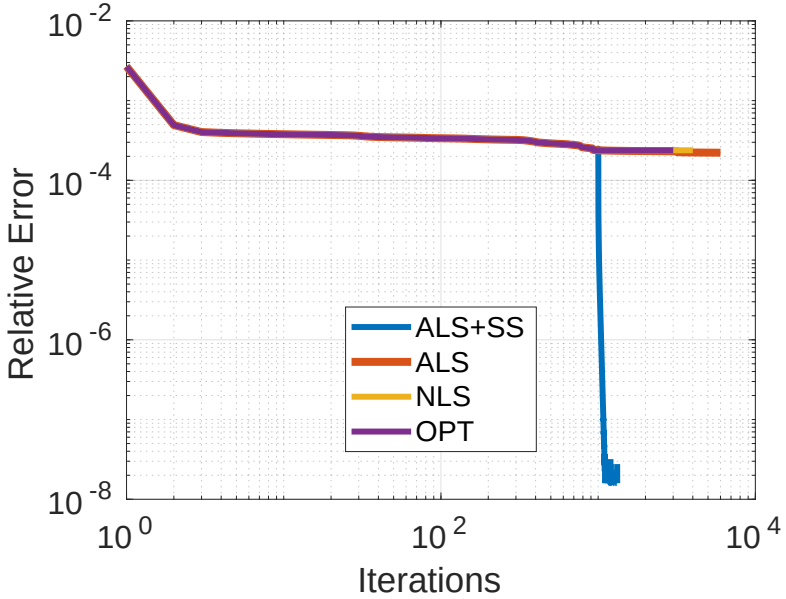
Why ALS and many other algorithms for TC fail

- Instability, especially when the ranks $R_{n-1}R_n > I_n$
- Sensitivity of the estimated model is significantly high, prevents the algorithm from converging to the exact model.

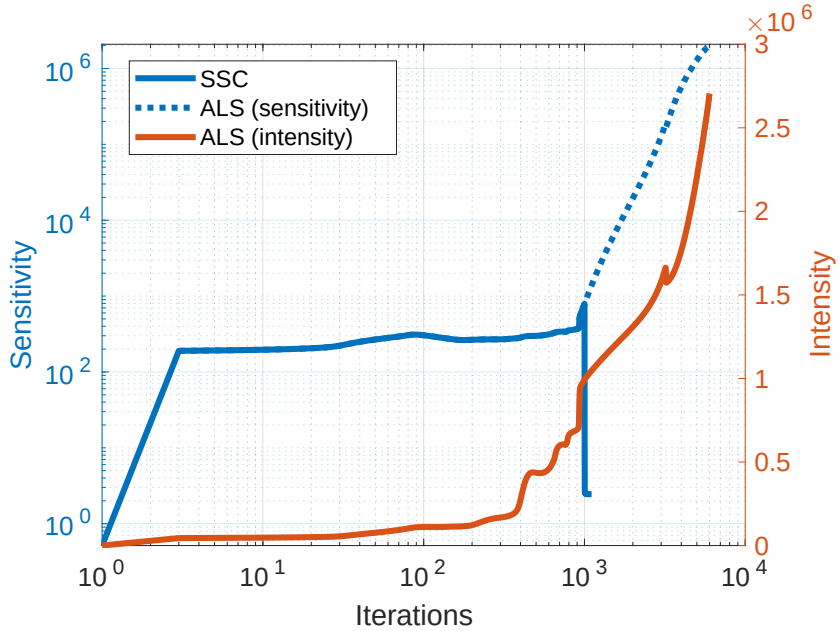
(Relative error vs Sensitivity over 10000 independent runs (100 tensors x 100 decompositions))

Example:TC with highly collinear loading components I

We decompose order-3 tensors of size $27 \times 27 \times 27$ which admit the TC model, $\mathcal{Y} = \llparenthesis \mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3 \rrparenthesis$, with bond dimensions $(5 - 5 - 5)$. The factor matrices $\mathbf{U}_n$ of size 27×25 have highly collinear loading components, $0.97 \leq \mathbf{U}_n^T(:, r)\mathbf{U}_n(:, s) \leq 0.99$, $n = 1, 2, 3$.



(e)



(f)

Example:TC with highly collinear loading components II

ALS failed in this example. The intensity (dashed red curve) of the estimated TC tensors increased quickly and exceeded 2.7×10^6, whereas its sensitivity (dotted blue curve) passed the level of 10^6 after 6000 iterations

The OPT(WOPT) **?** and NLS algorithms Sorber et al. (2013) also failed.

TC for incomplete data I

We demonstrate a simple TC decomposition for tensors of size $9 \times 9 \times 9$ with bond dimensions $(3 – 3 – 3)$.
The considered tensors can be factorized quickly. However, when 50% of the tensor elements are randomly removed, the tensors are challenging to any TC algorithms. The success rate for OPT**?** and ALS is less than 11%.
The algorithms get stuck in local minimal and cannot attain exact decomposition.

TC for incomplete data II

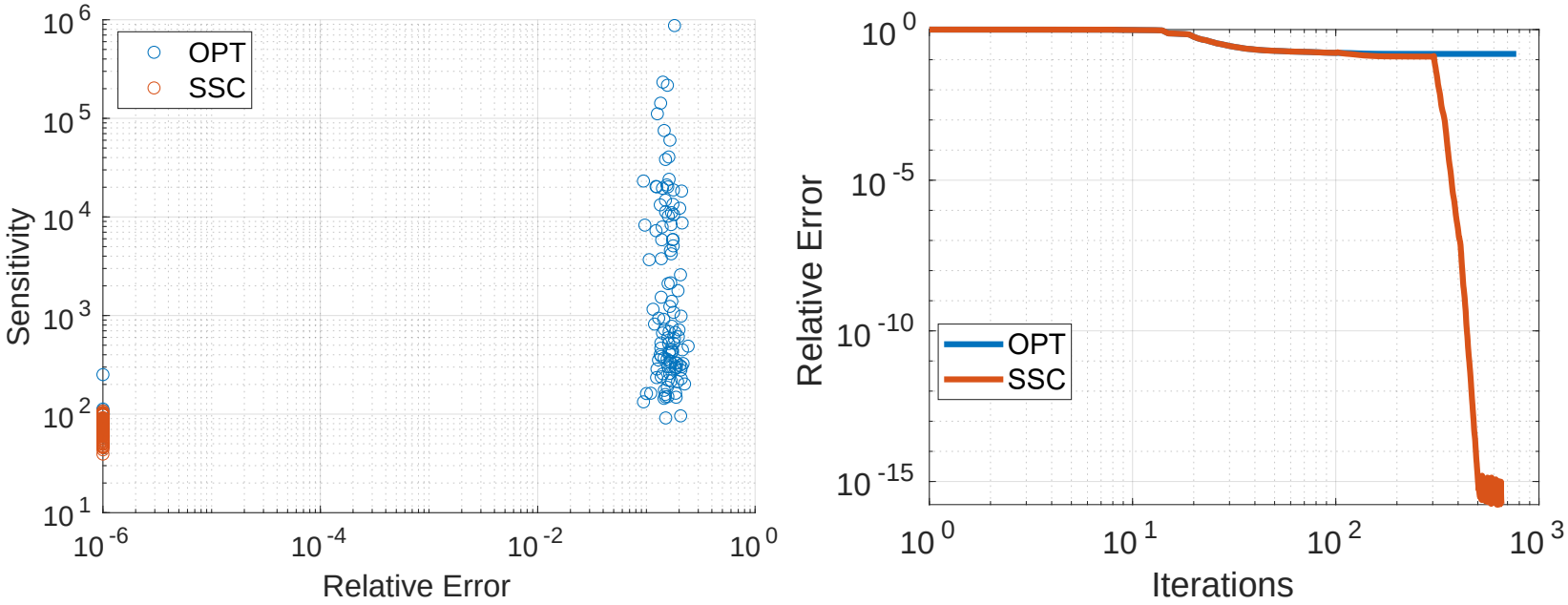


Figure: TC for incomplete tensor. With sensitivity correction, the decomposition can obtain the exact model.

Intensity and Sensitivity I

The TC model, $\mathcal{Y} = \llparenthesis \mathcal{A}, \mathcal{B}, \mathcal{C} \rrparenthesis$, is not unique up to scaling

$$\mathcal{Y} = \llparenthesis \alpha_1 \mathcal{A}, \alpha_2 \mathcal{B}, \alpha_3 \mathcal{C} \rrparenthesis \tag{2}$$

with arbitrary factors α_1, α_2, and α_3 such that $\alpha_1 \alpha_2 \alpha_3 = 1$.
The TC model, $\mathcal{Y} = \llparenthesis \mathcal{A}, \mathcal{B}, \mathcal{C} \rrparenthesis$, is also non-unique up to rotation

$$\mathcal{Y} = \llparenthesis \mathcal{A} \bullet \mathbf{Q}, \mathbf{Q}^{-1} \bullet \mathcal{B}, \mathcal{C} \rrparenthesis \tag{3}$$

where $\mathbf{Q}$ is an arbitrary invertible matrix of size $R_2 \times R_2$.

Definition (TC intensity.)

