Applications of tensor (multiway array) factorizations and decompositions in data mining

Morten Mørup*

Tensor (multiway array) factorization and decomposition has become an important tool for data mining. Fueled by the computational power of modern computer researchers can now analyze large-scale tensorial structured data that only a few years ago would have been impossible. Tensor factorizations have several advantages over two-way matrix factorizations including uniqueness of the optimal solution and component identification even when most of the data is missing. Furthermore, multiway decomposition techniques explicitly exploit the multiway structure that is lost when collapsing some of the modes of the tensor in order to analyze the data by regular matrix factorization approaches. Multiway decomposition is being applied to new fields every year and there is no doubt that the future will bring many exciting new applications. The aim of this overview is to introduce the basic concepts of tensor decompositions and demonstrate some of the many benefits and challenges of modeling data multiway for a wide variety of data and problem domains. © 2011 John Wiley & Sons, Inc. *WIREs Data Mining Knowl Discov* 2011 1 24–40 DOI: 10.1002/widm.1

INTRODUCTION

Tensors, or multiway arrays, are generalizations of vectors (first-order tensors) and matrices (second-order tensors) to arrays of higher orders ($N > 2$). Hence, a third-order tensor is an array with elements $x_{i,j,k}$. Tensor decompositions are in frequent use today in a variety of fields ranging from psychology, chemometrics, signal processing, bioinformatics, neuroscience, web mining, and computer vision to mention but a few.

Factorizing tensors have several advantages over two-way matrix factorization such as uniqueness of the optimal solution (without imposing constraints such as orthogonality and independence) and component identification even when only a relatively small fraction of all the data is observed (i.e., due to missing values). Furthermore, multiway decomposition techniques can explicitly take into account the multiway structure of the data that would otherwise be lost when analyzing the data by matrix factorization approaches by collapsing some of the modes. Tensor decompositions are in frequent use in psychometrics in order to address questions such as '*which group of subjects behave differently on which variables under which conditions?*'[1,2] In chemistry, tensor decomposition has been proven for low concentrations to be the physical model of fluorescence spectroscopy admitting unique recovery of the chemical compounds from sampled mixtures.[3,4] In neuroimaging, a tradition has been to average data across trials or groups of subjects for the extraction of the most reproducible neural activation. Here, tensor decomposition can efficiently extract the consistent patterns of activation, whereas noisy trials/subjects can be downweighted in the averaging process.[5–8] For signal processing, tensor decomposition forms an analysis framework to solve the blind source separation problem through the analysis of higher-order statistics,[9–11] whereas tensor decompositions have proven useful for the exploitation of different types of diversity in sensor array processing.[12,13] In computer vision, tensor decomposition enables the extraction of patterns that generalize well across common modes of variation,[14–16] whereas in bioinformatics, tensor factorization has proven useful for the understanding of cellular states and biological processes.[17–19] Lately, tensor factorization has become an important tool in web mining for exploratory analysis and comprehension of a large variety of data that are inherently multimodal.[20–23] As such, tensor decomposition is widely used in data

*Correspondence to: mm@imm.dtu.dk

Section for Cognitive Systems, DTU Informatics, Technical University of Denmark, Richard Petersens Plads, Bld. 321/118, 2800 Lyngby, Denmark

DOI: 10.1002/widm.1

mining and its importance is growing spurred by the computational power and storage capabilities of modern computers. Tensor decomposition is being applied to new fields every year and there is no doubt that tensor factorization will be an important framework for knowledge discovery of many types of modern large-scale data sets.

Tensor factorization has many challenges and open problems, particularly because its geometry is not yet fully understood, the occurrence of degenerate solutions, and no guarantee of finding the optimal solution. Furthermore, most tensor decomposition models impose a very restricted structure on the data which in turn require that data exhibit a strong degree of regularity. To overcome these limitations a variety of extensions and variants of tensor decomposition approaches have been proposed over the years. Thus, understanding the data generating process is key for the formulation of adequate tensor decomposition models that can well extract the inherent multimodal structure.

This overview will limit itself to the basic tensor decomposition models such as the Candecomp/Parafac (CP) and Tucker model, as well as their application in data mining. Other great introductory resources for tensor decomposition and their applications can be found in the recent review of Ref 24 the book on multiway analysis for the chemical sciences[28] as well as the book on applied multiway analysis of Ref 2. Furthermore, a good introduction to nonnegative tensors and their decompositions can be found in Ref 25. In the present paper, model estimation is reduced to a minimum considering only the simple and widely used alternating least squares (ALS) approach. For a thorough treatment of tensor model estimation approaches, we suggest that the reader consult Refs 24, 26, 27, and the references therein.

The paper is organized as follows: In '*Tensor Nomenclature*', we introduce standard tensor notation and operations, in '*The Tucker and Candecomp/Parafac Models*', we describe the two most widely used tensor decomposition approaches namely the Tucker and CP decompositions as well as some of their extensions. In '*Tensor Factorization for Data Mining*', we describe some of the applications of tensor factorization and decomposition in data mining. Because of space limitation, the aim of this article is to give an overview, thus, full credit cannot be given to all the many achievements made over the years of multiway/tensor analysis.

TENSOR NOMENCLATURE

Tensors and multiway arrays, also referred to as hypermatrices, are normally written in calligraphed letters. A general real tensor of order N is written $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$, we will use the following notation to more compactly denote the size of a tensor $\mathcal{X}^{I_1\times I_2\times\ldots\times I_N}$, whereas a given element of the tensor $\mathcal{X}$ is denoted by $x_{i_1,i_2,\ldots,i_N}$. The following section introduces the basic notation and operations that, for clarity, is given for third-order tensors, whereas they trivially generalize to tensors of arbitrary order.

Consider the third-order tensor $\mathcal{A}^{I\times J\times K}$ and $\mathcal{B}^{I\times J\times K}$. Scalar multiplication, addition of two tensors, and the inner product between two tensors are given by

$$\alpha\mathcal{B} = \mathcal{C}, \quad \text{where} \quad c_{i,j,k} = \alpha b_{i,j,k} \tag{1}$$

$$\mathcal{A} + \mathcal{B} = \mathcal{C}, \quad \text{where} \quad c_{i,j,k} = a_i + b_i \tag{2}$$

$$\langle \mathcal{A}, \mathcal{B}\rangle = \sum_{i,j,k} a_{i,j,k} b_{i,j,k} \tag{3}$$

As such, the Frobenius norm of a tensor is given by $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A}\rangle}$.

