Anh-Huy Phan
(joint work with Petr Tichavský)

Center for Computational and Data-Intensive Science and Engineering (CDISE)
Skolkowo Institute of Science and Technology (SKOLTECH), Moscow, Russia

May 4, 2021
• Looped Tensor network
• Relation between TC and Tucker, CP decompositions
• Sensitivity
• Algorithms
Model a tensor $Y$ of size $I_1 \times I_2 \times \cdots \times I_N$ by $N$ interconnected core tensors $G_n$ of size $R_n \times I_n \times R_{n+1}$

$$Y = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_{N-1}=1}^{R_{N-1}} G_1(:, r_1) \circ G_2(r_1, :, r_2) \circ \cdots \circ G_N(r_{N-1}, :),$$

or

$$Y = G_1 \cdot G_2 \cdot \cdots \cdot G_{N-1} \cdot G_N$$

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

Anh-Huy Phan  
Tensor Chain Decomposition
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.
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.
Similar to TT, core tensors $G_n$ of size $R_{n-1} \times I_n \times R_n$, $R_0 = R_N$, the TC model is written as

$$y = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_N=1}^{R_N} G_1(r_N, :, r_1) \circ G_2(r_1, :, r_2) \circ \cdots \circ G_N(r_{N-1}, :, r_N),$$

$$y_{i_1i_2\ldots i_N} = \text{tr}(G_1(:, i_1, :)G_2(:, i_2, :) \cdots G_3(:, i_N, :))$$

Shorthand notation for TC

$$y = \langle G_1, G_2, \ldots, G_N \rangle$$

When looped network opens

When one of the core tensors, e.g., $G_N$, in the TC model $y$ is a diagonal tensor of 1s, the TC becomes TT model

$$y = G_1 \circ G_2 \circ \cdots \circ G_{N-1}$$
Consider an order-3 TC decomposition
\[ \min \| Y \approx A, B, C \|_F^2 \]

In order to update the core tensor \( A \), we minimize the Frobenius norm Espig et al. (2011, 2012)
\[ \min \| Y^{(1)} - A^{(2)} Z \|_F^2 \]

where \( Y^{(1)} \) is mode-1 unfolding of \( Y \), \( A^{(2)} \) is mode-2 unfolding of \( A \), \( Z \) is mode-(1,3) unfolding of \( B \cdot C \),

The update rule reads
\[ A^{(2)} = Y^{(1)} Z^\dagger \]
• 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))
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%.
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%. 

Anh-Huy Phan
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.
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))
Matrix-multiplication tensor $\mathcal{M}$ obeys the relation

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

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

(a) TC-3

(b) Tucker-3
Applying this relation to order-3 TC tensor, \( Y = \mathcal{F}A, B, C \mathcal{F} \) gives

\[
y_{ijk} = \text{tr}(A_iB_jC_k) = \text{vec}(C_k)^T \text{vec}(A_iB_j)
\]

\[
= \text{vec}(C_k)^T (M \times_1 \text{vec}(A_i)^T \times_2 \text{vec}(B_j)^T)
\]

\[
= M \times_1 \text{vec}(A_i)^T \times_2 \text{vec}(B_j)^T \times_3 \text{vec}(C_k)^T.
\]

It follows that

\[
Y = M \times_1 A^{(2)} \times_1 B^{(2)} \times_3 C^{(2)}
\]

\[
= [M; A^{(2)}, B^{(2)}, C^{(2)}]
\]

where \( A^{(2)}, B^{(2)}, C^{(2)} \) are mode-2 unfoldings of \( A, B \) and \( C \), respectively.
Order-3 TC is a structured Tucker decomposition with fixed known core tensor.

Similar relation can be established for higher order tensors.
CANonical DEcomposition with LINear Constraints Carroll et al. (1970) PARallel FACtor with LINear DepedencesBro et al. (2009)

\[ \mathbf{y} = [\mathbf{AU}, \mathbf{BV}, \mathbf{CW}] \]

where \( \mathbf{U}, \mathbf{V}, \mathbf{W} \) are known dependence matrices.