For a given TC model, $\mathcal{Y} = \llparenthesis \mathcal{A}, \mathcal{B}, \mathcal{C} \rrparenthesis$, we can always normalize core tensors to unit norm, $\tilde{\mathcal{A}} = \mathcal{A}/\|\mathcal{A}\|_F$,
$\tilde{\mathcal{B}} = \mathcal{B}/\|\mathcal{B}\|_F, \tilde{\mathcal{C}} = \mathcal{C}/\|\mathcal{C}\|_F$ then $\mathcal{Y} = \alpha \llparenthesis \tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}} \rrparenthesis$ where
$\alpha = \|\mathcal{A}\|_F \|\mathcal{B}\|_F \|\mathcal{C}\|_F$ is called the TC intensity.

Intensity and Sensitivity II

Lemma (TC Degeneracy)

For a given TC model, $\mathcal{Y} = \llparenthesis \mathcal{A}, \mathcal{B}, \mathcal{C} \rrparenthesis$, there is always a sequence of equivalent TC models with diverging TC intensities.

We provide an example as a proof for the TC model with rank $R_2 = 2$. The other cases can be seen straightforwardly. Consider the sub-network $\mathcal{A} \bullet \mathcal{B}$, apply the DMRG-like update rule to split it to a sequence of three cores,

$$\mathcal{A} \bullet \mathcal{B} = \mathcal{U} \bullet \mathbf{S} \bullet \mathcal{V}$$

where $\mathbf{U}\,\mathbf{S}\,\mathbf{V}$ is thin-SVD of unfolding of $\mathcal{A} \bullet \mathcal{B}$ to a matrix of size $R_1 I_1 \times I_2 R_3$, $\mathbf{S} = \text{diag}(s_1, s_2)$ is a diagonal matrix of $R_2 = 2$ leading singular values, $\mathbf{U}$ and $\mathbf{V}$ are unfoldings of $\mathcal{U}$ and $\mathcal{V}$, respectively. The tensor $\mathcal{Y}$ has an equivalent TC model $\mathcal{Y} = \llparenthesis \mathcal{U}, \mathbf{S} \bullet \mathcal{V}, \mathcal{C} \rrparenthesis$.

Intensity and Sensitivity III

We next define a matrix $\mathbf{Q} = \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}$. Note that

$\mathbf{Q}^{-1} = \frac{1}{1-x^2} \begin{bmatrix} 1 & -x \\ -x & 1 \end{bmatrix}$

The tensor $\mathcal{Y}$ has another equivalent TC model given by

$$\mathcal{Y} = \llbracket \mathcal{U} \bullet \mathbf{Q}, \mathbf{Q}^{-1}\mathbf{S} \bullet \mathcal{V}, \mathcal{C} \rrbracket \tag{4}$$

but with an intensity

$$\begin{aligned} \alpha &= \|\mathcal{U} \bullet \mathbf{Q}\|_F \, \|\mathbf{Q}^{-1}\mathbf{S} \bullet \mathcal{V}\|_F \, \|\mathcal{C}\|_F \\ &= \|\mathbf{Q}\|_F \, \|\mathbf{Q}^{-1}\mathbf{S}\|_F \, \|\mathcal{C}\|_F \\ &= \frac{(1+x^2)\sqrt{2(s_1^2+s_2^2)}}{|1-x^2|} \|\mathcal{C}\|_F \, . \end{aligned} \tag{5}$$

Intensity and Sensitivity IV

It is obvious that when x approaches 1, the intensity α goes to infinity. For the general case, the proof can be derived similarly with a symmetric matrix $\mathbf{Q}$ of size $R_2 \times R_2$ which has ones on the diagonal and two non-zero off-diagonal elements x.

Intensity and Sensitivity V

TC instability

The first observation is that TC models estimated by any iterative algorithms can encounter large TC-intensity. In many cases, the TC-intensity increases quickly with the iterations. Without proper processing, the algorithm gets stuck in a false local minimum. The decomposition is more challenging, especially when the dimension of a core tensor is smaller than its ranks, e.g., $R_1 R_2 > I_1$, or when components of the core tensors are highly collinear.

Such a type of degeneracy in TC happens quite often and is quite similar to that in CPD.

Definition (Sensitivity)

Given a TC model $\mathcal{Y} = \llbracket \mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket$. Denote by $\delta\mathcal{A}, \delta\mathcal{B}, \delta\mathcal{C}$ random Gaussian distributed perturbations with element distributed independently with zero mean and variance σ^2.
Sensitivity of the model $\llbracket \mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket$ is defined as

$$ss(\theta) = \lim_{\sigma^2 \to 0} \frac{1}{\sigma^2} E\left\{ \|\mathcal{Y} - \llbracket \mathcal{A} + \delta\mathcal{A}, \mathcal{B} + \delta\mathcal{B}, \mathcal{C} + \delta\mathcal{C} \rrbracket\|_F^2 \right\}$$

Consider the error tensor

$$\begin{aligned} & \llbracket \mathcal{A} + \delta\mathcal{A}, \mathcal{B} + \delta\mathcal{B}, \mathcal{C} + \delta\mathcal{C} \rrbracket - \llbracket \mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket \\ = \; & \llbracket \delta\mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket + \llbracket \mathcal{A}, \delta\mathcal{B}, \mathcal{C} \rrbracket + \llbracket \mathcal{A}, \mathcal{B}, \delta\mathcal{C} \rrbracket + \llbracket \delta\mathcal{A}, \delta\mathcal{B}, \mathcal{C} \rrbracket \\ & + \llbracket \delta\mathcal{A}, \mathcal{B}, \delta\mathcal{C} \rrbracket + \llbracket \mathcal{A}, \delta\mathcal{B}, \delta\mathcal{C} \rrbracket + \llbracket \delta\mathcal{A}, \delta\mathcal{B}, \delta\mathcal{C} \rrbracket \end{aligned}$$

Sensitivity for TC II

Since these TC terms are uncorrelated and expectation of the terms consisting of two or three $\delta\mathbf{A}$, $\delta\mathbf{B}$ and $\delta\mathbf{C}$ are negligible

$$E\{\|\mathcal{Y} - [\![\mathcal{A} + \delta\mathcal{A}, \mathcal{B} + \delta\mathcal{B}, \mathcal{C} + \delta\mathcal{C}]\!]\|_F^2\}$$
$$= E\{\|[\![\delta\mathcal{A}, \mathcal{B}, \mathcal{C}]\!]\|_F^2\} + E\{\|[\![\mathcal{A}, \delta\mathcal{B}, \mathcal{C}]\!]\|_F^2\} + E\{\|[\![\mathcal{A}, \mathcal{B}, \delta\mathcal{C}]\!]\|_F^2\}.$$

Let $\mathcal{Z} = \mathcal{B} \bullet \mathcal{C}$ be a TT tensor of two cores $\mathcal{B}$ and $\mathcal{C}$. Then $\mathcal{Y} = [\![\mathcal{A}, \mathcal{Z}]\!]$.

Sensitivity for TC III

We expand the Frobenius norm $\|[\![\delta\mathcal{A}, \mathcal{B}, \mathcal{C}]\!]\|_F^2$

$$
\begin{aligned}
E\{\|[\![\delta\mathcal{A}, \mathcal{B}, \mathcal{C}]\!]\|_F^2\} &= E\{\|\delta\mathbf{A}_{(2)}\mathbf{Z}_{(1,4)}\|_F^2\} \\
&= E\{\mathrm{tr}((\delta\mathbf{A}_{(2)}^T\delta\mathbf{A}_{(2)})(\mathbf{Z}_{(1,4)}\mathbf{Z}_{(1,4)}^T)\} \\
&= \sigma^2 I_1 \,\mathrm{tr}(\mathbf{Z}_{(1,4)}\mathbf{Z}_{(1,4)}^T) \\
&= \sigma^2 I_1 \|\mathcal{B} \bullet \mathcal{C}\|_F^2 \\
E\{\|[\![\mathcal{A}, \delta\mathcal{B}, \mathcal{C}]\!]\|_F^2\} &= \sigma^2 I_2 \|\mathcal{C} \bullet \mathcal{A}\|_F^2 \\
E\{\|[\![\mathcal{A}, \mathcal{B}, \delta\mathcal{C}]\!]\|_F^2\} &= \sigma^2 I_3 \|\mathcal{A} \bullet \mathcal{B}\|_F^2
\end{aligned}
$$

Sensitivity

Sensitivity of a TC model $[\![\mathcal{A}, \mathcal{B}, \mathcal{C}]\!]$ is computed as

$$ss(\mathcal{A}, \mathcal{B}, \mathcal{C}) = I_1\|\mathcal{B} \bullet \mathcal{C}\|_F^2 + I_2\|\mathcal{C} \bullet \mathcal{A}\|_F^2 + I_3\|\mathcal{A} \bullet \mathcal{B}\|_F^2$$

Balanced norm for Minimal sensitivity I

Sensitivity

$$ss(\mathcal{A},\mathcal{B},\mathcal{C}) = l_1\|\mathcal{B}\bullet\mathcal{C}\|_F^2 + l_2\|\mathcal{C}\bullet\mathcal{A}\|_F^2 + l_3\|\mathcal{A}\bullet\mathcal{B}\|_F^2$$

Balanced norm for Minimal sensitivity

A TC model $[\![\mathcal{A},\mathcal{B},\mathcal{C}]\!]$ can be scaled to give a new equivalent model

$$[\![\mathcal{A},\mathcal{B},\mathcal{C}]\!] = [\![\alpha_1\mathcal{A},\alpha_2\mathcal{B},\alpha_3\mathcal{C}]\!]$$

with minimal sensitivity, where

$$\alpha_n = \frac{\beta_n}{\beta}$$

$\beta_1 = \sqrt{l_1}\|\mathcal{B}\bullet\mathcal{C}\|_F, \beta_2 = \sqrt{l_2}\|\mathcal{C}\bullet\mathcal{A}\|_F, \beta_3 = \sqrt{l_3}\|\mathcal{A}\bullet\mathcal{C}\|_F$ and $\beta = \sqrt[3]{\beta_1\beta_2\beta_3}$.

