Göran Bergqvist and
Erik G. Larsson

The Higher-Order Singular Value Decomposition: Theory and an Application

In many areas of science and technology, data structures have more than two dimensions, and are naturally represented by multidimensional arrays or tensors. Two-dimensional matrix methods, such as the singular value decomposition (SVD), are widespread and well studied mathematically. However, they do not take into account the multidimensionality of data. In some scientific areas, notably chemometrics and psychometrics, tensor methods have been developed and used with great success since the 1960s for the analysis of multidimensional data.

RELEVANCE

During the last decade, there has been a fast development of mathematical theory, new algorithms, and new application areas. The tensor view has been introduced in diverse applications, including signal and image processing, bioinformatics, visualization, pattern recognition, data mining, brain modeling, and environmental modeling. We give an introduction to state-of-the-art tensor methods, especially the higher-order SVD (HOSVD), with an application in signal processing.

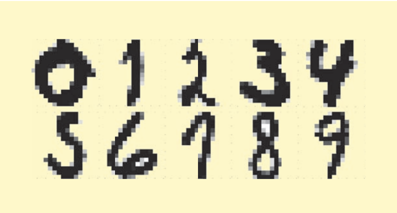
PREREQUISITES

The reader should be familiar with linear algebra, especially the SVD. We shall consider real-valued matrices and tensors unless stated otherwise.

PROBLEM STATEMENT

To illustrate the HOSVD, we will consider the problem of automatically recognizing a handwritten digit. We start with the training set of 7,291 digits in the U.S. postal service database, where each digit is 16×16 pixels, and each pixel has a gray scale intensity that is a real number between –1 and 1. See Figure 1 for examples of digits in this training set. For each $d = 0, 1, \ldots, 9$, one can define a $256 \times N(d)$ matrix A_d, where $N(d)$ is the number of appearances of d among the 7,291 test digits. Then, given a new test digit $z_i \in \mathbb{R}^{256}$, one can check for which d the new digit is closest to a linear combination of columns in A_d. To do this, for each d, find the coefficients $\alpha_j^{(d)}$ that minimize $\|z_i - \sum_{j=1}^{N(d)} \alpha_j^{(d)}(A_d)_{ij}\|$. With the standard norm on $\mathbb{R}^{256}$ these are least-squares problems. From a computational point of view, and to save memory, it is desirable to reduce the size of the matrices before beginning the process. A standard tool to obtain such a data compression of the A_ds, is to truncate the SVD of each one and obtain much smaller matrices $\widetilde{A}_d$, against which the new unknown digit should be tested.

DURING THE LAST DECADE, THERE HAS BEEN A FAST DEVELOPMENT OF MATHEMATICAL THEORY, NEW ALGORITHMS, AND NEW APPLICATION AREAS.



[FIG1] **Examples of digits of all ten classes in the training set. This figure was created by widely used data freely available at http://www-stat.stanford.edu/~tibs/ElemStatLearn/.**

Another possibility is to view the training set as a single $256 \times N^* \times 10$ three-way-array (3-array) T_{ijk}, where the three modes represent pixels, digits in training set, and class. Here, N^* is the largest $N(d)$ for $d = 0, \ldots, 9$, in fact $N^* = N(0) = 1194$. For values of d with $N(d) < N^*$, some digits need to be repeated to obtain a complete 3-array. The question is: can this higher-order array view of the data be the basis to produce algorithms with improved computational efficiency without a loss in accuracy?

SOLUTIONS

To solve the problem, we seek generalizations to q-arrays of the SVD, which is a standard matrix-algebraic tool in many applications. For an $(m \times n)$-matrix A we recall the SVD as being

$$\mathrm{A} = \mathrm{U}\Sigma\mathrm{V}^T = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \otimes \mathbf{v}_k. \tag{1}$$