Note that multiplication tensor, \( \mathbf{M} \), can be represented by a Canonical tensor model with binary factor matrices

For example, \( \mathbf{M} \) of size \( 4 \times 4 \times 4 \) has rank-7

\[
\mathbf{M}^{(1)} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]
$\mathcal{M} = [U, V, W]$ where

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

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

$$W = \begin{bmatrix}
1 & 0 & 0 & 1 & -1 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & -1 & 1 & 0 & 0 & 1 & 0
\end{bmatrix}.$$
### TC is PARALIND

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

\[
Y = \mathcal{F} \mathcal{A}, \mathcal{B}, \mathcal{C} \mathcal{F}
\]

\[
= [\mathcal{M}; A_{(2)}, B_{(2)}, C_{(2)}]
\]

\[
= [A_{(2)} U, B_{(2)} V, C_{(2)} W]
\]

(TC model)

(Tucker model)

(\( PARALIND \))

---

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

Given a TC model $\mathbf{Y} = \mathbb{F} \mathbf{A}, \mathbf{B}, \mathbf{C} \mathbb{F}$. Denote by $\delta \mathbf{A}, \delta \mathbf{B}, \delta \mathbf{C}$ random Gaussian distributed perturbations with element distributed independently with zero mean and variance $\sigma^2$.

Sensitivity of the model $\mathbb{F} \mathbf{A}, \mathbf{B}, \mathbf{C} \mathbb{F}$ is defined as

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

Sensitivity for TC can be derived from that for structured Tucker decomposition


Sensitivity of Kruskal model (for CPD) was proposed in Tichavský et al. (2019)

Tichavsky and Phan and Cichocki, Sensitivity, IEEE Signal processing letters, 2019
Consider the error tensor

\[
\mathcal{A} + \delta\mathcal{A}, \mathcal{B} + \delta\mathcal{B}, \mathcal{C} + \delta\mathcal{C} \rightleftharpoons \mathcal{A}, \mathcal{B}, \mathcal{C}
\]

\[
= \mathcal{A}, \mathcal{B}, \mathcal{C} \rightleftharpoons \mathcal{A}, \delta\mathcal{B}, \mathcal{C} \rightleftharpoons \mathcal{A}, \mathcal{B}, \delta\mathcal{C} \rightleftharpoons \delta\mathcal{A}, \delta\mathcal{B}, \mathcal{C}
\]

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

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

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

Let \(\mathcal{Z} = \mathcal{B} \bullet \mathcal{C}\) be a TT tensor of two cores \(\mathcal{B}\) and \(\mathcal{C}\). Then \(\mathcal{Y} = \mathcal{A}, \mathcal{Z}\).
We expand the Frobenius norm $\|\Phi \delta A, B, C \Phi\|_F^2$

\[
E\{\|\Phi \delta A, B, C \Phi\|_F^2\} = E\{\|\delta A(2)Z(1,4)\|_F^2\} \\
= E\{\text{tr}((\delta A_T(2)\delta A(2))(Z(1,4)Z_T^{(1,4)}))\} \\
= \sigma^2 l_1 \text{tr}(Z(1,4)Z_T^{(1,4)}) \\
= \sigma^2 l_1 \|B \cdot C\|_F^2 \\
E\{\|\Phi A, \delta B, C \Phi\|_F^2\} = \sigma^2 l_2 \|C \cdot A\|_F^2 \\
E\{\|\Phi A, B, \delta C \Phi\|_F^2\} = \sigma^2 l_3 \|A \cdot B\|_F^2
\]

Sensitivity

Sensitivity of a TC model $\Phi A, B, C \Phi$ is computed as

\[
ss(A, B, C) = l_1\|B \cdot C\|_F^2 + l_2\|C \cdot A\|_F^2 + l_3\|A \cdot B\|_F^2
\]
Sensitivity

\[ ss(A, B, C) = I_1 \| B \bullet C \|_F^2 + I_2 \| C \bullet A \|_F^2 + I_3 \| A \bullet B \|_F^2 \]