Balanced norm for Minimal sensitivity II

$$\begin{aligned} ss &= \beta_1^2\alpha_2^2\alpha_3^2 + \beta_2^2\alpha_1^2\alpha_3^2 + \beta_3^2\alpha_1^2\alpha_2^2 \\ &\geq 3\sqrt[3]{\beta_1^2\beta_2^2\beta_3^2\alpha_1^4\alpha_2^4\alpha_3^4} \\ &= 3\sqrt[3]{\beta_1^2\beta_2^2\beta_3^2} \end{aligned}$$

Note that $\alpha_1\alpha_2\alpha_3 = 1$.
Equality holds when

$$\frac{\beta_1}{\alpha_1} = \frac{\beta_2}{\alpha_2} = \frac{\beta_3}{\alpha_3}.$$

Hence $\alpha_n = \frac{\beta_n}{\sqrt[3]{\beta_1\beta_2\beta_3}}$.

Rotation method for Sensitivity Correction I

Due to non uniqueness of the model up to rotation, we can rotate core tensors by invertible matrices such that the new representation of the TC tensor has minimum sensitivity

$$\mathcal{Y} = \{\!\!\{\mathbf{Q}_N^{-1} \bullet \mathcal{A}_1 \bullet \mathbf{Q}_1, \mathbf{Q}_1^{-1} \bullet \mathcal{A}_2 \bullet \mathbf{Q}_2, \mathbf{Q}_2^{-1} \bullet \mathcal{A}_3 \bullet \mathbf{Q}_3, \ldots, \mathbf{Q}_{N-1}^{-1} \mathcal{A}_N \mathbf{Q}_N \}\!\!\}.$$

For simplicity, we derive the algorithm to find the optimal matrix, $\mathbf{Q}$ of size $R_2 \times R_2$ which rotates the first two core tensors, $\mathcal{A}_1$ and $\mathcal{A}_2$, and gives a new equivalent TC tensor
$\mathcal{Y}_{\mathbf{Q}} = \{\!\!\{\mathcal{A}_1 \bullet \mathbf{Q}, \mathbf{Q}^{-1} \bullet \mathcal{A}_2, \mathcal{A}_3, \ldots, \mathcal{A}_N \}\!\!\}$.
The optimal matrix $\mathbf{Q}$ minimizes the sensitivity of $\mathcal{Y}_{\mathbf{Q}}$

$$\min_{\mathbf{Q}} \quad ss(\mathcal{Y}_{\mathbf{Q}}) = l_1 \|\mathbf{Q}^{-1} \mathcal{A}_{-1}\|_F^2 + l_2 \|\mathcal{A}_{-2} \mathbf{Q}\|_F^2 + \sum_{n=3}^{N} l_n \|\mathcal{A}_{-n}\|_F^2 \qquad (7)$$

where $\mathcal{A}_{-n} = \mathcal{A}_{n+1} \bullet \cdots \bullet \mathcal{A}_N \bullet \mathcal{A}_1 \bullet \cdots \bullet \mathcal{A}_{n-1}$.

Rotation method for Sensitivity Correction II

We next define two matrices, $\mathbf{X}_1$ of size $R_3 \times R_3$ and $\mathbf{X}_2$ of size $R_1 \times R_1$, as self contraction of the tensor $\mathcal{A}_{-(1,2)} = \mathcal{A}_3 \bullet \cdots \bullet \mathcal{A}_N$ along all modes but mode-1 and mode-N, respectively

$$\mathbf{X}_1 = [\mathcal{A}_{-(1,2)}]_{(1)} [\mathcal{A}_{-(1,2)}]_{(1)}^T, \tag{8}$$

$$\mathbf{X}_2 = [\mathcal{A}_{-(1,2)}]_{(N)} [\mathcal{A}_{-(1,2)}]_{(N)}^T \tag{9}$$

and two square matrices, $\mathbf{T}_1$ and $\mathbf{T}_2$, of size $R_2 \times R_2$

$$\mathbf{T}_1 = \sum_{i_2=1}^{I_2} \mathcal{A}_2(:, i_2, :) \mathbf{X}_1 \mathcal{A}_2(:, i_2, :)^T, \tag{10}$$

$$\mathbf{T}_2 = \sum_{i_1=1}^{I_1} \mathcal{A}_1(:, i_1, :)^T \mathbf{X}_2 \mathcal{A}_1(:, i_1, :). \tag{11}$$

Rotation method for Sensitivity Correction III

We represent the matrix $\mathbf{Q}\mathbf{Q}^T = \mathbf{U}\mathbf{S}\mathbf{U}^T$ in form of its eigenvalue decomposition (EVD), where $\mathbf{U}$ is an orthogonal matrix of size $R_2 \times R_2$ and $\mathbf{S} = \text{diag}(s_1, \ldots, s_{R_2})$. The sensitivity in (7) is then computed as

$$\begin{aligned} ss(\mathcal{Y}_{\mathbf{Q}}) &= l_1 \|\mathbf{Q}^{-1}\mathcal{A}_{-1}\|_F^2 + l_2 \|\mathcal{A}_{-2}\mathbf{Q}\|_F^2 + \sum_{n=3}^{N} l_n \|\mathcal{A}_{-n}\|_F^2 \\ &= \sum_{n=3}^{N} l_n \|\mathcal{A}_{-n}\|_F^2 + l_1 \,\text{tr}(\mathbf{T}_1 \mathbf{Q}^{-1}\mathbf{Q}^{-1T}) + l_2 \,\text{tr}(\mathbf{T}_2 \mathbf{Q}\mathbf{Q}^T) \\ &= \sum_{n=3}^{N} l_n \|\mathcal{A}_{-n}\|_F^2 + l_1 \,\text{tr}((\mathbf{U}^T\mathbf{T}_1\mathbf{U})\mathbf{S}^{-1}) + l_2 \,\text{tr}((\mathbf{U}^T\mathbf{T}_2\mathbf{U})\mathbf{S}). \end{aligned}$$

Rotation method for Sensitivity Correction IV

Instead of seeking $\mathbf{Q}$, we find an orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{S}$

$$ss(\mathcal{Y}_{\mathbf{Q}}) = \sum_{n=3}^{N} I_n \|\mathcal{A}_{-n}\|_F^2 + \sum_{r=1}^{R_2} I_1 (\boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r) \frac{1}{s_r} + I_2 (\boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r) s_r$$

$$\geq \sum_{n=3}^{N} I_n \|\mathcal{A}_{-n}\|_F^2 + \sum_{r=1}^{R_2} 2\sqrt{I_1 I_2 (\boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r)(\boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r)}\,. \quad (12)$$

The equality holds when $s_r^\star = \sqrt{\dfrac{I_1 \boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r}{I_2 \boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r}}$, for $r = 1, \ldots, R_2$.

Given the optimal $s_r^\star$, we find the orthogonal matrix $\mathbf{U}$ in the following optimization problem

$$\min_{\mathbf{U} \in St_{R_2}} \quad \sum_{r=1}^{R_2} \sqrt{(\boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r)(\boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r)}\,, \quad (13)$$

which can be solved using the conjugate gradient algorithm on the Stiefel manifold Wen and Yin (2012).

Initialization. Applying the Cauchy-Schwarz inequality, the objective function in (**??**) is bounded above by $\frac{1}{2}\sum_{r=1}^{R_2}(\boldsymbol{u}_r^T\mathbf{T}_1\boldsymbol{u}_r) + (\boldsymbol{u}_r^T\mathbf{T}_2\boldsymbol{u}_r) = \frac{1}{2}\,\mathrm{tr}(\mathbf{U}^T(\mathbf{T}_1 + \mathbf{T}_2)\mathbf{U})$.
We can initialize $\mathbf{U}$ by eigenvectors of $(\mathbf{T}_1 + \mathbf{T}_2)$.
The rotation method is then applied to the next pair $\mathcal{A}_2$ and $\mathcal{A}_3$, $\mathcal{A}_3$ and $\mathcal{A}_4$, ..., $\mathcal{A}_N$ and $\mathcal{A}_1$, ... until the update reaches a stopping criterion. Pseudo-codes of the proposed algorithm for order-3 TC is listed in Algorithm 1.