The n^{th} mode matricizing and unmatricizing operation maps a tensor into a matrix and a matrix into a tensor, respectively, i.e.,

$$\mathcal{X}^{I_1\times I_2\times\ldots\times I_N} \underset{\text{matricizing}}{\rightarrow} \mathbf{X}_{(n)}^{I_n\times I_1\cdot I_2\cdots I_{n-1}\cdot I_{n+1}\cdots I_N} \tag{4}$$

$$\mathbf{X}_{(n)}^{I_n\times I_1\cdot I_2\cdots I_{n-1}\cdot I_{n+1}\cdots I_N} \underset{\text{un-matricizing}}{\rightarrow} \mathcal{X}^{I_1\times I_2\times\ldots\times I_N} \tag{5}$$

The matricizing operation for a third-order tensor is illustrated in Figure 1. The n-mode multiplication of an order N tensor $\mathcal{X}^{I_1\times I_2\times\ldots\times I_N}$ with a matrix $\mathbf{M}^{J\times I_n}$ is given by

$$\mathcal{X} \times_n \mathbf{M} = \mathcal{X} \bullet_n \mathbf{M} = \mathcal{Z}^{I_1\times\ldots\times I_{n-1}\times J\times I_{n+1}\times\ldots\times I_N}, \tag{6}$$

$$z_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} x_{i_1,\ldots,i_{n-1},i_n,i_{n+1},\ldots,i_N} m_{j,i_n}. \tag{7}$$

Using the matricizing operation, this operation corresponds to $\mathbf{Z}_{(n)} = \mathbf{M}\mathbf{X}_{(n)}$. As a result, the matrix products underlying the singular value decomposition (SVD) can be written as $\boldsymbol{USV}^\top = \boldsymbol{S} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V} = \boldsymbol{S} \times_2 \boldsymbol{V} \times_1 \boldsymbol{U}$ as the order of the multiplication does not matter. The outer product of the three vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ is given by

$$\mathcal{X} = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c}, \text{ such that } x_{i,j,k} = a_i b_j c_k \tag{8}$$

The Kronecker product is given by

$$\boldsymbol{P}^{I\times J} \otimes \boldsymbol{Q}^{K\times L} = \boldsymbol{R}^{IK\times JL},$$
$$\text{such that} \quad r_{k+K(i-1),l+L(j-1)} = p_{ij}q_{kl}, \tag{9}$$

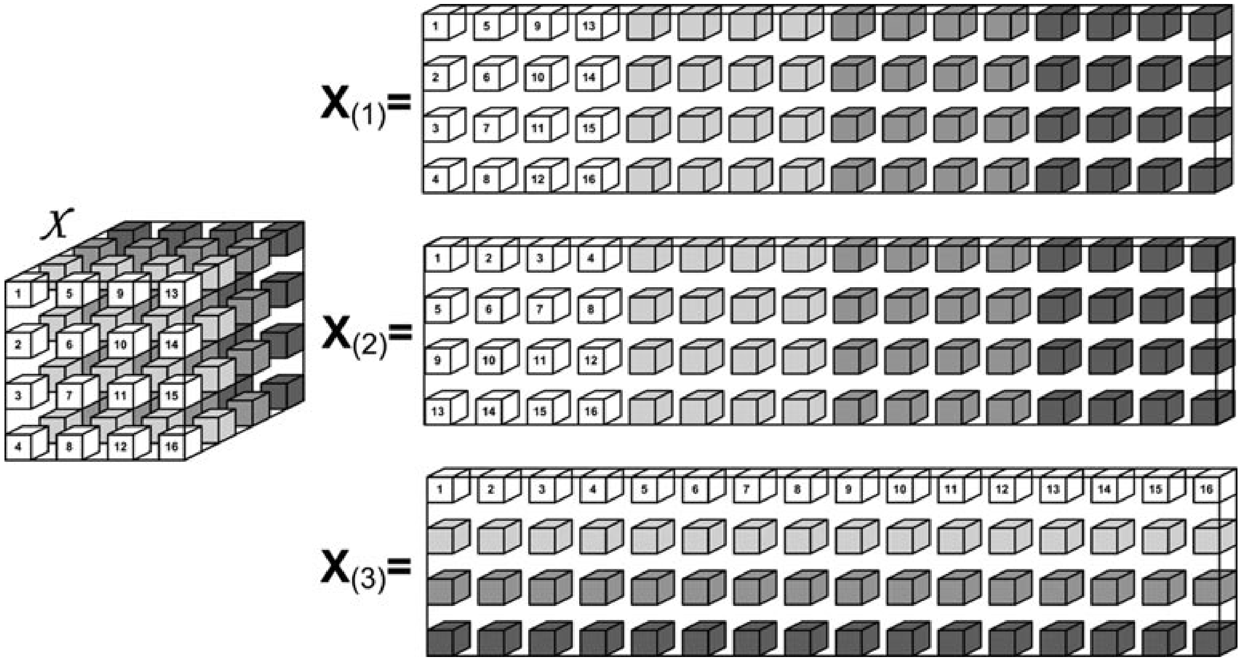

FIGURE 1 | The matricizing operation on a third-order tensor of size 4 × 4 × 4.

whereas the Khatri–Rao product is defined as a column-wise Kronecker product

$$\mathbf{A}^{I\times J}|\otimes|\mathbf{B}^{K\times J} = \mathbf{A}^{I\times J}\odot\mathbf{B}^{K\times J} = \mathbf{C}^{IK\times J},$$
$$\text{such that}\quad c_{k+K(i-1),j} = a_{ij}b_{kj}. \tag{10}$$

An important property when calculating the Moore–Penrose inverse (i.e., $\mathbf{A}^{\dagger} = (\mathbf{A}^{\top}\mathbf{A})^{-1}\mathbf{A}^{\top}$) of Kronecker and Khatri–Rao products are

$$(\mathbf{P}\otimes\mathbf{Q})^{\dagger} = (\mathbf{P}^{\dagger}\otimes\mathbf{Q}^{\dagger}) \tag{11}$$

$$(\mathbf{A}\odot\mathbf{B})^{\dagger} = [(\mathbf{A}^{\top}\mathbf{A})^{*}(\mathbf{B}^{\top}\mathbf{B})]^{-1}(\mathbf{A}\odot\mathbf{B})^{\top} \tag{12}$$

where $*$ denotes elementwise multiplication. This reduces the complexity from $\mathcal{O}(J^3L^3)$ to $\mathcal{O}(\max\{IJ^2, KJ^2, J^3, L^3\})$ and $\mathcal{O}(IKJ^2)$ to $\mathcal{O}(\max\{IKJ, IJ^2, KJ^2, J^3\})$, respectively. For additional properties of these matrix products see, also Ref 28. In Table 1, a summary of the operators described above can be found.