Balanced norm for Minimal sensitivity

A TC model \( \langle A, B, C \rangle \) can be scaled to give a new equivalent model

\[ \langle A, B, C \rangle \overset{\beta}{\longrightarrow} \langle \alpha_1 A, \alpha_2 B, \alpha_3 C \rangle \]

with minimal sensitivity, where

\[ \alpha_n = \frac{\beta_n}{\beta} \]

\[ \beta_1 = \sqrt{I_1 \| B \bullet C \|_F}, \quad \beta_2 = \sqrt{I_2 \| C \bullet A \|_F}, \quad \beta_3 = \sqrt{I_3 \| A \bullet B \|_F} \]

and

\[ \beta = \sqrt[3]{\beta_1 \beta_2 \beta_3}. \]
Balanced norm for Minimal sensitivity II

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

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 \( \alpha_n = \frac{\beta_n}{\sqrt[3]{\beta_1 \beta_2 \beta_3}} \).
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. (2018, 2020b)

Error Preserving Correction Method

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

$$\min \quad ss(\theta) = l_1 \|B \bullet C\|_F^2 + l_2 \|C \bullet A\|_F^2 + l_3 \|A \bullet B\|_F^2$$

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

where $\theta$ is vector of parameters and $\hat{Y} = \Psi A, B, C, \Psi$

Phan, Tichavsky and Cichocki, Error Preserving Correction: A Method for CP Decomposition at a Target Error Bound,

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

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

- Rewrite the objective function

\[
ss(\theta) = l_1 \|B \cdot C\|_F^2 + \text{tr}(QA^T(2)A(2))
\]

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

- In order to update the core tensor \(A\), we solve the minimization problem

\[
\min_{X} \quad \text{tr}(QX^TX)
\]

\[
\text{s.t.} \quad \|Y(1) - XZ^T\|_F^2 \leq \delta^2
\]

where \(Z\) is mode-(1,4) unfolding of \(B \cdot C\), \(X\) is mode-2 unfolding of the core tensor \(A\), i.e., \(A(2) = X\).
$A$ can be found in closed form as Spherical Constrained Quadratic Programming (SCQP) Gander et al. (1989); Phan et al. (2019).
Algorithm 1: Sensitivity Correction

Input: Data tensor $\mathbf{Y}$: $(l_1 \times l_2 \times l_3)$, and ranks $R$ and error bound $\delta$
Output: $\hat{\mathbf{Y}} = \langle \mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3 \rangle$: min $\text{ss}(\hat{\mathbf{Y}})$ \text{ s.t. } ||\mathbf{Y} - \hat{\mathbf{Y}}||^2_F \leq \delta^2$

begin
Initialize $\hat{\mathbf{Y}}$
repeat
for $n = 1, 2, 3$ do
$\mathbf{Z} = \mathcal{G}_2 \bullet \mathcal{G}_3$
$\mathbf{Q} = l_2 (I_{R_2} \otimes \mathcal{G}_{3,(3)} \mathcal{G}_{3,(3)}^T) + l_3 (\mathcal{G}_{2,(1)} \mathcal{G}_{2,(1)}^T \otimes I_{R_1})$
Solve $\mathcal{G}_1 = \arg \min_{\mathbf{X}} \text{tr}(\mathbf{XQX}^T)$ \text{ s.t. } ||\mathbf{Y}_{(1)} - \mathbf{XZ}_{(1)}||^2_F \leq \delta^2$
Cyclic-shift $\mathbf{Y}$ and $\hat{\mathbf{Y}}$
end
until a stopping criterion is met
Balance norm of core tensors
end
Algorithm 2: Sensitivity Correction Algorithm