Rotation method for Sensitivity Correction VI

Algorithm 1: `Rotation Method for SSC`

Input: $\mathcal{Y} = \llbracket \mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket$: $(I_1 \times I_2 \times I_3)$, and bond R
Output: $\hat{\mathcal{Y}} = \mathcal{Y}$ such that min $ss(\hat{\mathcal{Y}})$
begin
  repeat
    for $n = 1, 2, 3$ **do**
      $\mathbf{T}_1 = \sum_{i_2=1}^{I_2} \mathbf{B}_{i_2} (\mathbf{C}_{(1)} \mathbf{C}_{(1)}^T) \mathbf{B}_{i_2}^T$,
      $\mathbf{T}_2 = \sum_{i_1=1}^{I_1} \mathbf{A}_{i_1}^T (\mathbf{C}_{(3)} \mathbf{C}_{(3)}^T) \mathbf{A}_{i_1}$
      Solve $\mathbf{U}^* = \arg\min_{\mathbf{U} \in St_{R_2}} \sum_{r=1}^{R_2} \sqrt{(\boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r)(\boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r)}$
      for $r = 1 \ldots, R_2$ **do** $s_r^\star = \sqrt{\dfrac{I_1 \boldsymbol{u}_r^T \mathbf{T}_1 \boldsymbol{u}_r}{I_2 \boldsymbol{u}_r^T \mathbf{T}_2 \boldsymbol{u}_r}}$
      Rotate $\mathcal{A} \leftarrow \mathcal{A} \bullet \mathbf{U}\, \mathrm{diag}(\sqrt{s_1}, \ldots, \sqrt{s_{R_2}}, \ldots) \mathbf{U}^T$
      Rotate $\mathcal{B} \leftarrow \mathbf{U}\, \mathrm{diag}(1/\sqrt{s_1}, \ldots, 1/\sqrt{s_{R_2}}, \ldots) \mathbf{U}^T \bullet \mathcal{B}$
      Cyclic-shift $\hat{\mathcal{Y}} = \llbracket \mathcal{A}, \mathcal{B}, \mathcal{C} \rrbracket \leftarrow \llbracket \mathcal{B}, \mathcal{C}, \mathcal{A} \rrbracket$
    end
  until *a stopping criterion is met*
end

How to deal with Instability in TC I

- Instability in TC is like degeneracy in Canonical polyadic tensor decomposition, which is hard to avoid
- Instead we propose to correct the unstable estimated model Phan et al. (2019); Phan et al. (2020b)

Error Preserving Correction Method

- Seeking a new tensor, $\hat{\mathcal{Y}}$, which preserves the approximation error but has smaller sensitivity.

$$\begin{aligned} \min \quad & ss(\theta) = l_1\|\mathcal{B} \bullet \mathcal{C}\|_F^2 + l_2\|\mathcal{C} \bullet \mathcal{A}\|_F^2 + l_3\|\mathcal{A} \bullet \mathcal{B}\|_F^2 \\ \text{s.t.} \quad & c(\theta) = \|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2 \leq \delta^2, \end{aligned}$$

where θ is vector of parameters and $\hat{\mathcal{Y}} = [\![\mathcal{A}, \mathcal{B}, \mathcal{C}]\!]$

Phan, Tichavsky and Cichocki, Error Preserving Correction: A Method for CP Decomposition at a Target Error Bound, IEEE Transaction on Signal Processing, 2019.

How to deal with Instability in TC II

- Objective and constraint functions are nonlinear in all the factor matrices
- Rewrite the objective function and the constraint function for a single core tensor.
- Then solve the problem using the alternating update scheme

Alternating Sensitivity Correction Method I

$$\min \quad ss(\theta) \quad \text{s.t.} \quad c(\theta) = \|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2 \leq \delta^2,$$

- Rewrite the objective function

$$ss(\theta) = l_1 \|\mathcal{B} \bullet \mathcal{C}\|_F^2 + \text{tr}(\mathbf{Q}\mathbf{A}_{(2)}^T \mathbf{A}_{(2)})$$

where $\mathbf{Q} = l_2(\mathbf{I}_{R_2} \otimes \mathbf{C}_{(3)}\mathbf{C}_{(3)}^T) + l_3(\mathbf{B}_{(1)}\mathbf{B}_{(1)}^T \otimes \mathbf{I}_{R_1})$

- In order to update the core tensor $\mathcal{A}$, we solve the minimization problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathbf{Q}\mathbf{X}^T\mathbf{X}) \\ \text{s.t.} \quad & \|\mathbf{Y}_{(1)} - \mathbf{X}\mathbf{Z}^T\|_F^2 \leq \delta^2 \end{aligned}$$

where $\mathbf{Z}$ is mode-(1,4) unfolding of $\mathcal{B} \bullet \mathcal{C}$,
$\mathbf{X}$ is mode-2 unfolding of the core tensor $\mathcal{A}$, i.e., $\mathbf{A}_{(2)} = \mathbf{X}$.

Alternating Sensitivity Correction Method II

- $\mathcal{A}$ can be found in closed form as Spherical Constrained Quadratic Programming (SCQP) Gander et al. (1989); Phan et al. (2019).

Algorithm 2: Sensitivity Correction

Input: Data tensor $\mathcal{Y}$: $(I_1 \times I_2 \times I_3)$, and ranks R and error bound δ

Output: $\hat{\mathcal{Y}} = [\![\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3]\!]$: $\min ss(\hat{\mathcal{Y}})$ s.t. $\|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2 \leq \delta^2$

begin

1 Initialize $\hat{\mathcal{Y}}$

repeat

for $n = 1, 2, 3$ **do**

$\mathcal{Z} = \mathcal{G}_2 \bullet \mathcal{G}_3$

2 $\mathbf{Q} = I_2(\mathbf{I}_{R_2} \otimes \mathbf{G}_{3,(3)}\mathbf{G}_{3,(3)}^T) + I_3(\mathbf{G}_{2,(1)}\mathbf{G}_{2,(1)}^T \otimes \mathbf{I}_{R_1})$

3 Solve $\mathcal{G}_1 = \arg\min_{\mathbf{X}} \ \mathrm{tr}(\mathbf{X}\mathbf{Q}\mathbf{X}^T)$ s.t. $\|\mathbf{Y}_{(1)} - \mathbf{X}\mathbf{Z}_{(1,4)}\|_F^2 \leq \delta^2$

4 Cyclic-shift $\mathcal{Y}$ and $\hat{\mathcal{Y}}$

end

until *a stopping criterion is met*

5 Balance norm of core tensors

end

Algorithm 3: Sensitivity Correction Algorithm

Input: Data tensor $\mathcal{Y}$: $(I_1 \times I_2 \times I_3)$, and ranks R
Output: $\hat{\mathcal{Y}} = [\![\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3]\!]$
begin
  repeat
    Run ALS to update the TC tensor $\hat{\mathcal{Y}}$: min $\|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2$
    if $ss(\theta) \leq ss_{max}$ **then**
      Perform sensitivity correction with $\delta = \|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2$
    end
  until *a stopping criterion is met*
end

TC Decomposition with Sensitivity control I

- Different from the Sensitivity Correction method, we propose a TC decomposition with bounded sensitivity

$$\min \quad \|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2 \quad \text{s.t.} \quad ss(\theta) \leq \gamma,$$

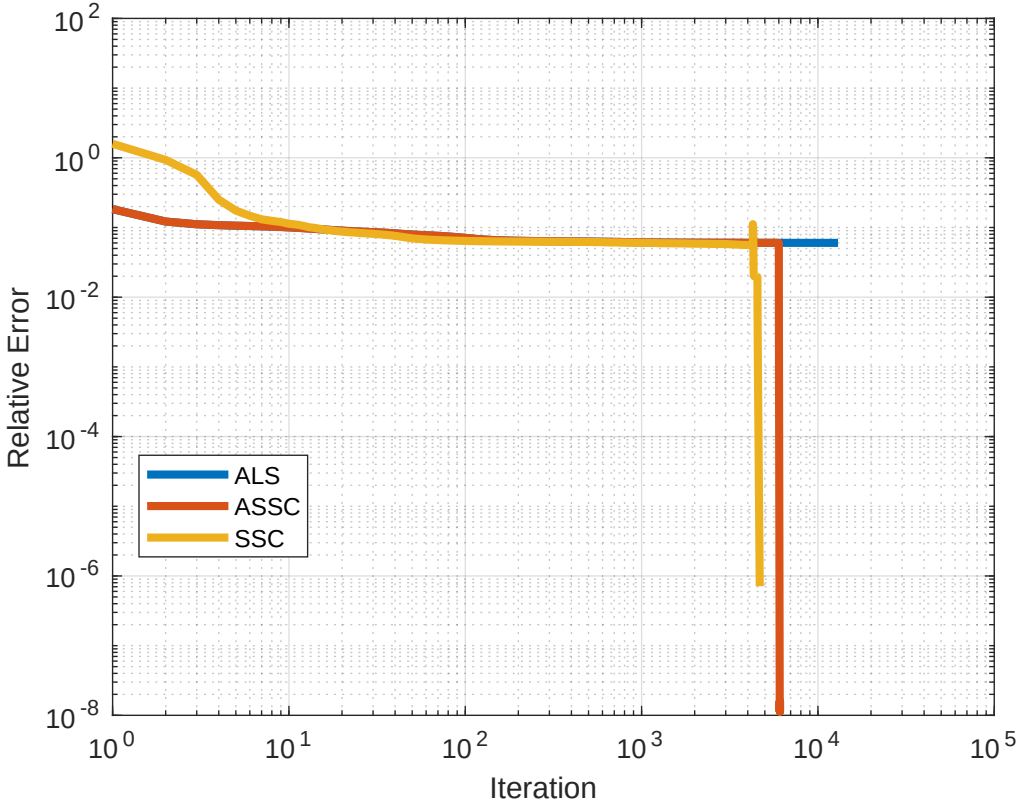
- In order to update the core tensor $\mathcal{A}$, we solve the minimization problem

$$\begin{aligned} \min_{\mathbf{X}} \quad & \|\mathbf{Y}_{(1)} - \mathbf{X}\mathbf{Z}^T\|_F^2 \\ \text{s.t.} \quad & \text{tr}(\mathbf{X}\mathbf{Q}\mathbf{X}^T) \leq \gamma_n \end{aligned}$$

where $\gamma_n = \gamma - l_1\|\mathcal{Z}\|_F^2$, $\mathbf{Z}$ is mode-(1,4) unfolding of $\mathcal{Z} = \mathcal{B} \bullet \mathcal{C}$,
$\mathbf{X}$ is mode-2 unfolding of the core tensor $\mathcal{A}$, i.e., $\mathbf{A}_{(2)} = \mathbf{X}$,
$\mathbf{Q} = l_2(\mathbf{I}_{R_2} \otimes \mathbf{C}_{(3)}\mathbf{C}_{(3)}^T) + l_3(\mathbf{B}_{(1)}\mathbf{B}_{(1)}^T \otimes \mathbf{I}_{R_1})$