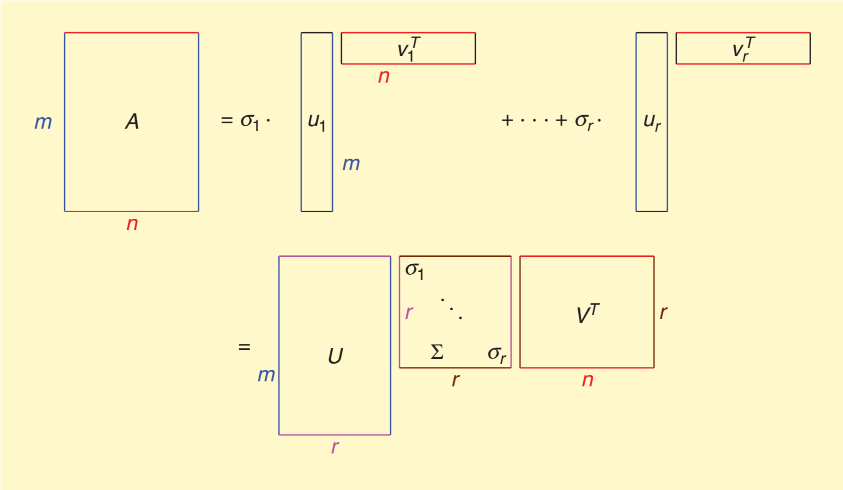
That is, for the elements A_{ij} of A,

$$A_{ij} = \sum_{k=1}^{r} U_{ik}\Sigma_{kk}V_{jk} = \sum_{k=1}^{r} \sigma_k U_k V_{jk}.$$

The SVD is illustrated in Figure 2. Here $\otimes$ denotes the tensor (or outer) product: $\boldsymbol{x} \otimes \boldsymbol{y} \triangleq \boldsymbol{x}\boldsymbol{y}^T$. Also, $r \leq \min(m, n)$ is the rank of A, that is, the dimension of the space spanned by the columns of A or equivalently the dimension of the space spanned by its rows. Σ is a diagonal $(r \times r)$ matrix with the nonzero singular values of A (the square roots of the eigenvalues of $\mathrm{A}^T\mathrm{A}$) on its diagonal. The singular values are real valued and nonnegative, being adopted the following convention $\sigma_1 > \cdots > \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_n$. $\mathbf{u}_k$ and $\mathbf{v}_k$ are the orthonormal columns of the matrices U $(m \times r)$ and V $(n \times r)$, respectively, with $\mathbf{v}_k$ being eigenvectors of $\mathrm{A}^T\mathrm{A}$ and $\mathbf{u}_k = \mathrm{A}\mathbf{v}_k/\sigma_k$. U and V can be

Digital Object Identifier 10.1109/MSP.2010.936030

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY KANPUR. Downloaded on March 23,2024 at 06:18:38 UTC from IEEE Xplore. Restrictions apply.



[FIG2] The SVD of a matrix A. The matrices U and V can be expanded with columns to quadratic orthogonal matrices. Σ is then augmented with zero elements so that its size becomes equal to that of A.

augmented with columns to square and orthogonal $(m \times m)$ and $(n \times n)$ matrices. Σ is then expanded with zero elements to an $(m \times n)$ matrix.

A fundamental fact is that the set of truncated series

$$A \approx \sum_{k=1}^{s} \sigma_k \mathrm{u}_k \mathrm{v}_k^T \quad (s < r) \qquad (2)$$

is the best rank-s approximation of A, both with respect to the operator norm and to the Frobenius norm. In many applications, one wants to approximate a data matrix with a low-rank matrix. The SVD does this in the best way. Such low-rank approximations can be used, for example, for denoising or data compression. In statistics, the SVD is also called the principal component expansion of the matrix $\mathbf{A}^T\mathbf{A}$.

The SVD is useful whenever we have a two-dimensional data set $\{A_{ij}\}$, which is naturally expressed in terms of a matrix A. In many applications, such as the one we consider here, we have a multidimensional data set $\{T_{i_1 \ldots i_q}\}$, $1 \le i_j \le n_j$, of dimension q, say. In this case, the data may be arranged into a multiway array (also known as multiarray or q-mode array or tensor) $\mathbf{T}$. The basic problem is whether tensors can be approximated in a fashion similar to the truncated-SVD expansion in (2). The case of interest is $q > 2$ since for $q = 2$, $\mathbf{T}$ is a conventional matrix and we can use the SVD. What are the possible generalizations of the SVD to $q > 2$?