1. **Input**: Data tensor $\mathbf{Y}: (I_1 \times I_2 \times I_3)$, and ranks $R$
2. **Output**: $\hat{\mathbf{Y}} = \{G_1, G_2, G_3\}$
3. **begin**
4. **repeat**
5. Run ALS to update the TC tensor $\hat{\mathbf{Y}}$: $\min \| \mathbf{Y} - \hat{\mathbf{Y}} \|_F$
6. **if** $\text{ss}(\theta) \leq \text{ss}_{\text{max}}$ **then**
7. Perform sensitivity correction with $\delta = \| \mathbf{Y} - \hat{\mathbf{Y}} \|_F$
8. **end**
9. **until** a stopping criterion is met
10. **end**
Different from the Sensitivity Correction method, we propose a TC decomposition with bounded sensitivity

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

In order to update the core tensor $A$, we solve the minimization problem

\[
\min_X \|Y^{(1)} - XZ^T\|_F^2
\]
\[
\text{s.t.} \quad \text{tr}(XQX^T) \leq \gamma_n
\]

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

Similar to the SSC algorithm, $A$ can be found in closed form.
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.
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%.
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

$$\|Y - X\|_F \leq \varepsilon \|Y\|_F$$

Example: Fitting Images by TC-decomposition II

(a) Mandrill

(b) Peppers

Anh-Huy Phan

Tensor Chain Decomposition
Example: Fitting Images by TC-decomposition III

(c) Lena

(d) Barbara
Example: Fitting Images by TC-decomposition IV

(e) Tiffany

(f) House
**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.
Low-rank approximation reduces the number of parameters in convolutional layers. Thereby accelerate the inference of the network.

In Phan et al. (2020b), we show that CPD with sensitivity control can significantly improve compression of widely used CNN 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).
Figure: Decomposition of Layer2 convolutional kernel, rank-(6,10,5). Sensitivity of the estimated TC model after 3000 iterations was $6.5989e+06$, and increased to $1.8938e+07$ after 13000 iterations. SSC was applied after 3000 ALS updates and yielded a model with sensitivity of 32146.
Figure: Decomposition of Layer2 convolutional kernel, rank-(10,10,10). SSC was applied after 3000 ALS updates.
Figure: Decomposition of Layer2 convolutional kernel, rank-(9,18,12). SSC was applied after 3000 ALS updates.
Figure: Decomposition of Layer3 convolutional kernel, rank-(10,10,10). SSC was applied after 3000 ALS updates.
## Table: Relative approximation errors obtained using ALS and ALS+SSC

<table>
<thead>
<tr>
<th>Layer</th>
<th>Method</th>
<th>Setting 1</th>
<th>Setting 2</th>
<th>Setting 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ALS</td>
<td>0.6623</td>
<td>0.5116</td>
<td>0.3559</td>
</tr>
<tr>
<td></td>
<td>SSC</td>
<td><strong>0.6543</strong></td>
<td><strong>0.4917</strong></td>
<td><strong>0.3281</strong></td>
</tr>
<tr>
<td>3</td>
<td>ALS</td>
<td>0.8319</td>
<td>0.6856</td>
<td>0.5792</td>
</tr>
<tr>
<td></td>
<td>SSC</td>
<td><strong>0.8317</strong></td>
<td><strong>0.6779</strong></td>
<td><strong>0.5728</strong></td>
</tr>
<tr>
<td>4</td>
<td>ALS</td>
<td>0.8670</td>
<td>0.7932</td>
<td>0.6321</td>
</tr>
<tr>
<td></td>
<td>SSC</td>
<td><strong>0.8659</strong></td>
<td><strong>0.7928</strong></td>
<td><strong>0.6274</strong></td>
</tr>
<tr>
<td>5</td>
<td>ALS</td>
<td>0.8121</td>
<td>0.6833</td>
<td>0.5525</td>
</tr>
<tr>
<td></td>
<td>SSC</td>
<td><strong>0.8117</strong></td>
<td><strong>0.6817</strong></td>
<td><strong>0.5492</strong></td>
</tr>
</tbody>
</table>
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.