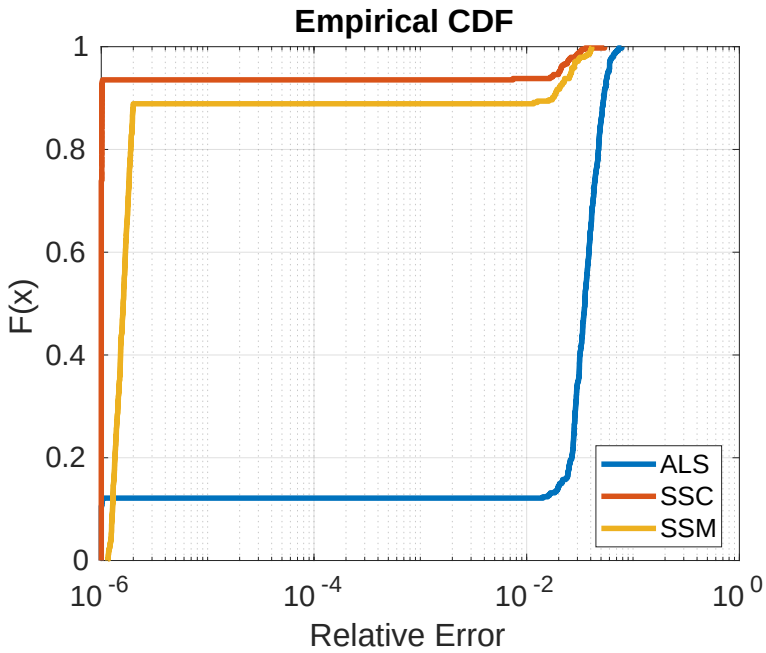
- Similar to the SSC algorithm, $\mathcal{A}$ can be found in closed form.

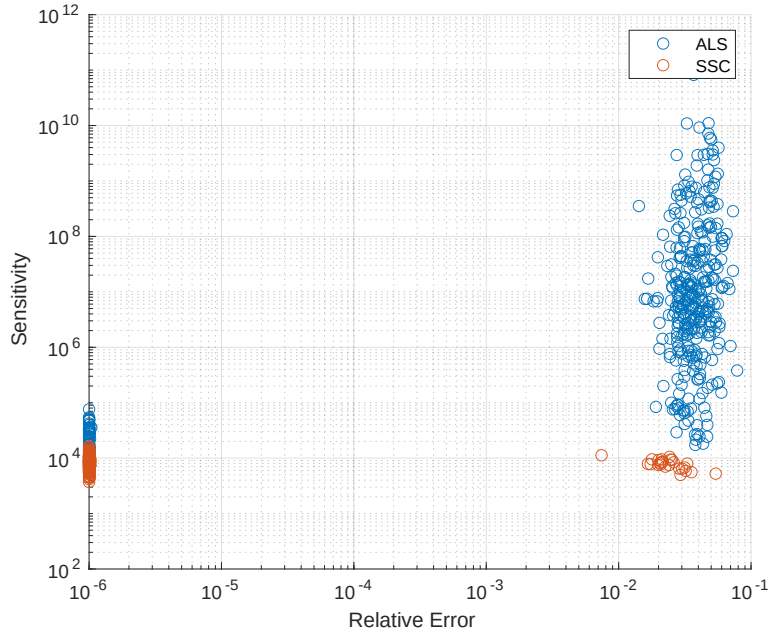
Second Example (cont') I



- Noise-free tensors of size $7 \times 7 \times 7$
- We applied Sensitivity correction after 3000 ALS updates, then continue the decomposition.
- ALS converged quickly after sensitivity correction.

Second Example (cont') II





- In 10000 iterations, ALS succeeded in fitting the tensor in less than 12% of runs.
- With Sensitivity correction (SSC), ALS attained a much higher success rate of 94%.
- Sensitivity minimization (SSM) works well as SSC, with a success rate of 89%.

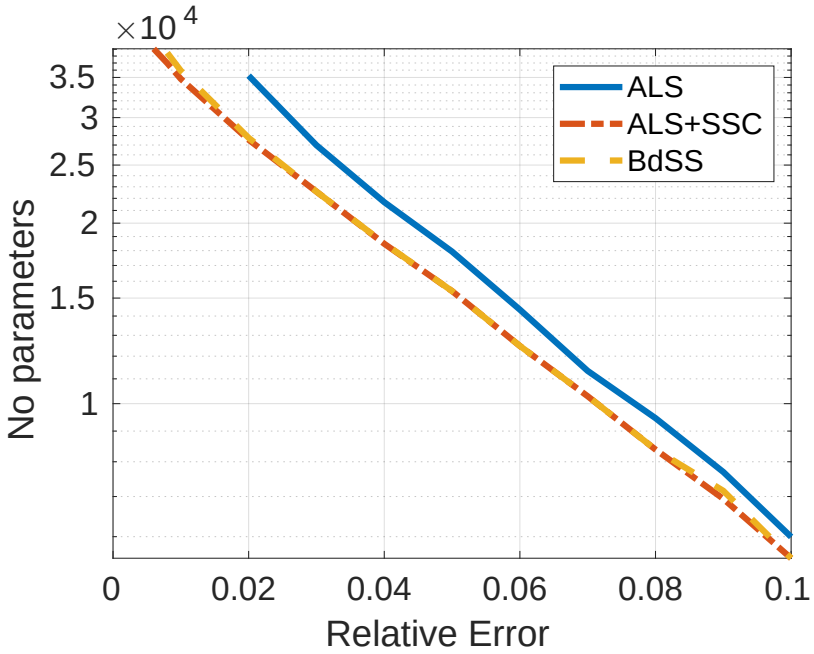
Example: Fitting Images by TC-decomposition I

- Decompose color images of size $128 \times 128 \times 3$ by TC model with ranks $R_1 = R_2$
- For the same approximation bound, we compare three models obtained using ALS, ALS with sensitivity correction, and SSM

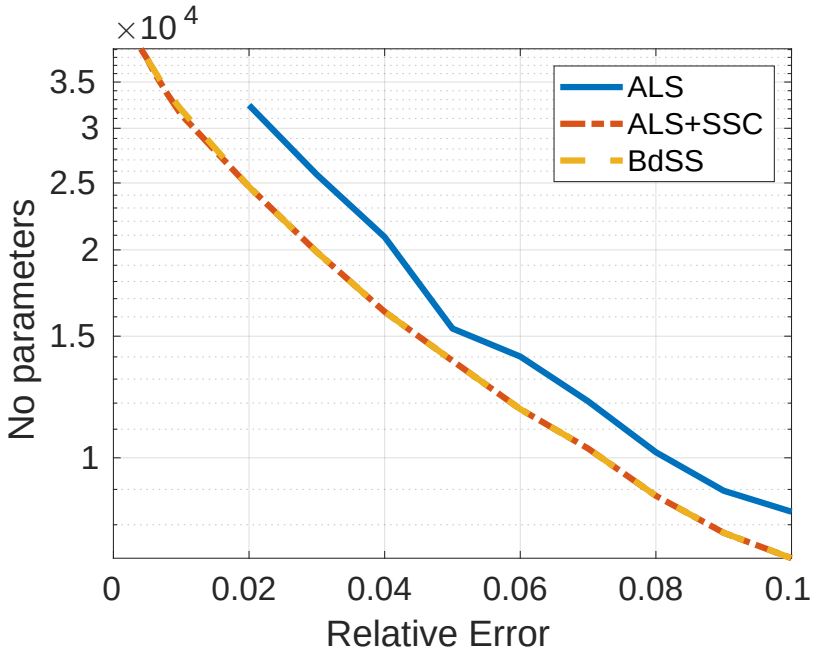
$$\|\mathcal{Y} - \mathcal{X}\|_F \leq \varepsilon \|\mathcal{Y}\|_F$$

Sedighin, Cichocki and Phan, Adaptive Rank Selection for Tensor Ring Decomposition, IEEE Journal of Selected Topics in Signal Processing, 2021