GENERALIZATIONS OF THE SVD

The SVD may be generalized to higher-order tensors or multiway arrays in several ways. The two main approaches are the so-called Tucker/HOSVD decomposition and the CP expansion (from canonical decomposition (CANDECOMP) and parallel factors (PARAFAC) [4]). The CP expansion is a special case of the Tucker/HOSVD decompositions. For simplicity, we present these decompositions for tensors of order $q = 3$. This shows all fundamental differences to the case of a conventional matrix $(q = 2)$, and generalizations to $q > 3$ are rather direct.

[THE SVD MAY BE GENERALIZED TO HIGHER-ORDER TENSORS OR MULTIWAY ARRAYS IN SEVERAL WAYS.]

The Tucker decomposition of an $(m \times n \times p)$ tensor T of order $q = 3$ is

$$T = \sum_{I=1}^{M} \sum_{J=1}^{N} \sum_{K=1}^{P} G_{IJK}\, \mathrm{u}_I \otimes \mathrm{v}_J \otimes \mathrm{w}_K$$

or for the components,

$$T_{ijk} = \sum_{I=1}^{M} \sum_{J=1}^{N} \sum_{K=1}^{P} G_{IJK} U_{iI} V_{jJ} W_{kK}. \qquad (3)$$

The HOSVD is a special case of (3) when the matrices involved are orthogonal and matrix slices of G are mutually orthogonal; we return to this shortly. The CP expansion of $\mathbf{T}$ is

$$\mathbf{T} = \sum_{l=1}^{r} x_l \otimes y_l \otimes z_l \quad \text{or}$$
$$T_{ijk} = \sum_{l=1}^{r} X_{il} Y_{jl} Z_{kl}. \qquad (4)$$

The Tucker and CP decompositions are illustrated in Figure 3. Here, x_l and u_I are the columns of the matrices $\mathbf{X}$ and $\mathbf{U}$, and G_{IJK} are the components of an $(M \times N \times P)$-core tensor $\mathbf{G}$. We will discuss the Tucker and CP decompositions separately, but note that the CP expansion is the special case of the Tucker expansion when G is superdiagonal, i.e., $G_{IJK} = 0$ if any two indices are distinct, and $M = N = P = r$. In principle, G may be larger than T.

The are several rank concepts for tensors [1], [3], [4]. The rank of T is the minimal possible value of r in the CP expansion (4). This rank is always well defined. The column (mode-1) rank of T is the dimension of the subspace of $\mathbb{R}^m$ spanned by the np columns of T (for every fixed pair of values of jk we have such a column). The row (mode-2) rank r_2 and the mode-3 rank r_3 are defined analogously. The triple (r_1, r_2, r_3) is called the multirank of T. A typical rank of T is a rank that appears with nonzero probability if the elements T_{ijk} are randomly chosen according to a continuous probability distribution. A generic rank is a typical rank which appears with probability one. In the matrix case $(q = 2)$, the number of terms r in the SVD expansion is always equal to the column rank of the matrix, which in turn is equal to the row rank. However, for tensors, r, r_1, r_2, and r_3 can all be different. For matrices $(q = 2)$, the typical and generic ranks of an $(m \times n)$-matrix are always min (m, n). However, for a higher-order tensor a generic rank over the real numbers does not necessarily exist. Both the typical and generic ranks of an $(m \times n \times p)$-tensor may be strictly greater than min (m, n, p), and are hard to calculate.