THE TUCKER AND CANDECOMP/PARAFAC MODELS

The two most widely used tensor decomposition methods are the Tucker model[29] and Canonical Decomposition (CANDECOMP)[30] also known as Parallel Factor Analysis (PARAFAC)[31] jointly abbreviated CP. In the following section, we describe the models for

TABLE 1 | Summary of the Utilized Variables and Operations. $\mathcal{X}$, $\mathbf{X}$, $\boldsymbol{x}$, and x are Used to Denote Tensors, Matrices, Vectors, and Scalars Respectively.

Operator	Name	Operation
$\langle\mathcal{A},\mathcal{B}\rangle$	Inner product	$\langle\mathcal{A},\mathcal{B}\rangle = \sum_{i,j,k} a_{i,j,k}b_{i,j,k}$
$\|\mathcal{A}\|_F$	Frobenius norm	$\sqrt{\langle\mathcal{A},\mathcal{A}\rangle}$
$\mathbf{X}_{(n)}$	Matricizing	$\mathcal{X}^{I_1\times I_2\times\ldots\times I_N} \rightarrow \mathbf{X}_{(n)}^{I_n\times I_1\cdot I_2\cdots I_{n-1}\cdot I_{n+1}\cdots I_N}$
$\times_n$ or $\bullet_n$	*n*-mode product	$\mathcal{X}\times_n\mathbf{M} = \mathcal{Z}$ where $\mathbf{Z}_{(n)} = \mathbf{M}\mathbf{X}_{(n)}$
$\circ$	outer product	$\boldsymbol{a}\circ\boldsymbol{b} = \mathbf{Z}$ where $z_{i,j} = a_ib_j$
$\otimes$	Kronecker product	$\mathbf{A}\otimes\mathbf{B} = \mathbf{Z}$ where $z_{k+K(i-1),l+L(j-1)} = a_{ij}b_{kl}$
$\odot$ or $\|\otimes\|$	Khatri–Rao product	$\mathbf{A}\odot\mathbf{B} = \mathbf{Z},$ where $z_{k+K(i-1),j} = a_{ij}b_{kj}$.
k_A	*k*-rank	Maximal number of columns of $\mathbf{A}$ guaranteed to be linearly independent.

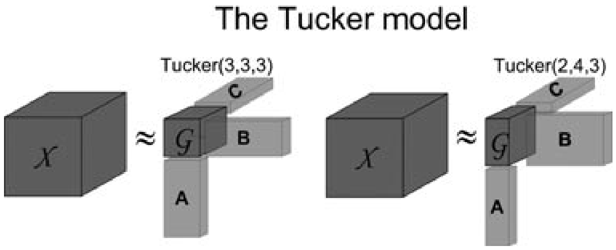

FIGURE 2 | Illustration of the Tucker model of a third-order tensor $\mathcal{X}$. The model decomposes the tensor into loading matrices with a mode specific number of components as well as a core array accounting for all multilinear interactions between the components of each mode. The Tucker model is particularly useful for compressing tensors into a reduced representation given by the smaller core array $\mathcal{G}$.

a third-order tensor but they trivially generalize to general Nth order arrays by introducing additional mode-specific loadings.

Tucker Model

The Tucker model proposed in Ref 29 reads for a third-order tensor $\mathcal{X}^{I\times J\times K}$

$$\mathcal{X}^{I\times J\times K} \approx \sum_{lmn} g_{l,m,n}\mathbf{a}_l^{\mathrm{I}} \circ \mathbf{b}_m^{\mathrm{J}} \circ \mathbf{c}_n^{\mathrm{K}}, \qquad \text{such that}$$

$$x_{i,j,k} \approx \sum_{lmn} g_{l,m,n} a_{i,l} b_{j,m} c_{k,n},$$

where the so-called core array $\mathcal{G}^{L\times M\times N}$ with elements $g_{l,m,n}$ accounts for all possible linear interactions between the components of each mode. To indicate how many vectors pertain to each modality, it is customary also to denote the model a Tucker(L, M, N) model. Using the n-mode tensor product $\times_n$,[29,32] the model can be written as

$$\mathcal{X}^{I\times J\times K} \approx \mathcal{G}^{L\times M\times N} \times_1 \mathbf{A}^{I\times L} \times_2 \mathbf{B}^{J\times M} \times_3 \mathbf{C}^{K\times N}.$$

Each mode of the array is approximately spanned by given loading matrices for that mode such that the vectors of each modality interact with the vectors of all remaining modalities with strengths given by the core tensor $\mathcal{G}$, see, also Figure 2.

The Tucker model is not unique. As such, multiplying by invertible matrices $\mathbf{Q}^{L\times L}$, $\mathbf{R}^{M\times M}$, and $\mathbf{S}^{N\times N}$ gives an equivalent representation, i.e.,

$$\begin{aligned}\mathcal{X} \approx (\mathcal{G} \times_1 \mathbf{Q} \times_2 \mathbf{R} \times_3 \mathbf{S}) \times_1 (\mathbf{A}\mathbf{Q}^{-1}) \times_2 (\mathbf{B}\mathbf{R}^{-1}))\\ \times_3(\mathbf{C}\mathbf{S}^{-1})) = \widetilde{\mathcal{G}} \times_1 \widetilde{\mathbf{A}} \times_2 \widetilde{\mathbf{B}} \times_3 \widetilde{\mathbf{C}}.\end{aligned}$$

As a result, the factors of the unconstrained Tucker model can be constrained orthogonal or orthonormal (which is useful for compression) without hampering the reconstruction error. However, imposing orthogonality/orthonormalty does not resolve the lack of uniqueness as the solution is still ambiguous to multiplication by orthogonal/orthonormal matrices $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{S}$. Using the n-mode matricizing and Kronecker product operation, the Tucker model can be written as

$$\begin{aligned}\mathbf{X}_{(1)} &\approx \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})^\top\\ \mathbf{X}_{(2)} &\approx \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C}\otimes\mathbf{A})^\top\\ \mathbf{X}_{(3)} &\approx \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B}\otimes\mathbf{A})^\top.\end{aligned}$$

The above decomposition for a third-order tensor is also denoted a Tucker3 model, the Tucker2 model and Tucker1 models are given by

$$\begin{aligned}\text{Tucker2:}\quad & \mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{I},\\ \text{Tucker1:}\quad & \mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{I} \times_3 \mathbf{I},\end{aligned}$$

where $\mathbf{I}$ is the identity matrix. Thus, the Tucker1 model is equivalent to regular matrix decomposition based on the representation $\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}$.