Example: Fitting Images by TC-decomposition II

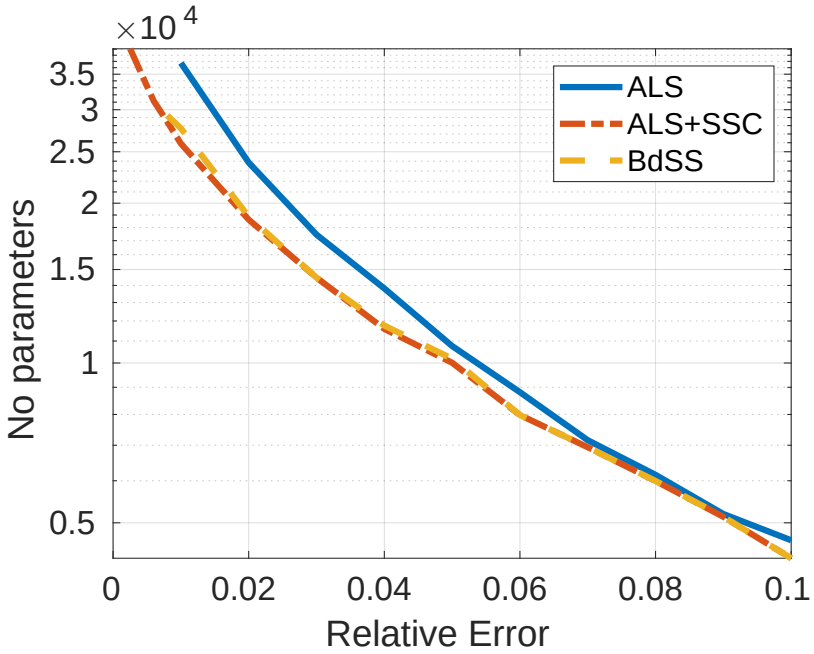


(a) Mandrill

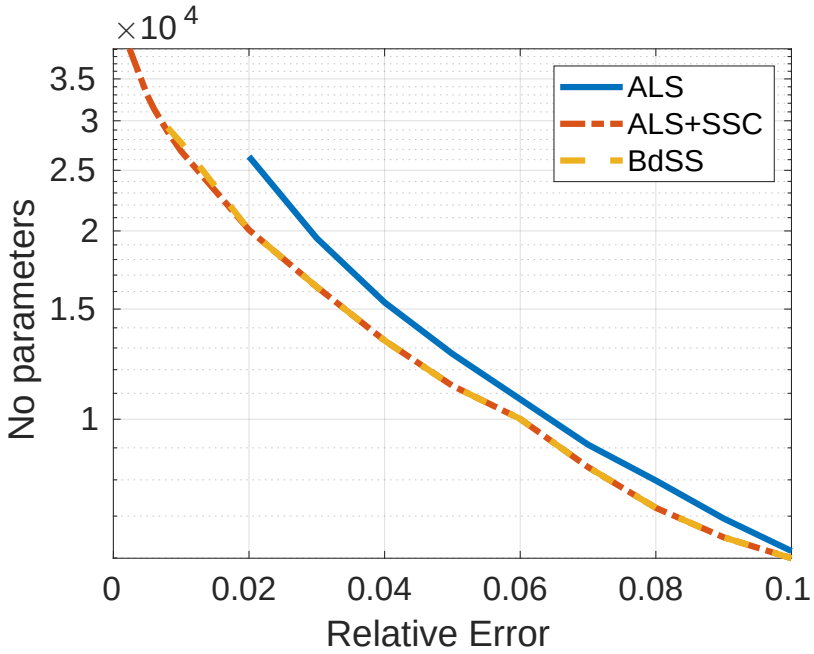


(b) Peppers

Example: Fitting Images by TC-decomposition III

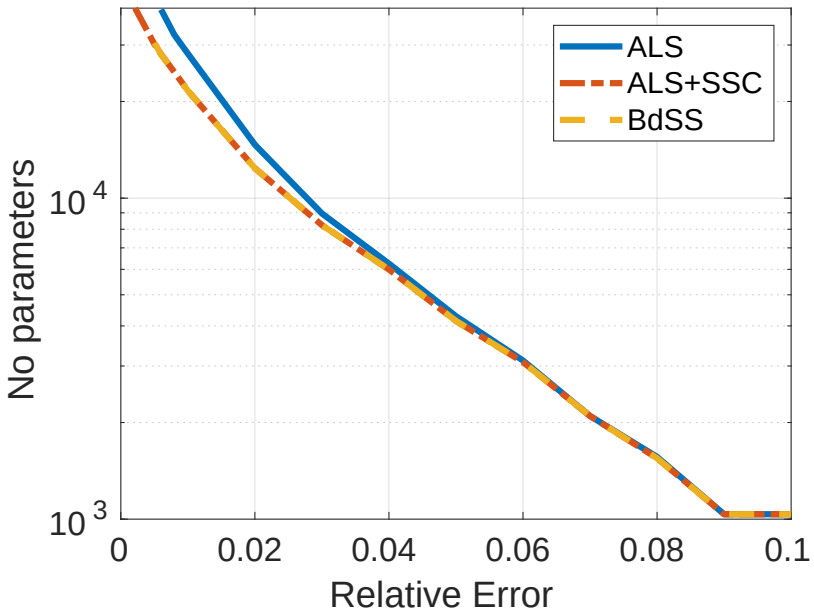


(c) Lena

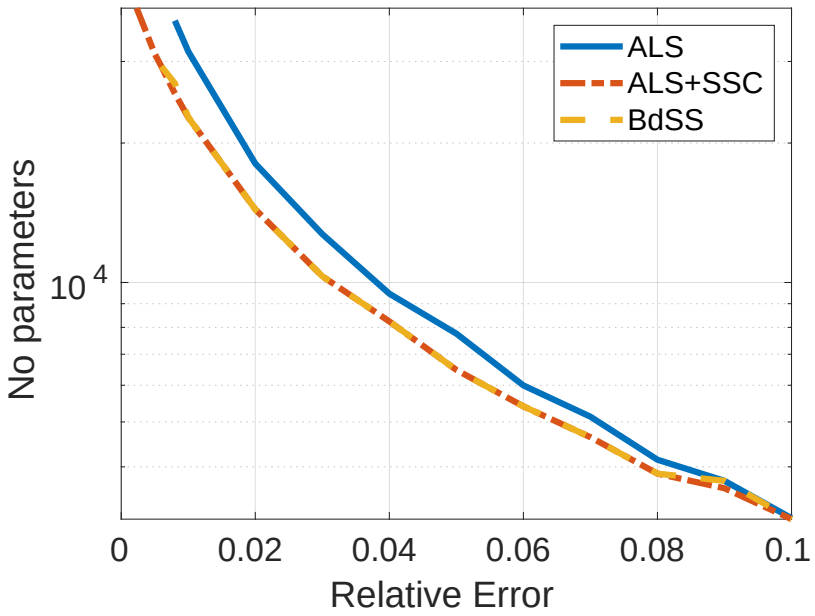


(d) Barbara

Example: Fitting Images by TC-decomposition IV

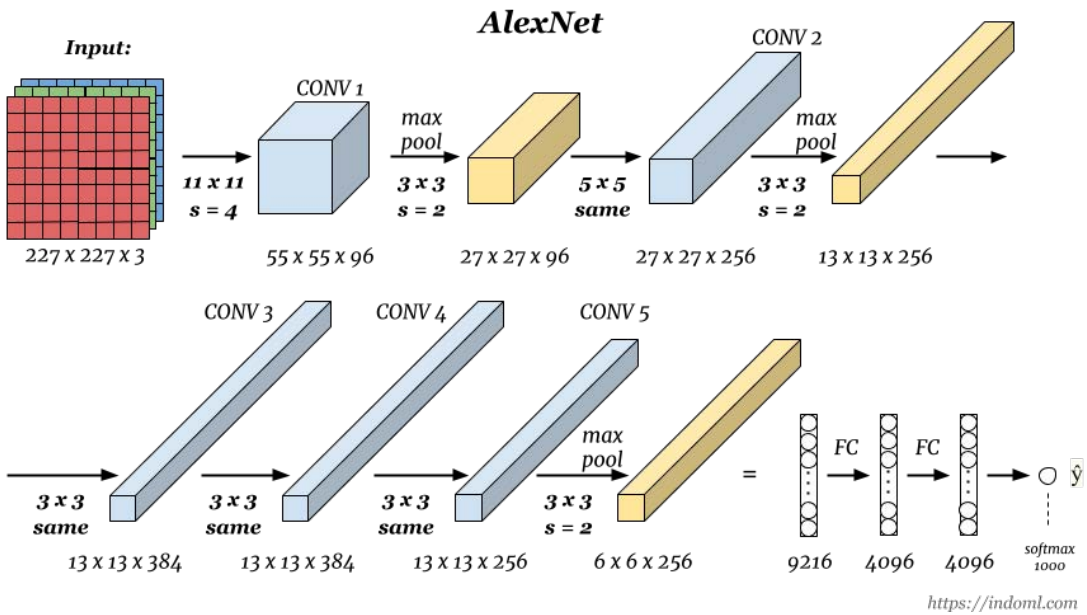


(e) Tiffany



(f) House

Example: Approximation of Convolutional Kernels I



- **AlexNet** one of the first deep convolutional neural networks beating traditional computer vision methodologies, composed of 5 convolutional layers followed by 3 fully connected layers
- Most convolutional neural networks are overparameterized and exhibit high computational cost.