THE TUCKER AND HOSVD EXPANSIONS

Most common for the Tucker decomposition is to assume that $M \le m$, $N \le n$,

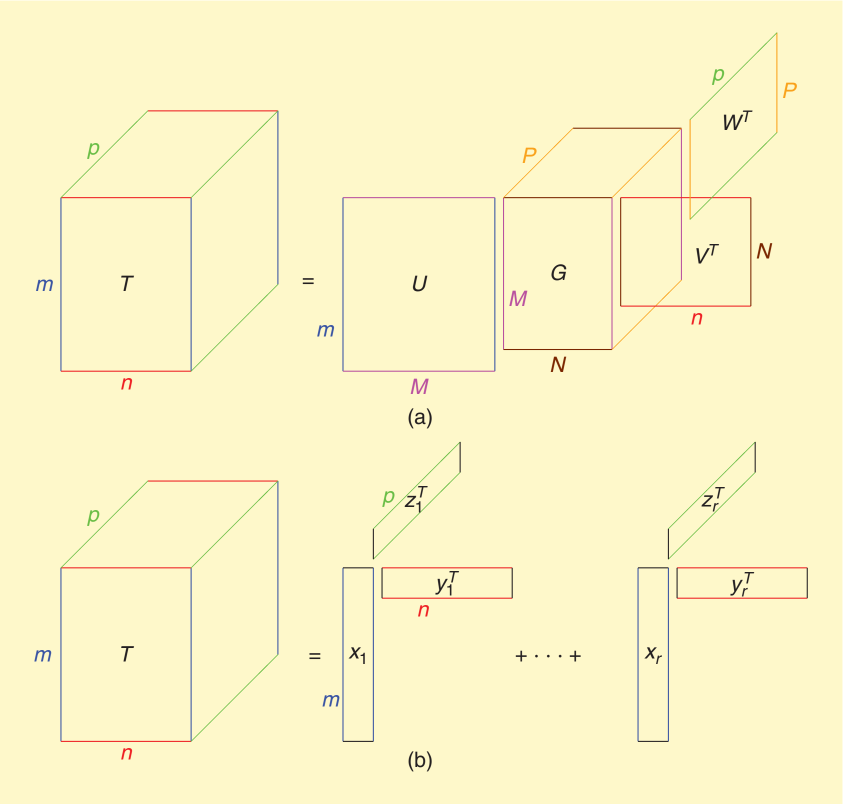
Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY KANPUR. Downloaded on March 23,2024 at 06:18:38 UTC from IEEE Xplore. Restrictions apply.

and $P \leq p$, so that $\mathbf{G}$ is a compression of $\mathbf{T}$. The vectors $\mathbf{u}_i$, $\mathbf{v}_j$, and $\mathbf{w}_k$ are seen as columns of matrices $\mathbf{U}$ $(m \times M)$, $\mathbf{V}$ $(n \times N)$, and $\mathbf{W}$ $(p \times P)$ respectively, and they are usually assumed to be orthonormal. The matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ are sometimes seen as generalized principal components.

In [2], it was shown that $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ can be taken to be orthogonal matrices, so that $\mathbf{G}$ has the same size as $\mathbf{T}$. Simultaneously, the different matrix slices of $\mathbf{G}$ along any mode can be chosen to be mutually orthogonal (with respect to the standard inner product on matrix spaces), and with decreasing Frobenius norm. This is clearly a generalization of the matrix SVD, in which rows and columns of the singular value matrix Σ are mutually orthogonal and with decreasing norm. In this case, the Tucker decomposition is called the HOSVD. Owing to the orthogonality conditions, the HOSVD is essentially unique. The HOSVD is "rank revealing," which means that if $\mathbf{T}$ has multirank (r_1, r_2, r_3), then the last $m - r_1$, $n - r_2$, and $p - r_3$ slices along the different modes in $\mathbf{G}$ are zero matrices. Then one can use thin matrices $\mathbf{U}$ $(m \times r_1)$, $\mathbf{V}$ $(n \times r_2)$, $\mathbf{W}$ $(p \times r_3)$, and also a smaller $(r_1 \times r_2 \times r_3)$ core tensor, to write the expansion.

There are several algorithms for calculating Tucker expansions and the HOSVD [2], [4]. In HOSVD, $\mathbf{U}$ can be calculated by performing a matrix SVD on the $(m \times np)$ matrix obtained by a flattening or matricization of $\mathbf{T}$. $\mathbf{V}$ and $\mathbf{W}$ are found in the same way. Since $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ are orthogonal, $\mathbf{G}$ is then easily calculated from $G_{IJK} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} T_{ijk} U_{iI} V_{jJ} W_{kK}$.

As opposed to the SVD for matrices, the order-$(s_1 \times s_2 \times s_3)$ truncation of the HOSVD is not the best multirank-(s_1, s_2, s_3) approximation of $\mathbf{T}$. However, for many applications, it is considered to be sufficiently good, or else it can serve as an initial value in algorithms for finding the best approximation. For the problem of finding the best multirank- (s_1, s_2, s_3) approximation of $\mathbf{T}$, alternating least-squares has been the traditional method but very recently improved methods have been developed [4].