Model Estimation

Traditionally, the Tucker model has been estimated on the basis of updating the elements of each mode in turn that for the least squares objective commonly is denoted ALS. By fitting the model using ALS, the estimation reduces to a sequence of regular matrix factorization problems. As a result, for least squares minimization, the solution of each mode can be solved by pseudoinverses, i.e.,

$$\begin{aligned}\mathbf{A} &\leftarrow \mathbf{X}_{(1)}(\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})^\top)^\dagger\\ \mathbf{B} &\leftarrow \mathbf{X}_{(2)}(\mathbf{G}_{(2)}(\mathbf{C}\otimes\mathbf{A})^\top)^\dagger\\ \mathbf{C} &\leftarrow \mathbf{X}_{(3)}(\mathbf{G}_{(3)}(\mathbf{B}\otimes\mathbf{A})^\top)^\dagger\\ \mathcal{G} &\leftarrow \mathcal{X} \times_1 \mathbf{A}^\dagger \times_2 \mathbf{B}^\dagger \times_3 \mathbf{C}^\dagger.\end{aligned}$$

The analysis simplifies when orthogonality is imposed[24] such that the estimation of the core can be omitted. Orthogonality can be imposed by estimating the loadings of each mode through the SVD forming the Higher-order Orthogonal Iteration (HOOI),[10,24] i.e.,

$$\begin{aligned}\mathbf{A}\mathbf{S}^{(1)}\mathbf{V}^{(1)^\top} &= \mathbf{X}_{(1)}(\mathbf{C}\otimes\mathbf{B}),\\ \mathbf{B}\mathbf{S}^{(2)}\mathbf{V}^{(2)^\top} &= \mathbf{X}_{(2)}(\mathbf{C}\otimes\mathbf{A}),\\ \mathbf{C}\mathbf{S}^{(3)}\mathbf{V}^{(3)^\top} &= \mathbf{X}_{(3)}(\mathbf{B}\otimes\mathbf{A}).\end{aligned}$$

such that $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are found as the first L, M, and N left singular vectors given by solving the right hand side by SVD. The core array is estimated upon convergence by $\mathcal{G} \leftarrow \mathcal{X} \times_1 \mathbf{A}^\dagger \times_2 \mathbf{B}^\dagger \times_3 \mathbf{C}^\dagger$. The above procedures are unfortunately not guaranteed to converge to the global optimum.

A special case of the Tucker model is given by the HOSVD[29,32] where the loadings of each mode is

TABLE 2 | Overview of the Most Common Tensor Decomposition Models, Details of the Models as well as References to Their Literature can be Found in Refs 24, 28, and 44

Model name	Decomposition	Unique
CP	$x_{i,j,k} \approx \sum_d a_{i,d} b_{j,d} c_{k,d}$	Yes
The minimal D for which approximation is exact is called the rank of a tensor, model in general unique.		
Tucker	$x_{i,j,k} \approx \sum_{l,m,n} g_{l,m,n} a_{i,l} b_{j,m} c_{k,n}$	No
The minimal L, M, N for which approximation is exact is called the multilinear rank of a tensor.		
Tucker2	$x_{i,j,k} \approx \sum_{lm} g_{l,m,k} a_{i,l} b_{j,m}$	No
Tucker model with identity loading matrix along one of the modes.		
Tucker1	$x_{i,j,k} \approx \sum_{l,m,n} g_{l,j,k} a_{i,l}$	No
Tucker model with identity loading matrices along two of the modes. The model is equivalent to regular matrix decomposition.		
PARAFAC2	$x_{i,j,k} \approx \sum_d^D a_{i,d} b^{(k)}_{j,d} c_{k,d}$, s.t. $\sum_j b^{(k)}_{j,d} b^{(k)}_{j,d'} = \psi_{d,d'}$	Yes
Imposes consistency in the covariance structure of one of the modes. The model is well suited to account for shape changes; furthermore, the second mode can potentially vary in dimensionality.		
INDSCAL	$x_{i,j,k} \approx \sum_d a_{i,d} a_{j,d} c_{k,d}$	Yes
Imposing symmetry on two modes of the CP model.		
Symmetric CP	$x_{i,j,k} \approx \sum_d a_{i,d} a_{j,d} a_{k,d}$	Yes
Imposing symmetry on all modes in the CP model useful in the analysis of higher order statistics.		
CANDELINC	$x_{i,j,k} \approx \sum_{lmn} (\sum_d \hat{a}_{l,d} \hat{b}_{m,d} \hat{c}_{n,d}) a_{i,l} b_{j,m} c_{k,n}$	No
CP with linear constraints can be considered a Tucker decomposition where the Tucker core has CP structure.		
DEDICOM	$x_{i,j,k} \approx \sum_{d,d'} a_{i,d} b_{k,d} r_{d,d'} b_{k,d'} a_{j,d'}$	Yes
Can capture asymmetric relationships between two modes that index the same type of object.		
PARATUCK2	$x_{i,j,k} \approx \sum_{d,e} a_{i,d} b_{k,d} r_{d,e} s_{k,e} t_{j,e}$	Yes[55]
A generalization of DEDICOM that can consider interactions between two possible different sets of objects.		
Block Term Decomp.	$x_{i,j,k} \approx \sum_r \sum_{lmn} g^{(r)}_{lmn} a^{(r)}_{i,n} b^{(r)}_{j,m} c^{(r)}_{k,n}$	Yes[56]
A sum over R Tucker models of varying sizes where the CP and Tucker models are natural special cases.		
ShiftCP	$x_{i,j,k} \approx \sum_d a_{i,d} b_{j-\tau_{i,d},d} c_{k,d}$	Yes[6]
Can model latency changes across one of the modes.		
ConvCP	$x_{i,j,k} \approx \sum_\tau^T \sum_d a_{i,d,\tau} b_{j-\tau,d} c_{k,d}$	Yes
Can model shape and latency changes across one of the modes. When $T = J$ the model can be reduced to regular matrix factorization; therefore, uniqueness is dependent on T.		

determined solely by the SVD of the matricized array,

$$
\begin{aligned}
\boldsymbol{A}\boldsymbol{S}^{(1)}\boldsymbol{V}^{(1)^\top} &= \boldsymbol{X}_{(1)},\\
\boldsymbol{B}\boldsymbol{S}^{(2)}\boldsymbol{V}^{(2)^\top} &= \boldsymbol{X}_{(2)},\\
\boldsymbol{C}\boldsymbol{S}^{(3)}\boldsymbol{V}^{(3)^\top} &= \boldsymbol{X}_{(3)}.
\end{aligned}
$$

Although this approach strikingly resembles the SVD,[32] it is not guaranteed to find an optimal compression. In particular, the approach does not take the (Tucker) structure of the remaining modes into account when solving for the loadings of a given mode, see also Ref 10. Therefore, the HOSVD is commonly used as an initialization method that is refined by other Tucker estimation approaches. If the matricized ranks in the three modes are found to be L, M, and N respectively, then a Tucker(L, M, N) model fits the data perfect.