Example: Approximation of Convolutional Kernels II

- Low-rank approximation reduces the number of parameters in convolutional layers.
  Thereby accelerate the inference of the network.
- Phan et al. (2020b) show that CPD with sensitivity control can significantly improve compression of CNNs including ResNet, VGG, over ordinary CPD methods due to severe degeneracy of the decomposition results.
- By keeping the decomposition at low sensitivity, the compressed CNNs retain their original accuracy (with a minor loss).

Example: Approximation of Convolutional Kernels III

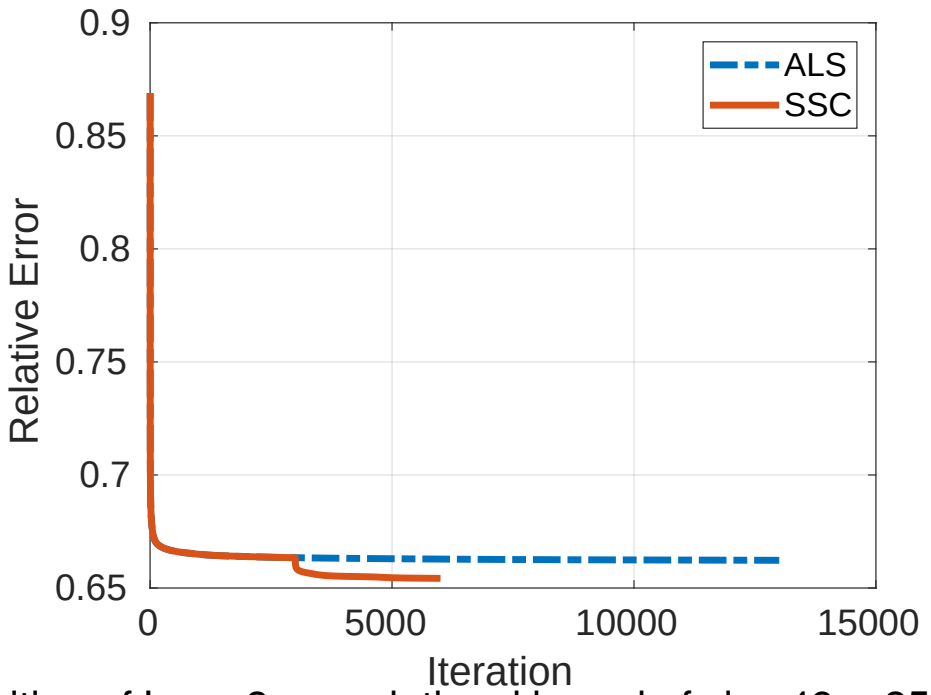


Figure: Decomposition of Layer2 convolutional kernel of size $48 \times 256 \times 25$, rank-(6,10,5).

- Sensitivity of the estimated TC model after 3000 iterations was 6.6e+06, and increased to 1.9e+07 after 13000 iterations.
- SSC was applied after 3000 ALS updates and yielded a model with sensitivity of 32146.

Example: Approximation of Convolutional Kernels IV

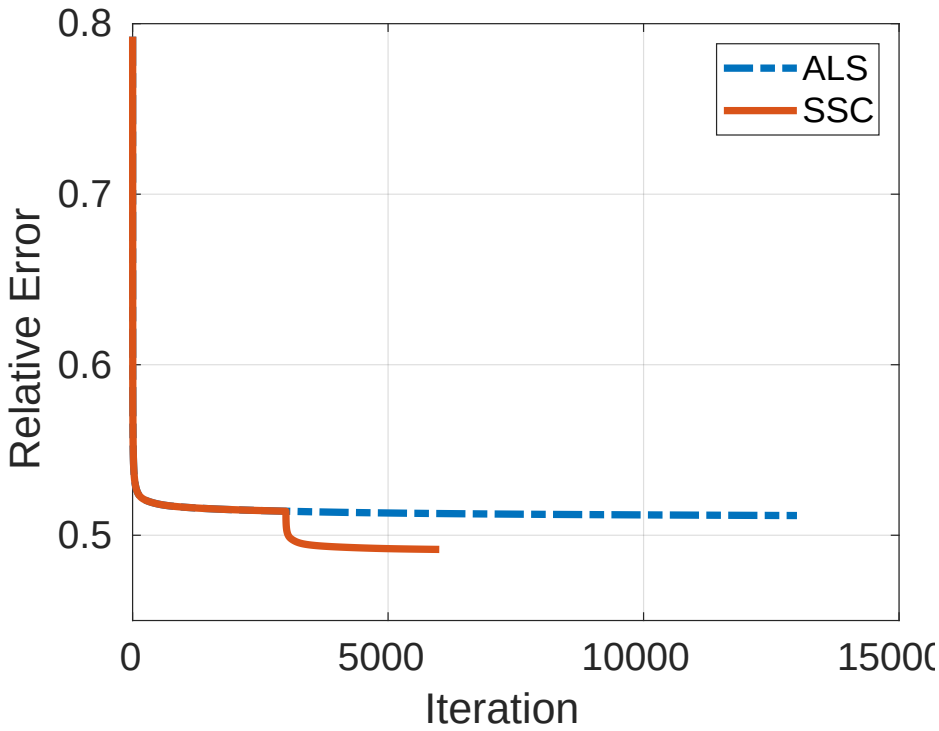


Figure: Decomposition of Layer2 convolutional kernel, rank-(10,10,10).

- Sensitivity of the estimated TC model after 3000 iterations was 6.2e+07, and increased to 1.9e+08 after 13000 iterations.
- SSC was applied after 3000 ALS updates and yielded a model with sensitivity of 46988.

Example: Approximation of Convolutional Kernels V

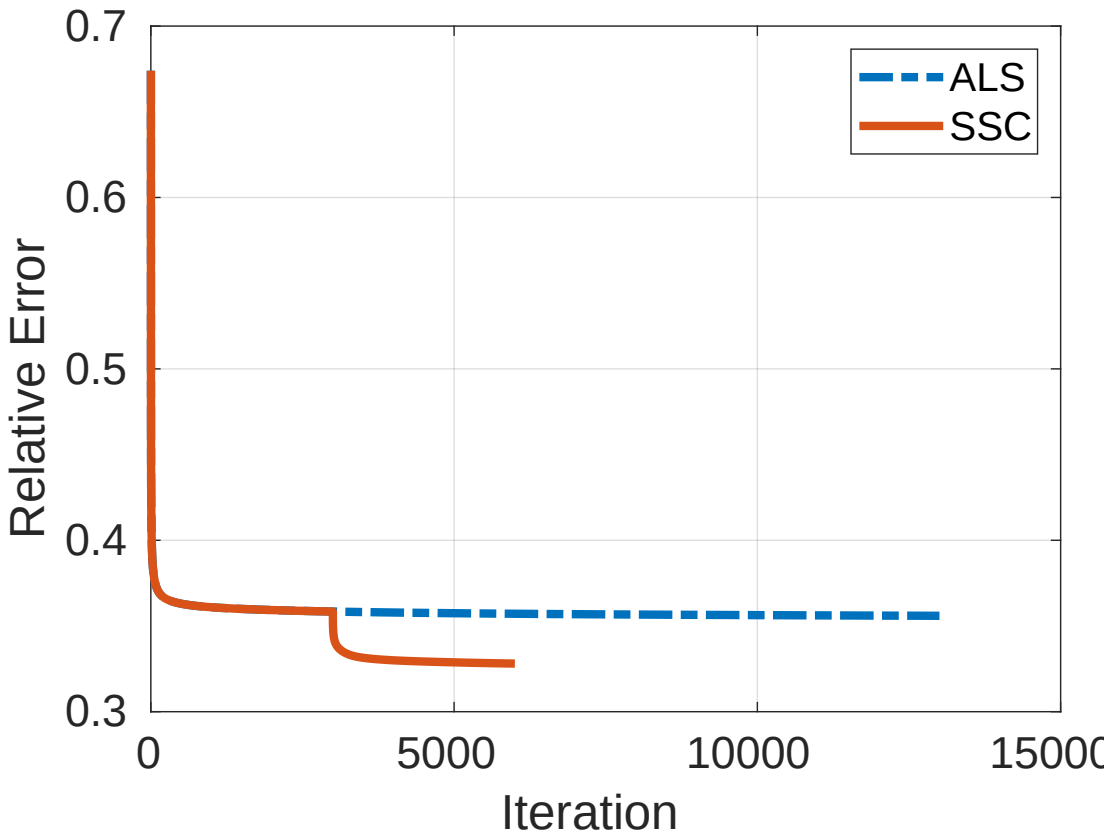


Figure: Decomposition of Layer2 convolutional kernel, rank-(9,18,12). SSC was applied after 3000 ALS updates.