[FIG3] (a) The Tucker and HOSVD expansions of a tensor T. For data compression, the core tensor G is smaller than T (that is, $M < m$, $N < n$, $P < p$), and U, V and W are thin matrices. In HOSVD, G has the same size as T ($M = m$, $N = n$, $P = p$). U, V, and W are then quadratic orthogonal matrices. (b) The CP (CANDECOMP/PARAFAC) expansion of a rank-r tensor T.

THE CP EXPANSION

Formally, the CP expansion (4) is the special case of the Tucker decomposition (3) when $\mathbf{G}$ is superdiagonal. It is important to know to what degree the terms in the CP expansion (4) are unique. The uniqueness of the SVD for matrices (1) essentially depends on the orthogonality conditions on $\mathbf{U}$ and $\mathbf{V}$.

THERE ARE SEVERAL ALGORITHMS FOR CALCULATING TUCKER EXPANSIONS AND THE HOSVD.

The uniqueness of the CP expansion (4) is, up to a trivial rescaling and reordering, guaranteed under milder assumptions, and orthogonality cannot in general be imposed. One sufficient condition for uniqueness is $k_X + k_Y + k_Z \geq 2(r+1)$ [5], where k_X is the largest number such that any k_X columns of $\mathbf{X}$ are linearly independent. Various other conditions have also been derived [4]. To calculate the CP expansion, one can use alternating least-squares methods to minimize the difference between $\mathbf{T}$ and an expansion with a fixed number of terms. One then increases the number of terms until a match between $\mathbf{T}$ and the series is obtained.

The truncated CP expansion with $s < r$ terms is in general not the best rank-s approximation of $\mathbf{T}$ but, again, for many applications it is sufficiently good. To calculate the best rank-s approximation one can use the same method as for determining the entire expansion but with the number of terms kept to s. An important difference to the SVD for conventional matrices is that the best rank-1 approximation may not be one of the terms in the best rank-2 approximation, and so on. Even more importantly, the rank-s

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY KANPUR. Downloaded on March 23,2024 at 06:18:38 UTC from IEEE Xplore. Restrictions apply.

approximation problem is ill-posed [3]. Specifically, a sequence of tensors of rank s, such that $\inf(\|\mathrm{T} - \widetilde{\mathrm{T}}\|; \operatorname{rank}(\widetilde{\mathrm{T}}) = s)$ is approached, may converge to a tensor of rank greater than s. That is, the infimum is not attained by any tensor of rank $\leq s$. This problem is, of course, relevant for stability of algorithms and for applications.

THE APPLICATION OF HOSVD TO HANDWRITTEN DIGIT CLASSIFICATION

In [7], the HOSVD was applied to handwritten digit classification. The $256 \times N^* \times 10$ 3-array T_{ijk} of data from the training set is by HOSVD truncation reduced to a $m \times n \times 10$ 3-array. The HOSVD of T_{ijk} is

$$\begin{aligned} T_{ijk} &= \sum_{I=1}^{256} \sum_{J=1}^{1194} \sum_{K=0}^{9} G_{IJK} U_{iI} V_{jJ} W_{kK} \\ &\approx \sum_{I=1}^{m} \sum_{J=1}^{n} \sum_{K=0}^{9} G_{IJK} U_{iI} V_{jJ} W_{kK} \\ &= \sum_{I=1}^{m} \sum_{J=1}^{n} F_{IJk} U_{iI} V_{jJ}, \end{aligned} \qquad (5)$$

where $F_{Ijk} = \sum_{K=0}^{9} G_{IJK} W_{kK}$. Values of m and n between 30 and 60 were used, which means that the data were compressed by about 99%. Only the first m and n columns of U and V, respectively, need to be calculated. The reduced $m \times n \times 10$ tensor F_{IJk} is computed by $F_{Ijk} = \sum_{i=1}^{256} \sum_{j=1}^{1194} T_{ijk} U_{iI} V_{jJ}$.

