2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

Compact matrix form of the d-dimensional tensor decomposition

Ivan Oseledets[†]

[†]Institute of numerical mathematics, 119333, Moscow, Gubkina 8.
Email: ivan.oseledets@gmail.com

Abstract—We present a new tensor decomposition, which is nonrecursive, does not suffer from the curse of dimensionality, but has a viable stability properties and can be computed by a robust algorithm via a sequence of SVD decompositions. The new form gives a clear and convenient way to implement all basic operations efficiently. A fast recompression procedure is presented, as well as basic linear algebra operations.

1. Introduction

A new decomposition for multiway arrays is presented, which generalizes the singular value decomposition of a matrix, inherits its "good" properties (for example, existence of best approximation with fixed rank) stability with respect to perturbations and allows to use well established algorithms for the SVD to compute this new decomposition. The number of parameters is linear in the dimension d and it is not recursive and has a very simple form. Instead of the canonical decomposition (known also as CANDECOMP/PARAFAC model) [4, 3] written as

$$A(i_1, ..., i_d) = \sum_{s=1}^{R} u_1(i_1, s), \dots u_d(i_d, s) \qquad (1)$$

(with total number of parameters dRn), we consider the representation of a tensor in form

$$\begin{aligned} &A(i_1, \dots, i_d) = \\ &= \sum_{\alpha_1, \dots \alpha_{d-1}} G_1(i_1, \alpha_1) G_2(\alpha_1, i_2, \alpha_2) G_3(\alpha_2, i_3, \alpha_3) \dots \qquad (2) \\ &\dots G_{d-1}(\alpha_{d-2}, i_{d-1}, \alpha_{d-1}) G_d(\alpha_{d-1}, i_d), \end{aligned}$$

where G_1 has size $n_1 \times r_1$, G_d has size $n_d \times r_d$, and for $2 \leq k \leq (d-1)$ G_k is a three-dimensional tensor of size $n_k \times r_{k-1} \times r_k$. This decomposition is called *TT-decompositions* (from *tensor-train* decomposition) Numbers r_k will be called *compression ranks*, and for simplicity assume that they are all of the same order: $r_k \sim r$.

Also (2) can be written in the equivalent matrix form:

$$A(i_1, \dots, i_d) = G_1^{(i_1)} G_2^{(i_2)} \dots G_d^{(i_d)}, \qquad (3)$$

where $G_1^{(i_1)}$ is $1 \times r_1$ row, $G_k^{(i_k)}$ is a $r_{k-1} \times r_k$ matrix, and $G_d^{(i_d)}$ is $r_{d-1} \times 1$ column.

2. Computing compression ranks

Compression ranks are *computable* and their computation is reduced to the estimation of the ranks of the special *unfolding matrices* of a tensor, defined as

$$A_k = A(i_1, i_2, \dots, i_k; i_{k+1}, \dots, i_d),$$

i.e. the first k indices enumerate the rows of A_k and the last $(d-k)$ enumerate its columns. Then the following theorem holds:

Theorem 1. *There exists a decomposition of form* (2) *with*

$$r_k = \operatorname{rank} A_k.$$

The compression ranks are bounded from above by the canonical rank of a tensor.

Moreover, instead of the canonical rank R we can take so-called effective tensor rank. **A** [2, 7].

Using theorems 1 and 2, the following estimate is obtained (2)

Theorem 3. *If tensor* **A** *has canonical rank R, then there exists a representation* (2) *with the number of parameters*

$$(d-2)nR^2 + 2nR.$$

Using additionaly the Tucker decomposition [6, 5], this estimate can be improved to $(d-2)R^3 + dnR$.

3. Basic linear algebra operations

As an example of how to use TT-decomposition (2), consider the evaluation of the multidimensional contraction:

$$W = \sum_{i_1, i_2, \dots, i_d} A(i_1, i_2, \dots, i_d) u_1(i_1) u_2(i_2) \dots u_d(i_d),$$

which appears for the numerical computation of the multidimensional integrals. Using the matrix representation (3) the problem is reduced to the sequence of one-dimensional convolutions

$$\Gamma_k(\alpha_{k-1}, \alpha_k) = \sum_{i_k=1}^{n_k} G_k(\alpha_{k-1}, i_k, \alpha_k) u_k(i_k),$$

and to the evaluation of the product

$$v_1\Gamma_2 \dots \Gamma_{d-1} v_d^\top,$$

where v_1, v_d are rows of the corresponding vectors.. That is the problem is reduced to d matrix-by-vector multiplications. The total cost is $O(dnr^2 + dr^2)$ also using preliminary Tucker decomposition this number can be reduced to $O(dnr + dr^3)$.

All basic algorithms can be developed: additions of tensor, matrix-by-vector product, norm.