Example: Approximation of Convolutional Kernels VI

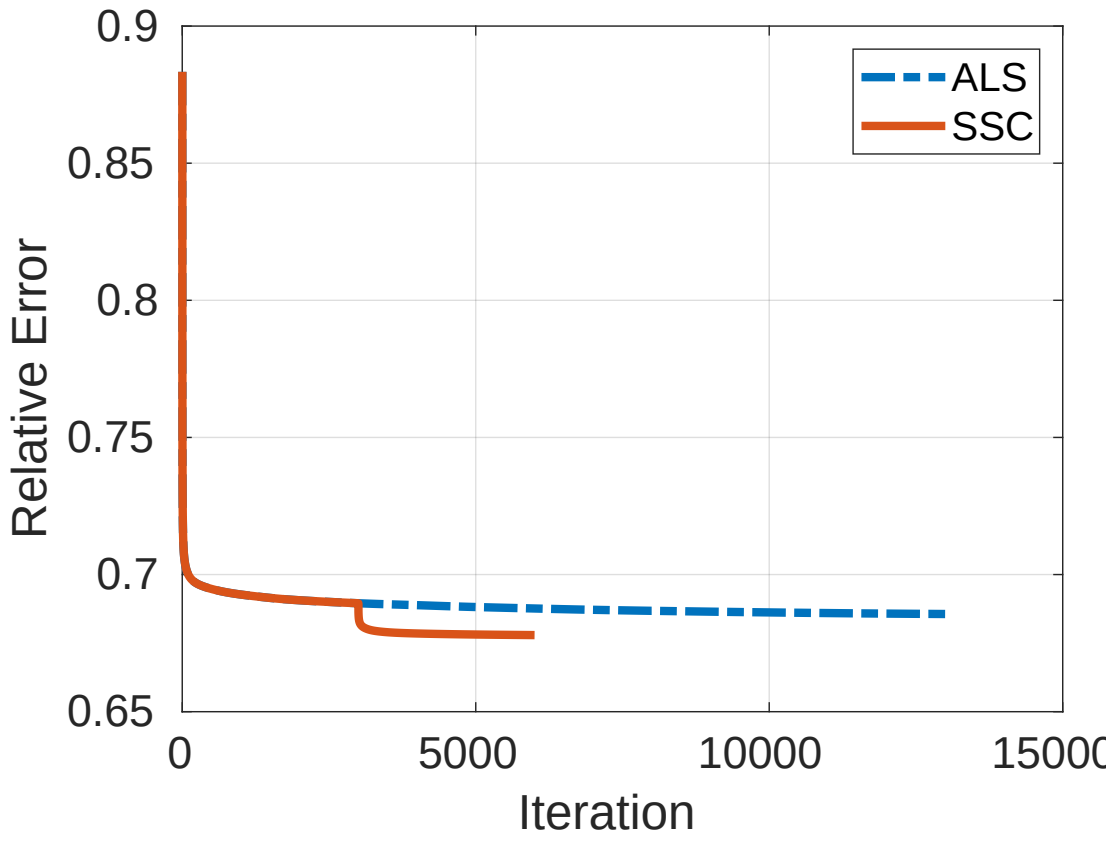


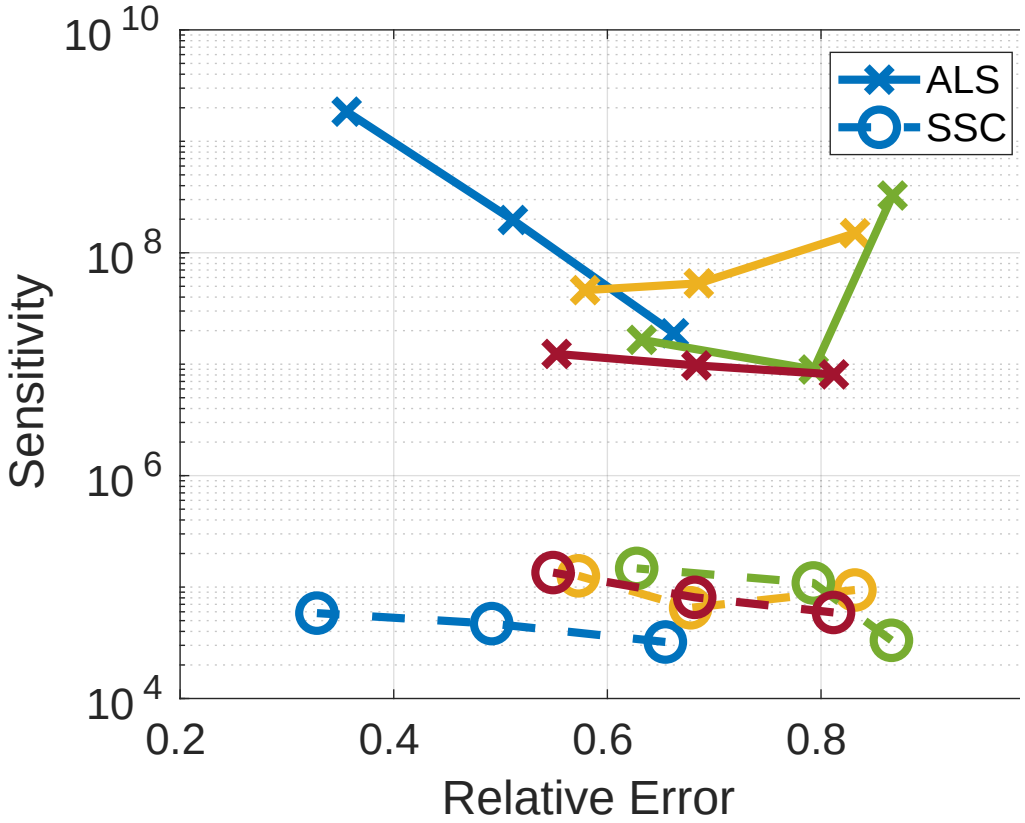
Figure: Decomposition of Layer3 convolutional kernel, rank-(10,10,10). SSC was applied after 3000 ALS updates.

Example: Approximation of Convolutional Kernels VII

Table: Relative approximation errors obtained using ALS and ALS+SSC

Layer	Method	Setting 1	Setting 2	Setting 3
2	ALS	0.6623	0.5116	0.3559
	SSC	**0.6543**	**0.4917**	**0.3281**
3	ALS	0.8319	0.6856	0.5792
	SSC	**0.8317**	**0.6779**	**0.5728**
4	ALS	0.8670	0.7932	0.6321
	SSC	**0.8659**	**0.7928**	**0.6274**
5	ALS	0.8121	0.6833	0.5525
	SSC	**0.8117**	**0.6817**	**0.5492**

Example: Approximation of Convolutional Kernels VIII



Without sensitivity correction or control, TC decomposition often converges to false local minima with very high sensitivity.

TC can encounter severe instability problem, which prevents the compressed CNNs from achieving its original accuracy

Conclusions

- Sensitivity of the estimated TC tensor can be very high, makes the algorithms get stuck in local minima
- With sensitivity correction, the decomposition can start from an equivalent model but with smaller sensitivity. This helps the decomposition get through the unstable easily.
- Sensitivity correction or decomposition with minimal sensitivity can find a TC model with smaller approximation error or lower sensitivity.

References I

Bro, R., Harshman, R. A., Sidiropoulos, N. D., and Lundy, M. E. (2009). Modeling multi-way data with linearly dependent loadings. *Journal of Chemometrics*, 23(7-8):324–340.

Carroll, J., Pruzansky, S., and Kruskal, J. (1970). Candelinc: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika*, 45(1):3–24.

Espig, M., Hackbusch, W., Handschuh, S., and Schneider, R. (2011). Optimization problems in contracted tensor networks. *Comput. Visual. Sci.*, 14(6):271–285.

Espig, M., Naraparaju, K. K., and Schneider, J. (2012). A note on tensor chain approximation. *Computing and Visualization in Science*, 15(6):331–344.

References II

Gander, W., Golub, G. H., and von Matt, U. (1989). A constrained eigenvalue problem. *Special Issue Dedicated to Alan J. Hoffman, Linear Algebra and its Applications*, 114:815 – 839.

Handschuh, S. (2015). *Numerical Methods in Tensor Networks*. PhD thesis, Facualty of Mathematics and Informatics, University Leipzig, Germany, Leipzig, Germany.

Holtz, S., Rohwedder, T., and Schneider, R. (2012). The alternating linear scheme for tensor optimization in the tensor train format. *SIAM J. Scientific Computing*, 34(2).

Kressner, D., Steinlechner, M., and Uschmajew, A. (2014). Low-rank tensor methods with subspace correction for symmetric eigenvalue problems. *SIAM Journal on Scientific Computing*, 36(5):A2346–A2368.

References III

Landsberg, J. M. (2012). *Tensors: Geometry and Applications*, volume 128. American Mathematical Society, Providence, RI, USA.

Oseledets, I. and Tyrtyshnikov, E. (2009). Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM Journal on Scientific Computing*, 31(5):3744–3759.

Phan, A.-H., Cichocki, A., Uschmajew, A., TichavskÃ½, P., Luta, G., and Mandic, D. P. (2020a). Tensor networks for latent variable analysis: Novel algorithms for tensor train approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4622–4636.

Phan, A.-H., Sobolev, K., Sozykin, K., Ermilov, D., Gusak, J., Tichavský, P., Glukhov, V., Oseledets, I., and Cichocki, A. (2020b). Stable low-rank tensor decomposition for compression of convolutional neural network. In *Computer Vision – ECCV 2020*, pages 522–539, Cham. Springer International Publishing.

References IV

Phan, A.-H., Tichavský, P., and Cichocki, A. (2019). Error preserving correction: A method for CP decomposition at a target error bound. *IEEE Transactions on Signal Processing*, 67(5):1175–1190.

Phan, A.-H., Yamagishi, M., and Cichocki, A. (2019). Quadratic programming over ellipsoids and its applications to linear regression and tensor decomposition. *Neural Computing and Applications*.

Sorber, L., Van Barel, M., and De Lathauwer, L. (2013). Tensorlab v1.0.

Vidal, G. (2003). Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902.

Wen, Z. and Yin, W. (2012). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, pages 1–38.