D5: Databases and Information Systems
Tensors in Data Analysis, WS 2017–18
Homework #3: Variations and applications of the CP decomposition
Tutorial: **12 October 2017** at 14:30

Problem 1 (Maximum rank). It was stated in the lectures that the rank of a tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{I \times J \times K}$ is never more than

$$\min\{IJ, IK, JK\} .$$

Let I, J, and K be that $JK = \min\{IJ, IK, JK\}$ and let $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{I \times J \times K}$ be arbitrary. Your task is to construct $\boldsymbol{A} \in \mathbb{R}^{I \times JK}$, $\boldsymbol{B} \in \mathbb{R}^{J \times JK}$, and $\boldsymbol{C} \in \mathbb{R}^{K \times JK}$ such that

$$\boldsymbol{T}_{(1)} = \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T .$$

Hint: Construct $\boldsymbol{B}$ from identity matrices.

Solution. Let

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{T}_{(1)} \\
\boldsymbol{B} &= \underbrace{[\boldsymbol{I}_J\, \boldsymbol{I}_J \cdots \boldsymbol{I}_J]}_{K \text{ times}} \\
\boldsymbol{C} &= \begin{pmatrix} \boldsymbol{j}_J^T & 0 & \cdots & 0 \\ 0 & \boldsymbol{j}_J^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{j}_J^T \end{pmatrix} (K \text{ rows}) ,
\end{aligned}$$

where $\boldsymbol{I}_J$ is J-by-J identity matrix and $\boldsymbol{j}_J^T$ is J-dimensional row vector of all 1s. Now

$$\begin{aligned}
\boldsymbol{C} \odot \boldsymbol{B} &= [\boldsymbol{c}_1 \otimes \boldsymbol{b}_1\; \boldsymbol{c}_2 \otimes \boldsymbol{b}_2\; \cdots\; \boldsymbol{c}_{JK} \otimes \boldsymbol{b}_{JK}] \\
&= \begin{pmatrix} \boldsymbol{I}_J & 0 & \cdots & 0 \\ 0 & \boldsymbol{I}_J & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{I}_J \end{pmatrix} \\
&= \boldsymbol{I}_{JK} = (\boldsymbol{C} \odot \boldsymbol{B})^T .
\end{aligned}$$

Hence, $\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T = \boldsymbol{A}\boldsymbol{I}_{JK} = \boldsymbol{A} = \boldsymbol{T}_{(1)}$.

One can also show that this construct admits $\boldsymbol{T}_{(2)} = \boldsymbol{B}(\boldsymbol{C} \odot \boldsymbol{A})^T$ and $\boldsymbol{T}_{(3)} = \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^T$, though those proofs require much more complex subscripting.

Problem 2 (Nonnegative INDSCAL). Present an algorithm for nonnegative 3-way INDSCAL. That is, given a nonnegative 3-way tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}_{\geqslant 0}^{I \times J \times K}$ and an integer R, find matrices $\boldsymbol{A} \in \mathbb{R}_{\geqslant 0}^{I \times R}$, $\boldsymbol{B} \in \mathbb{R}_{\geqslant 0}^{J \times R}$, and $\boldsymbol{C} \in \mathbb{R}_{\geqslant 0}^{K \times R}$ that aim at minimizing

$$\|\boldsymbol{\mathcal{T}} - [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]\| \ .$$

Solution. *The problem statement is wrong. The real problem should be: Given* $\boldsymbol{\mathcal{T}} \in \mathbb{R}_{\geqslant 0}^{I \times I \times K}$*, find* $\boldsymbol{A} \in \mathbb{R}_{\geqslant 0}^{I \times R}$ *and* $\boldsymbol{C} \in \mathbb{R}_{\geqslant 0}^{K \times R}$ *that minimize* $\|\boldsymbol{\mathcal{T}} - [\![\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{C}]\!]\|$.

One option is to use the multiplicative update rules for NCP. If we let $\boldsymbol{\mathcal{Q}} = [\![\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{C}]\!]$, we have

$$\begin{aligned} a_{ir} &\leftarrow a_{ir} \frac{\sum_{j,k} a_{jr} c_{kr} (t_{ijk}/q_{ijk})}{\sum_{j,k} a_{jr} c_{kr}} \\ c_{kr} &\leftarrow c_{kr} \frac{\sum_{i,j} a_{ir} a_{jr} (t_{ijk}/q_{ijk})}{\sum_{i,j} a_{ir} a_{jr}} \ . \end{aligned}$$

The initialization of the factor matrices requires some attention. They should naturally be nonnegative, and we should aim to have them in a correct scale. For multiplicative update rules we also cannot have zero entries. For example, we can sample from uniform distribution over $(0, u)$, where we set u so that the expected value of the CP product of the random matrices, $\mathbb{E}[[\![\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{C}]\!]_{ijk}]$, is equal to the average value in the tensor, $\frac{1}{IJK} \sum_{i,j,k} t_{ijk}$. We can obtain this by setting

$$u = \frac{2}{R\sqrt[3]{IJK}} \sqrt[3]{\sum_{i,j,k} t_{ijk}} \ .$$

Problem 3 (CP-APR for KL-divergence). In CP-APR, we need to find a matrix $\boldsymbol{A}$ that minimizes

$$L(\boldsymbol{A}) = \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T - \boldsymbol{T}_{(1)} * \log(\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T) \ .$$

This is a type of a KL divergence. In nonnegative matrix factorization (NMF), we are given a nonnegative matrix $\boldsymbol{A} \in \mathbb{R}_{\geqslant 0}^{I \times J}$ and an integer K and we have to find nonnegative matrices $\boldsymbol{W} \in \mathbb{R}_{\geqslant 0}^{I \times K}$ and $\boldsymbol{H} \in \mathbb{R}_{\geqslant 0}^{K \times J}$ such that $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$.