4. Recompression in TT format

The most important procedure is the *recompression procedure* which consists of the following. Suppose some decomposition of form (2) is known and we want to get another decomposition with a fewer number of parameters. Such algorithm is absent for the canonical format and that is probably the most serious drawback of this format. For TT decomposition the situation is perfect and such an algorithm exists and it is based on the standard algorithms, SVD and QR decomposition. The idea is based on the following. If some TT representation is given then for any selected mode k it gives a skeleton (dyadic) approximation of the corresponding unfolding A_k:

$$A_k = U_k V_k^\top,$$

where U_k is $n^k \times r_k$ and V_k is $n^{d-k} \times r_k$. For the matrix case the recompression consists of two steps. First, QR decompositions of U_k and V_k are computed,

$$U_k = Q_u R_u, ;\ V_k = Q_v R_v,$$

and then for a "small" $r_k \times r_k$ matrix $R_u R_v^\top$, its truncated singular value decomposition is computed, and that gives the truncated SVD of the initial matrix. The problem is that the row dimensions of U_k and V_k are large and depend on d expontially. However, they have a special TT structure and its QR decomposition can be computed fast in a structured way. U_k is represented as

$$U_k(i_1, i_2, \dots, i_k, \alpha_k) = \sum_{\alpha_1,\dots,\alpha_{k-1}} G_1(i_1, \alpha_1) \quad (4)$$
$$G_2(\alpha_1, i_2, \alpha_2) \dots G_k(\alpha_{k-1}, i_k, \alpha_k).$$

To compute QR decomposition, first QR decomposition of the $n_1 \times r_1$ matrix U_1 is computed, yielding a $r_1 \times r_1$ matrix R_1, which is transfered to the second core:

$$G_2'(\alpha_1', i_1, \alpha_2) = \sum_{\alpha_1} G(\alpha_1, i_1, \alpha_2) R(\alpha_1, \alpha_1').$$

Then the second core is treated as a $\alpha_1 n_1 \times r_2$ matrix, $G_2'(\alpha_1' i_1, \alpha_2)$, its QR decomposition is computed, the Q factor is reshaped into a new core $Q_2(\alpha_1', i_1, \alpha_2')$ and the R factor is transfered to the right core G_3 and so on. It can be shown that this simple to implement algorithm gives an exact QR-decomposition of the matrix U_k with a Q factor in a TT format. The same holds for V_k, so the final recompression algorithm works from left-to-right, successively computing QR decompositions in the TT format, and truncated SVD decompositions.

5. Comparison of two formats

In the end, let us compare the two formats. r can be much smaller that R and the new format will be more effective than the old one. Moreover, the estimate for the recompression procedure in the canonical format is given as in the work [8], where no theoretical estimates are presented (for some cases the method may not converge, or converge to a local minimum due to the unstable nature of the canonical decomposition).

	Canonical	TT
Number of parameters	$O(dnR)$	$O(dnr + (d-2)r^3)$
Matrix-by-vector	$O(dn^2R^2)$	$O(dn^2r^2 + dr^6)$
Addition	$O(dnR)$	$O(dnr)$
Recompression	$O(dnR^2 + d^3R^3)$	$O(dnr^2 + dr^4)$
Convolution	$O(dnR)$	$O(dnr + dr^3)$

Table 1: Format comparison.

References

[1] Beylkin G., Mohlenkamp M. J. "Numerical analysis in higher dimensions" *Proc. Natl. Acad. Sci. USA*, 2002, vol. 99, No. 16, pp. 1046–10251.

[2] Bini D., Capovani M. "Tensor rank and border rank of band Toeplitz matrices" *SIAM J. Comput.*, 1987, vol. 2, pp. 252–258.

[3] Carroll J.D., Chang J.J. "Analysis of individual differences in multidimensional scal- ing via n-way generalization of Eckart-Young decomposition" *Psychometrica*, vol. 35, 1977, pp. 283-319.

[4] Harshman R.A. "Foundations of the Parafac procedure: models and conditions for an explanatory multimodal factor analysis" *UCLA Working Papers in Phonetics*, vol. 16, 1970, pp. 1–84.

[5] Oseledets I., Savostyanov D., Tyrtyshnikov E. "Tucker dimensionality reduction in linear time" *SIAM J. Matrix Anal. Appl.*, 2008, vol. 30, No. 3, pp. 939–956.

[6] Tucker L.R. "Some mathematical notes on three-mode factor analysis", *Psychometrika*, 1966, vol. 31, pp. 279–311.

[7] Oseledets I.V., Tyrtyshnikov E.E, "On the recursive representation of multidimensional tensors", *Doklady RAS*, vol. 427, No. 1.

[8] Espig M. “Effiziente Bestapproximation mittels Summen von Elementartensoren in hohen Dimensionen” Phd thesis, 2007.