CP Model

The CP model independently proposed in Refs 30, 31, 33 can be considered a special case of the Tucker model where the size of each modality of the core array $\mathcal{G}$ is the same, i.e., $L = M = N$, whereas interactions are only between columns of same indices such that the only nonzero elements are found along the diagonal of the core, i.e., $g_{l,m,n} \neq 0$ if and only if $l = m = n$, see also Figure 3. As a result, the CP model can be written as

$$
\mathcal{X}^{I\times J\times K} \approx \mathcal{D}^{D\times D\times D} \times_1 \boldsymbol{A}^{I\times D} \times_2 \boldsymbol{B}^{J\times D} \times_3 \boldsymbol{C}^{K\times D},
$$

where $\mathcal{D}$ is a diagonal tensor. An important property of the CP model is that the restriction imposed on the Tucker core leads to uniqueness of the representation. When multiplying by invertible matrices $\boldsymbol{Q}^{D\times D}$,

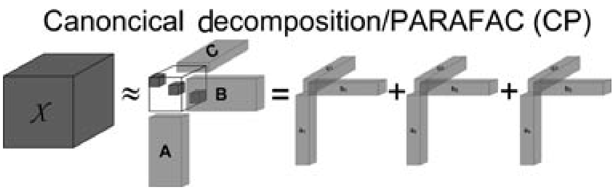

FIGURE 3 | Illustration of the CANDECOMP/PARAFAC (CP) model of a third-order tensor $\mathcal{X}$. The model decomposes a tensor into a sum of rank one components and the model is very appealing due to its uniqueness properties.

$\boldsymbol{R}^{D\times D}$, and $\boldsymbol{S}^{D\times D}$, we find

$$\mathcal{X} \approx (\mathcal{D} \times_1 \boldsymbol{Q} \times_2 \boldsymbol{R} \times_3 \boldsymbol{S}) \times_1 (\boldsymbol{A}\boldsymbol{Q}^{-1}) \times_2 (\boldsymbol{B}\boldsymbol{R}^{-1}) \\ \times_3 (\boldsymbol{C}\boldsymbol{S}^{-1}) = \widetilde{\mathcal{D}} \times_1 \widetilde{\boldsymbol{A}} \times_2 \widetilde{\boldsymbol{B}} \times_3 \widetilde{\boldsymbol{C}}.$$

As such, the new core $\widetilde{\mathcal{D}}$ must be nonzero only along the diagonal for the representation to remain a CP model. This, in practice, has the consequence that $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{S}$ can only be scale and permutation matrices with identical permutation. In Refs 34, 35, the uniqueness properties of the CP model were thoroughly investigated and among several results, the following uniqueness criterion derived

$$k_A + k_B + k_C \geq 2D + 2. \tag{13}$$

Here, the Kruskal rank or k-rank k_A of a matrix $\boldsymbol{A}$ is the maximal number r such that any set of r columns of the matrix $\boldsymbol{A}$ is linearly independent; therefore, $k_A \leq rank(\boldsymbol{A}) \leq D$ where D is the number of components.

The notion of k-rank is closely related to the notion of spark in compressed sensing[36] and while k-rank is NP hard to compute it can be bounded by measures of coherence.[36] In the presence of noise with continuous probability distribution and when all the dimensions of the tensor are larger than D, we have in practice $k_A = k_B = k_C = D$. As Kruskal wrote in Ref 34, struck by his own uniqueness criterion,

> *'A surprising fact is that the nonrotatability characteristic can hold even when the number of factors extracted is greater than every dimension of the three-way array.'*

The criterion has been generalized to order N arrays in Ref 37.

The uniqueness property of the optimal CP solution is perhaps the most appealing aspect of the CP model. Uniqueness of matrix decomposition has been a longstanding challenge that has spurred a great deal of research early on in the psychometrics literature where rotational approaches such as VARIMAX were proposed (see, also Refs 1, 2, 31, and references therein) and lately in the signal processing literature where methods based on statistical independence[9] (see, also, '*Signal Processing*') have been derived in order to disambiguate matrix decompositions. Thus, contrary to the regular matrix factorization approaches, the CP model admit a unique representation of the data.

Unfortunately, uniqueness sometimes comes at a price as CP degenerate solutions are known to occur, i.e., solutions in which the component loadings are highly correlated in all the modes and where the solution is not physically meaningful but rather a mathematical artifact caused by the inability of CP to model that particular tensor with that particular number of components. This makes the CP estimation unstable, slow in convergence, and difficult to interpret because the components are dominated by strong cancelations effects between the various components in the model.[38,39] An example of CP degeneracy is given in Figure 6.

Model Estimation

As with the Tucker model, the CP decomposition does not admit any known closed form solution, and in general, there is no guarantee that the optimal solution, even if it exists, can be identified. Although a direct fitting approach based on a generalized eigenvalue problem with fixed computational complexity can be imposed, the optimization criterion is not strictly well defined in terms of the least squares objective, whereas the solutions obtained have been found to be inferior to the following ALS approach.[27]

Parameter estimation by ALS is widely used because of its ease of implementation by the use of the Khatri–Rao product and the matricizing operations. For a third-order CP model, we can disambiguat the scaling ambiguity between the diagonal elements of the core and the loadings by fixing the diagonal core elements to 1 such that the CP model can be written as

$$\mathcal{X}^{I\times J\times K} \approx \sum_d \boldsymbol{a}_d^{\mathrm{I}} \circ \boldsymbol{b}_d^{\mathrm{J}} \circ \boldsymbol{c}_d^{\mathrm{K}}, \quad \text{such that}$$

$$x_{i,j,k} \approx \sum_d a_{i,d} b_{j,d} c_{k,d}.$$

Using the matricizing and Khatri–Rao product, this is equivalent to

$$\boldsymbol{X}_{(1)} \approx \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^\top,$$
$$\boldsymbol{X}_{(2)} \approx \boldsymbol{B}(\boldsymbol{C} \odot \boldsymbol{A})^\top,$$
$$\boldsymbol{X}_{(3)} \approx \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top.$$