For an unknown digit $z_i \in \mathbb{R}^{256}$, the low dimensional representation $(U^T z)_i = \sum_{j=1}^{256} U_{\mathrm{ji}} z_j \in \mathbb{R}^{\mathrm{m}}$ is calculated. For each $d = 0, \ldots, 9$ it is straightforward, by least-squares, to see how well $(U^T z)_i$ is approximated by the columns of the $m \times n$ matrix $(\widetilde{F}_d)_{ij} = F_{ijd}$. By a matrix SVD, $\widetilde{F}_d$ is further reduced to an $m \times k$ matrix with $k \approx 10$ during this process. The value of d with the smallest residual determines the classification of the digit.

With the tensor model, all digits from different classes are projected to a common subspace. Only one projection of a test digit z_i is then needed rather the ten, one for each d, needed if the training data is modeled as ten matrices. This saves memory in the test phase. Various tests run in [7] show that compared to other methods the algorithm is computationally efficient (and simple), and has a satisfactory error rate of only 5% when the data is compressed by 99%.

[THE TUCKER DECOMPOSITION AND THE HOSVD ARE MORE RECENT TOOLS THAN THE CP, AND THEREFORE THEY ARE NOT AS WIDELY KNOWN IN THE SP COMMUNITY.]

OTHER APPLICATIONS TO SIGNAL PROCESSING

The CP decomposition has been used to solve various problems in the statistical signal processing literature. For example, [8] considers a sensor array composed of several subarrays that receive a linear superposition of signals emitted by r sources. The model for the received signals has the precise form of (4) and the received tensor T has three dimensions: time, antenna index, and subarray index. Another example is blind multiantenna receivers for code-division multiple-access systems [6]. Here, the CP model (4) applies with r being the number of users whose signals are simultaneously received, and T representing the received data along the dimensions antenna, chip, and symbol index.

The Tucker decomposition and the HOSVD are more recent tools than the CP, and therefore they are not as widely known in the SP community. The HOSVD has, however, been used before in other related applications. For example, the recent paper [9] shows how the HOSVD can be used in image processing and face recognition. Therein, face image data were modeled via three tensors with texels, illuminations and views as the three modes, and recognition algorithms that exploit this structure were presented.

CONCLUSIONS: WHAT WE HAVE LEARNED

Tensor modeling and algorithms for computing various tensor decompositions (the Tucker/HOSVD and CP decompositions, as discussed here, most notably) constitute a very active research area in mathematics. Most of this research has been driven by applications. There is also much software available, including MATLAB toolboxes [4]. The objective of this lecture has been to provide an accessible introduction to state of the art in the field, written for a signal processing audience. We believe that there is good potential to find further applications of tensor modeling techniques in the signal processing field.

AUTHORS

Göran Bergqvist (gober@mai.liu.se) is a professor of applied mathematics in the Department of Mathematics at Linköping University in Sweden.

Erik G. Larsson (erik.larsson@isy.liu.se) is a professor and head of the Division for Communication Systems in the Department of Electrical Engineering at Linköping University.

REFERENCES

[1] P. Comon, J. M. F. ten Berge, L. De Lathauwer, and J. Castaing, "Generic and typical ranks of multiway arrays," *Linear Algebra Applicat.*, vol. 430, no. 11, pp. 2997–3007, 2009.

[2] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Applicat.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[3] V. De Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Applicat.*, vol. 30, no. 3, pp. 1084–1127, 2008.

[4] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sept. 2009.

[5] J. B. Kruskal, "Rank, decomposition, and uniqueness for 3-way and N-way arrays," in *Multiway Data Analysis*, R. Coppi and S. Bolasco, Eds. Amsterdam, The Netherlands: North-Holland, 1989, pp. 7–18.

[6] D. Nion and L. De Lathauwer, "A block component model-based blind DS-CDMA receiver," *IEEE Trans. Signal Processing*, vol. 56, no. 11, pp. 5567–5579, 2008.

[7] B. Savas and L. Eldén, "Handwritten digit classification using higher order singular value decomposition," *Pattern Recognit.*, vol. 40, no. 3, pp. 993–1003, 2007.

[8] N. D. Sidiropoulos, R. Bro, and G. B. Giannikis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[9] M. Vasilescu and D. Terzopoulos, "Multilinear (tensor) image synthesis, analysis and recognition," *IEEE Signal Processing Mag.*, vol. 24, no. 6, pp. 118–123, 2007.

[SP]

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY KANPUR. Downloaded on March 28,2024 at 06:18:38 UTC from IEEE Xplore. Restrictions apply.