On Avoiding Diverging Components in the Computation of the Best Low Rank Approximation of Higher-Order Tensors

Lieven De Lathauwer

Tech. Report 05-269, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2005

This report was written as a contribution to an e-mail discussion between Rasmus Bro, Lieven De Lathauwer, Richard Harshman and Lek-Heng Lim.

1. Orthogonality in one of the modes

Consider the approximation of a tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ by a rank-R tensor $\hat{\mathcal{A}}$, given by

$$\hat{\mathcal{A}} = \sum_{r=1}^{R} \lambda_r \, \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \cdots \circ \mathbf{u}_r^{(N)},$$

where $R \leqslant I_1$. Denote $X = (\lambda_1, \ldots, \lambda_R; \mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_R^{(1)}; \ldots; \mathbf{u}_1^{(N)}, \ldots, \mathbf{u}_R^{(N)})$. We have the following theorem.

Theorem 1 *Under the condition that the vectors $\mathbf{u}_r^{(1)}$, $r = 1, \ldots, R$, are mutually orthonormal, the function*

$$f(X) = \|\mathcal{A} - \sum_{r=1}^{R} \lambda_r \, \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \cdots \circ \mathbf{u}_r^{(N)}\|^2 \tag{1}$$

attains its infimum.

Proof: Let $\mathbf{U}^{(n)} = [\mathbf{u}_1^{(n)} \ldots \mathbf{u}_R^{(n)}]$, $n = 1, \ldots, N$, and let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_R)$. Matricizing (1), we obtain

$$f(X) = \|\mathbf{A}_{(1)} - \mathbf{U}^{(1)} \cdot \mathbf{\Lambda} \cdot (\mathbf{U}^{(2)} \odot \cdots \odot \mathbf{U}^{(N)})^T\|^2. \tag{2}$$

Let $\overline{\mathbf{U}}^{(1)} = (\mathbf{U}^{(1)}, \mathbf{U}^{(1)^\perp})$ be (square) unitary. For any choice of $\overline{\mathbf{U}}^{(1)}$, we have

$$f(X) = \|\overline{\mathbf{U}}^{(1)^H} \cdot \mathbf{A}_{(1)} - \begin{pmatrix} \mathbf{I}_{R \times R} \\ \mathbf{O}_{(I_1 - R) \times R} \end{pmatrix} \cdot \mathbf{\Lambda} \cdot (\mathbf{U}^{(2)} \odot \cdots \odot \mathbf{U}^{(N)})^T\|^2. \tag{3}$$

Define $\mathcal{B} = \mathcal{A} \times_1 \overline{\mathbf{U}}^{(1)^H}$. Denote the $(I_2 \times I_3 \times \ldots \times I_N)$-slices of $\mathcal{B}$ by $\mathcal{B}_1$, $\ldots$, $\mathcal{B}_{I_1}$. Then

$$f(X) = \sum_{r=1}^{R} \|\mathcal{B}_r - \lambda_r \, \mathbf{u}_r^{(2)} \circ \mathbf{u}_r^{(3)} \circ \cdots \circ \mathbf{u}_r^{(N)}\|^2 + c(\mathbf{U}^{(1)}), \tag{4}$$

in which $c(\mathbf{U}^{(1)}) = \|\mathcal{A} \times_1 (\mathbf{U}^{(1)^\perp})^H\|^2$. We conclude that the problem reduces to a set of best rank-1 approximation problems, for which the infimum is attained. ∎

2. Bounded condition number in one mode

In the previous section the condition number of $\mathbf{U}^{(1)}$ was taken equal to one. This constraint can be relaxed. We have the following theorem.

Theorem 2 *Let the vectors $\mathbf{u}_r^{(n)}$, $r = 1, \ldots, R$, have unit norm. Under the condition that the condition number $\kappa(\mathbf{U}^{(1)}) \leqslant k$, the function $f(X)$ attains its infimum.*

Proof: All level sets of the cost function are compact. Closed: cf. note Lek-Heng. Bounded: below.

We assume that all vectors $\mathbf{u}_r^{(n)}$, $r = 1, \ldots, R$, $n = 1, \ldots, N$, have unit-length. Hence, we have to show that $\lambda_r \to \infty$ implies that $f(X) \to \infty$. Let $\lambda = (\lambda_1, \ldots, \lambda_R)$. We have

$$f(X) = \|\mathrm{vec}(\mathcal{A}) - (\mathbf{U}^{(1)} \odot \cdots \odot \mathbf{U}^{(N)}) \cdot \lambda\|^2. \tag{5}$$

Reasoning by contradiction, we see that the condition number of $\mathbf{U}^{(1)} \odot \cdots \odot \mathbf{U}^{(N)}$ is bounded because the condition number of $\mathbf{U}^{(1)}$ is bounded. This implies that $\|(\mathbf{U}^{(1)} \odot \cdots \odot \mathbf{U}^{(N)}) \cdot \lambda\| \to \infty$ whenever $\lambda_r \to \infty$. As a result, $f(X) \to \infty$ whenever $\lambda_r \to \infty$. ∎

From a practical point of view, in an ALS algorithm, the constraint could be imposed by replacing the current estimate of $\mathbf{U}^{(1)}$ by its best approximation with condition number at most k. This approximation is simply obtained by replacing the singular values that are more than k times smaller than the dominant singular value σ_1, by σ_1/k.

3. Zero-correlation in one mode

One can also impose that the factors in one mode are uncorrelated. In that case, the matrix $\mathbf{U}^{(1)}$ is of the form $\mathbf{1} \cdot \mathbf{m}^T + \mathbf{Q} \cdot \mathbf{\Omega}$, in which $\mathbf{1}$ is a vector that contains only ones, $\mathbf{m}$ contains the means of the different factors, $\mathbf{Q}$

is column-wise orthogonal and $\mathbf{\Omega}$ is diagonal. Even when some entries of $\mathbf{\Omega}$ become big, this cannot lead to degeneracy, cf. Section 1. However, entries of $\mathbf{m}$ may also become big and may mutually cancel. This means that degeneracy cannot be completely avoided. If it occurs, two or more columns of $\mathbf{U}^{(1)}$ become proportional to $\mathbf{1}$. This is an event that happens with probability zero.