For the least squares objective we, thus, find

$$A \leftarrow \mathbf{X}_{(1)}(C \odot B)(C^{\top}C * B^{\top}B)^{-1}$$
$$B \leftarrow \mathbf{X}_{(2)}(C \odot A)(C^{\top}C * A^{\top}A)^{-1}$$
$$C \leftarrow \mathbf{X}_{(3)}(B \odot A)(B^{\top}B * A^{\top}A)^{-1}$$

However, some calculations are redundant between the alternating steps. Thus, the following approach based on premultiplying the largest mode(s) with the data is more computationally efficient.[27] Multiplying the first mode with the data when updating for the second and third mode of a third-order array gives

$$A \leftarrow \mathbf{X}_{(1)}(C \odot B)(C^{\top}C * B^{\top}B)^{-1}, \quad \widehat{\mathbf{X}}_{(1)} = A^{\top}\mathbf{X}_{(1)}$$
$$B \leftarrow \widehat{\mathbf{X}}_{(2)}(C \odot I)(C^{\top}C * A^{\top}A)^{-1}$$
$$C \leftarrow \widehat{\mathbf{X}}_{(3)}(B \odot I)(B^{\top}B * A^{\top}A)^{-1}.$$

We will, without loss of generality, assume $I \geq J \geq K$, hence, the above approach reduces the cost invoked for the above updates of B and C from $O(IJKD)$ to $O(max\{JD^2, D^3\})$ when taking advantage of the sparsity structure of the Khatri–Rao products where D is the number of components in the CP model. Commonly, the above ALS algorithm is run multiple times with different initializations to avoid the identification of a local minima solution.

Although the ALS algorithm for CP and Tucker estimation are widely used, ALS can suffer from slow convergence. Although ALS converges at most linearly in practice, it can be extremely slow particularly in cases of high collinearity between the factors.[27] Alternative estimation approaches such as Levenberg–Marquardt, conjugate gradient, and enhanced line search have been shown to improve convergence.[26] For further details on CP and Tucker model estimation and alternatives to ALS optimization including complexity analysis and performance evaluation, see also Refs 24, 26, 27, and references therein.

Rank and Multilinear Rank of a Tensor

The rank of a tensor is given by its minimal sum of rank one components R such that

$$\mathcal{X} = \sum_{r=1}^{R} a_r \circ b_r \circ c_r. \tag{14}$$

Notice, contrary to the matrix case, the rank of a tensor can be greater than $\min(I, J, K)$. Furthermore, for tensor decomposition over the real field a $2 \times 2 \times 2$ tensor can both be rank 2 and rank 3 but in the complex field $2 \times 2 \times 2$ tensors generically have rank 2.[26] The fact that the typical rank of a tensor can take more than one value is specific to the real field. One major difference between matrix and tensor rank is that there is no straightforward algorithm to determine the rank of a tensor. In practice, the rank of a tensor is determined numerically by fitting various CP models for different R.[24] Using the Tucker model representation, a third-order tensor is said to have multilinear rank-(L, M, N) if its mode-1 rank, mode-2 rank, and mode-3 rank are equal to L, M, and N, respectively[10,32,39]

$$\mathcal{X} = \sum_{lmn}^{LMN} g_{l,m,n} a_l \circ b_m \circ c_n. \tag{15}$$

Although the Tucker model, due to its orthogonal representation, is useful for projection onto tensorial subspaces (i.e., compression), the CP model by definition is outer product rank revealing and often of interest because of its unique and easily interpreted representations.

Missing Values

Missing data can arise in a variety of settings because of loss of information, errors in the data collection process, or costly experiments.[40] In the case of missing data, a standard practice is to impute missing data values, i.e., the missing data element $x_{i,j,k}$ is replaced by the estimated value $r_{i,j,k}$ of that element from the decomposition model, i.e., $x_{i,j,k} = r_{i,j,k}$ starting from an initial guess of these missing values (this is also referred to as expectation maximization). An alternative approach is to use marginalization where the missing values are ignored (i.e., marginalized) during optimization, i.e., for least squares estimation by considering the objective $\sum_{i,j,k} w_{i,j,k}(x_{i,j,k} - r_{i,j,k})^2$, where $w_{i,j,k} = 1$ if $x_{i,j,k}$ is present and $w_{i,j,k} = 0$ if $x_{i,j,k}$ is missing. Although methods based on imputation often are easier to implement (i.e., alternating least squares can be directly applied), they are useful only as long as the amount of missing data is relatively small as their performance tend to degrade for large amounts of missing data as the intermediate models used for imputation have increased risk of convergence to a wrong solution (see also Refs 27, 40, and references therein). Factorizing tensors based on the CP model have been shown to recover the true underlying components from noisy data with up to 99% data missing for third-order tensors,[40] whereas the corresponding two-way methods become rather unstable already with 25–40% of data missing.[27,40] This has been attributed to the fact that there are fewer free parameters p relative to observations for models accounting for multilinear dynamics, i.e., the CP model has $p = D(I + J + K)$, the Tucker model has $p = IL + JM + KN$ (when considering the core a deterministic function of the loadings), whereas the corresponding two-way analysis requires $p = D(I + JK)$ fitted parameters for a third-order tensor $\mathcal{X}^{I\times J\times K}$.

Model Order Estimation

Determining the number of components for the CP and Tucker model is challenging. Contrary to the SVD decomposition, the CP and Tucker models are not in general nested, hence, the extracted features change with the number of components extracted. As such, determining the model order is important to interpret the CP and Tucker decompositions in a viable way. Model order estimation is particularly a challenge for the Tucker model as the number of components is specified for each mode separately resulting in a large combinatorial explosion in the number of potential models, i.e., for CP up to $D^{\max}$ models have to be evaluated, whereas for the third-order Tucker model $L^{\max}M^{\max}N^{\max}$ potential models have to be considered. A commonly used approach is to evaluate the number of components for all potential models in terms of their ability to account for the data relative to the number of parameters used in the model, see also Ref 41 and references therein.