The standard NMF algorithm for KL divergence has the following update rule:

$$\boldsymbol{W}_{ik} \leftarrow \boldsymbol{W}_{ik} \frac{\sum_{j=1}^{m} (\boldsymbol{A}_{ij}/(\boldsymbol{W}\boldsymbol{H})_{ij}) \boldsymbol{H}_{kj}}{\sum_{j=1}^{m} \boldsymbol{H}_{kj}} \ .$$

Adapt this update rule for the factor matrix $\boldsymbol{A}$ in the CP decomposition. How does it relate to the update rule

$$\boldsymbol{A} \leftarrow \boldsymbol{A} * \big(\boldsymbol{T}_{(1)} \oslash (\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T)\big)(\boldsymbol{C} \odot \boldsymbol{B})^T \ ,$$

presented in the lecture? (To recap, $\oslash$ is the element-wise division.)

Solution. The NMF KL update rule adapted to matrix $\boldsymbol{A}$ in NCP is

$$a_{ir} \leftarrow a_{ir} \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} \big(t_{ijk}/(\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T)_{ijk}\big)(\boldsymbol{C} \odot \boldsymbol{B})_{r,(jk)}}{\sum_{j=1}^{J} \sum_{k=1}^{K} (\boldsymbol{C} \odot \boldsymbol{B})_{r,(jk)}}$$

We can write this in a matrix format:

$$\boldsymbol{A} \leftarrow \boldsymbol{A} * \Big(\big(\boldsymbol{T}_{(1)} \oslash (\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T)\big)(\boldsymbol{C} \odot \boldsymbol{B})^T \operatorname{diag}((\boldsymbol{C} \odot \boldsymbol{B})^T \boldsymbol{1}_{JK})^{-1}\Big) \ .$$

Compared to the update rule for CP-APR, this has a normalization factor $\operatorname{diag}((\boldsymbol{C} \odot \boldsymbol{B})^T \boldsymbol{1}_{JK})^{-1}$.

Problem 4 (PARAFAC2). The PARAFAC2 decomposition is another variant of the CP decomposition, defined slice-wise as follows. Given K matrices $\boldsymbol{X}_k \in \mathbb{R}^{I_k \times J}$ and rank R, find K matrices $\boldsymbol{U}_k \in \mathbb{R}^{I_k \times R}$, diagonal matrices $\boldsymbol{S}_k \in \mathbb{R}^{R \times R}$, and a matrix $\boldsymbol{V} \in \mathbb{R}^{J \times R}$ such that

$$\sum_{k=1}^{K} \left\| \boldsymbol{X}_k - \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}^T \right\|_F$$

is minimized.

a) PARAFAC2 is related to CP, but how? Under which conditions is PARAFAC2 the same as the CP decomposition?

b) Consider following kind of health records data: We have longitudinal health records data over K patients and J attributes, such as diagnoses and medication. For each patient, we have collected these attributes over different time span and at different times, and each patient k is represented by a I_k-by-J matrix $\boldsymbol{X}_k$, where I_k is the number of observations for this patient, and $(\boldsymbol{X}_K)_{ij}$ is the value of variable j that observation point i. Notice that the observation points do not align between the users, that is, they correspond to different points in time. Assume we do rank-R PARAFAC2 to the collection of such matrices $\{\boldsymbol{X}_k\}_{k=1}^K$ and obtain $\{\boldsymbol{U}_k, \boldsymbol{S}_k\}_{k=1}^K$, and $\boldsymbol{V}$.

   We can assume that the columns of the J-by-R matrix $\boldsymbol{V}$ corresponds to some latent *phenotypes*, that is, they encode which diagnoses and medication "go together." How would you interpret the other factors?

Solution.

a) For PARAFAC2 to be equal to CP, it has to be that (1) $I_1 = I_2 = \cdots = I_K = I$, (2) the rows of $\boldsymbol{X}_k$ correspond to each other, and (3) $\boldsymbol{U}_1 = \boldsymbol{U}_2 = \cdots = \boldsymbol{U}_k = \boldsymbol{U}$. Then we can take the matrices $\boldsymbol{X}_k$ as the frontal slices of tensor $\boldsymbol{\mathcal{X}}$, set $\boldsymbol{A} = \boldsymbol{U}$, $\boldsymbol{B} = \boldsymbol{V}$, and arrange the values in the diagonal of $\boldsymbol{S}_k$ as the kth row of $\boldsymbol{C}$. In this case $\boldsymbol{X}_k \approx \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}^T$ for all k is equivalent to $\boldsymbol{\mathcal{X}} \approx [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$.

b) The diagonal matrices $\boldsymbol{S}_k$ indicate the *importance* or *strength* of each of the R phenotypes in the kth subject. The most relevant phenotype is the one with the highest value. Each column of $\boldsymbol{U}_k$ provides a *temporal signature* for each of the phenotypes in patient k, that is, they indicate when the phenotype has been observed and at which level.