For the CP model, a widely used heuristic for evaluating the number of components is based on the so-called core consistency diagnostic (CORCONDIAG) proposed in Ref 42

$$\mathcal{G} \leftarrow \mathcal{X} \times_1 \boldsymbol{A}_{\text{CP}}^{\dagger} \times_2 \boldsymbol{B}_{\text{CP}}^{\dagger} \times_3 \boldsymbol{C}_{\text{CP}}^{\dagger}, \tag{16}$$

$$\text{CORCONDIAG} = 100 \times \left(1 - \frac{\|\mathcal{I} - \mathcal{G}\|_F^2}{\|\mathcal{I}\|_F^2}\right). \tag{17}$$

Where $\mathcal{I}$ is the (diagonal) CP core array and $\mathcal{G}$ the corresponding Tucker core array obtained from the loadings $\boldsymbol{A}_{\text{CP}}$, $\boldsymbol{B}_{\text{CP}}$, and $\boldsymbol{C}_{\text{CP}}$ extracted from the CP model. Too many components will result in a strong degree of cross-talk across the loadings of the modes and will yield a low value of the CORCONDIAG. Too few components will not have any cross-talk at all. Thus, the 'correct' number of components is taken to be just before a major drop-off in the curve of $(d, \text{CORCONDIAG})$. As explained in Ref 42 *'As a rule of thumb, a core consistency above 90% can be interpreted as "very trilinear", whereas a core consistency in the neighborhood of 50% would mean a problematic model with signs of both trilinear variation and variation which is not trilinear'. A core consistency close to zero or even negative implies an invalid model, because the space covered by the component matrices is then not primarily describing trilinear variation.'*

An alternative procedure for model order estimation uses missing value estimation in conjunction with crossvalidation, see also Ref 43 and references therein. Recently, a hierarchical Bayesian approach based on automatic relevance determination has been proposed to estimate the model order of the CP and Tucker model without having to exhaustively evaluate all potential model orders.[41] Here, priors on the model parameters are given hyperparameters that represent the scale of each component by defining their range of variation. By optimizing these hyperparameters, components can be removed if their scale goes below some threshold. This results in an estimate of the model order when the model is initialized with 'too many' components. Furthermore, sparsity imposed on the core can be used to interpolate between a CP and Tucker representation (see, also Ref 41 and references therein).

Common Constraints

When optimizing the CP and Tucker model on the basis of alternating least squares, the estimation of the parameters of each mode form a regular matrix decomposition problem. Constraints such as orthogonality and nonnegativity from matrix decomposition analysis can be directly imposed on the components to further improve their identification. A benefit for the CP decomposition of imposing such constraints is that degeneracy no longer can occur. A detailed account of nonnegative tensor factorization can be found in Ref 25.

Other Tensor Factorization Methods

A multitude of other tensor factorization models beyond the CP and Tucker models have been proposed over the years. Ranging from models exploiting various kinds of symmetry in the tensors (INDSCAL, Symmetric CP, and DEDICOM) to generalizations of the CP model to account for latency and shape changes (shiftCP, convCP, and PARAFAC2) to decompositions that can be considered models interpolating between or combining the Tucker and CP models (Block Term Decompositions and CANDELINC). For an overview of these approaches, see also Table 2 as well as Refs 24, 28, 44.

Some Available Tensor Software

Several software packages are available online for tensor decomposition. The *n*-way toolbox[45] is a great starting point providing Matlab algorithms for model estimation of the CP and Tucker model including decomposition under nonnegativity and orthogonality constraints. Fast prototyping and handling of sparse multiway arrays in Matlab is provided by the TensorToolbox.[46] For additional software, see also Refs 4, 24.

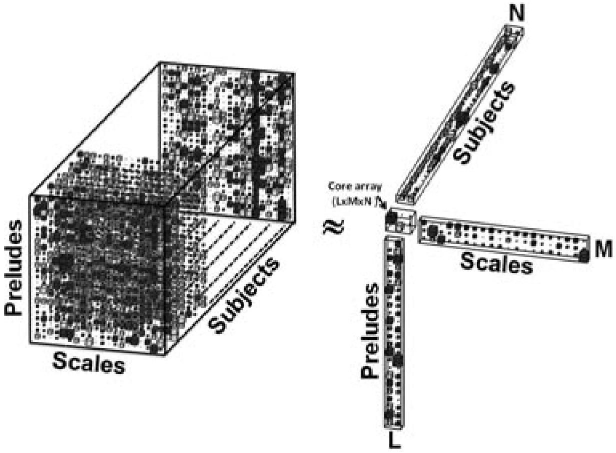

FIGURE 4 | Example of a Tucker(2, 3, 2) analysis of the chopin data $\mathcal{X}^{24\ \text{Preludes}\times 20\ \text{Scales}\times 38\ \text{Subjects}}$ described in Ref 49. The overall mean of the data has been subtracted prior to analysis. Black and white boxes indicate negative and positive variables, whereas the size of the boxes their absolute value. The model accounts for 40.42% of the variation in the data, whereas the model on the same data random permuted accounts for 2.41 ± 0.09% of the variation. As such, the data are very structured and compressible by the Tucker model.

TENSOR FACTORIZATION FOR DATA MINING

The first applications of Tensor decomposition was within the field of Psychology in the 1970s when the CP model was demonstrated to alleviate the rotational ambiguity in factor analysis, whereas the framework enabled to address higher order interactions. In 1981 Appellof and Davidson[3] pioneered the use of the CP model in chemistry for the analysis of fluorescence data, whereas Möcks[47] demonstrated in 1988 how the CP model was useful in the analysis of multisubject-evoked potentials of electroencephalography (EEG) data by reinventing the model under the name topographic component model. Since then tensor decompositions have found wide use in practically all fields of science ranging from signal processing, computer vision, bioinformatics to web mining. In many of these studies, it has been proven that the use of tensor decomposition can explore relations and interactions between the modes of the data that are lost when resorting to traditional matrix approaches. In particular, tensor decomposition efficiently extract the consistent patterns of activation while giving an intuitive account of how the measurements of each mode interact. However, tensor decomposition has not only proven useful for redundancy reduction (i.e., compression) but also for many types of data proven to account well for the underlying physics/dynamics of the system generating the data. In the following, some of the key applications of tensor factorization in data mining is given across a multitude of scientific fields given more or less in their historical order. This is in no way an exhaustive account of the many applications of tensor decomposition; however, the examples given will hopefully demonstrate some of the many benefits of multiway modeling for a variety of data and problem domains.

Psychology

The first applications of CP was within the field of psychometrics in 1970 pioneered by the work of Carroll and Chang[30] and Harshman.[31] Ref 30 introduced Canonical Decomposition in the context of analyzing multiple similarity or dissimilarity matrices from a variety of subjects. They applied the method to one data set on auditory tones from Bell Labs and to another data set of comparisons of countries based on the idea that simply averaging the data removed the different aspects present in the data,[31] introduced PARAFAC because it eliminated the rotational ambiguity associated with two-dimensional PCA and thus has better uniqueness properties motivated by Cattells principle of parallel proportional profiles.[48] PARAFAC was here applied to vowel-sound data where different individuals spoke different vowels and the formant (i.e., the pitch) was measured, i.e.,

$$\mathcal{X}^{\text{Subject}\times\text{Vowel}\times\text{Pitch}} \approx \sum_d \boldsymbol{a}_d^{\text{Subject}} \circ \boldsymbol{b}_d^{\text{Vowel}} \circ \boldsymbol{c}_d^{\text{Pitch}}. \qquad (18)$$

Since these initial works both the CP as well as Tucker model also referred to as N-mode PCA[2] have had a widespread application within social and behavioral sciences addressing questions such as '*Which group of subjects behave differently on which variables under which conditions?*'[2] In Figure 4 is given a Tucker$(2, 3, 2)$ analysis of 24 chopin preludes based on 20 types of scoring scales evaluated by 38 judges/subjects,[49] i.e.,

$$\mathcal{X}^{\text{Predude}\times\text{Scales}\times\text{Subject}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}_l^{\text{Prelude}} \circ \boldsymbol{b}_m^{\text{Scales}} \circ \boldsymbol{c}_n^{\text{Subject}}. \qquad (19)$$

The analysis extracts loadings that well span the dynamics of each mode, whereas the core array accounts

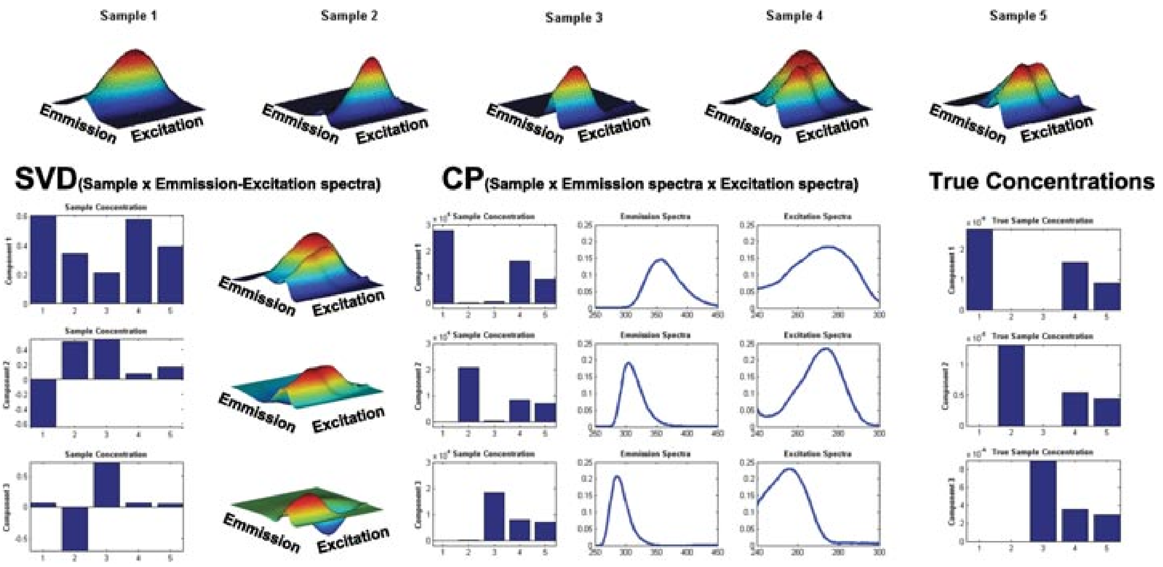

FIGURE 5 | Example of SVD (bottom left) and CP analysis (bottom middle) of the *claus.mat* fluorescence data given at the top provided by the *N*-way toolbox.[45] Both the three component SVD and CP model accounts for more than 99.9% of the variation in the data. However, the CP decomposition admits a unique account of the data, resulting in the identification of the true underlying chemical compounds and their relative concentrations (bottom right).

for all linear interactions between the components of each mode. The decomposition account for 40.42% of the variation in the data, whereas an equivalent analysis of the corresponding randomly permuted data having no correspondence between the entries in the tensor only accounts for 2.41% of the variation. Thus, the data are very structured and compressible by the Tucker model. Great introductions to multiway modeling within psychology can be found in Refs 1, 2.

Chemistry

Appellof and Davidson[3] pioneered the use of CP in chemometrics in 1981 for the use in analysis of fluorescence spectroscopy. As stated in Beer–Lambert's law, there is a linear relation between absorbance of light and the concentration of a compound. Measuring samples of mixed compounds such that the concentration of each compound vary across the samples admit unique recovery of the spectral profiles of the compounds according to the CP model

$$\mathcal{X}^{\text{Exication}\times\text{Emission}\times\text{Samples}} \approx \sum_{d=1}^{D^{\#Compounds}} \boldsymbol{a}_d^{\text{Excitation}} \circ \boldsymbol{b}_d^{\text{Emmision}} \circ \boldsymbol{c}_d^{\text{Samples}}. \quad (20)$$

One of the most interesting aspects of the application of the CP model for fluorescence spectroscopy is that it enables the so-called second-order advantage, making it possible to do quantitative chemical analytes even in the presence of un-calibrated interferents.[4] This is not possible with traditional regression-based calibration. Apart from fluorescence, multiway decomposition is widely used in chromatography, flow injection analysis, and nuclear magnetic resonance (NMR) as well as in the analysis of environmental data.[4] An illustration of CP analysis of fluorescence data is given in Figure 5. A good introduction to multiway analysis in chemistry can be found in Ref 28, whereas an extensive review is given in Ref 4.

Neuroscience

When Harshman proposed the PARAFAC model in Ref 31, one of his suggested applications was to use the model in the analysis of EEG data. However, it wasn't until the reinvention of the CP model by Möcks,[47] naming it the topographic components model that the model was used in the analysis of event-related EEG data. Since the work of Möcks[47] multiway analysis has been used in the analysis of time-frequency-transformed EEG[7,8,25,50,51] data as well as functional magnetic resonance imaging (fMRI), see also Refs 5, 6 and references therein. Neuroscience data are naturally multimodal. In event-related designs, spatial activation is measured over time and trials forming a three-way array of $\mathcal{X}^{\text{Space}\times\text{Time}\times\text{Trials}}$ and often these measurements are also recorded across multiple subjects and conditions, which naturally form even higher order arrays. The consistent patterns of activation of a five-way

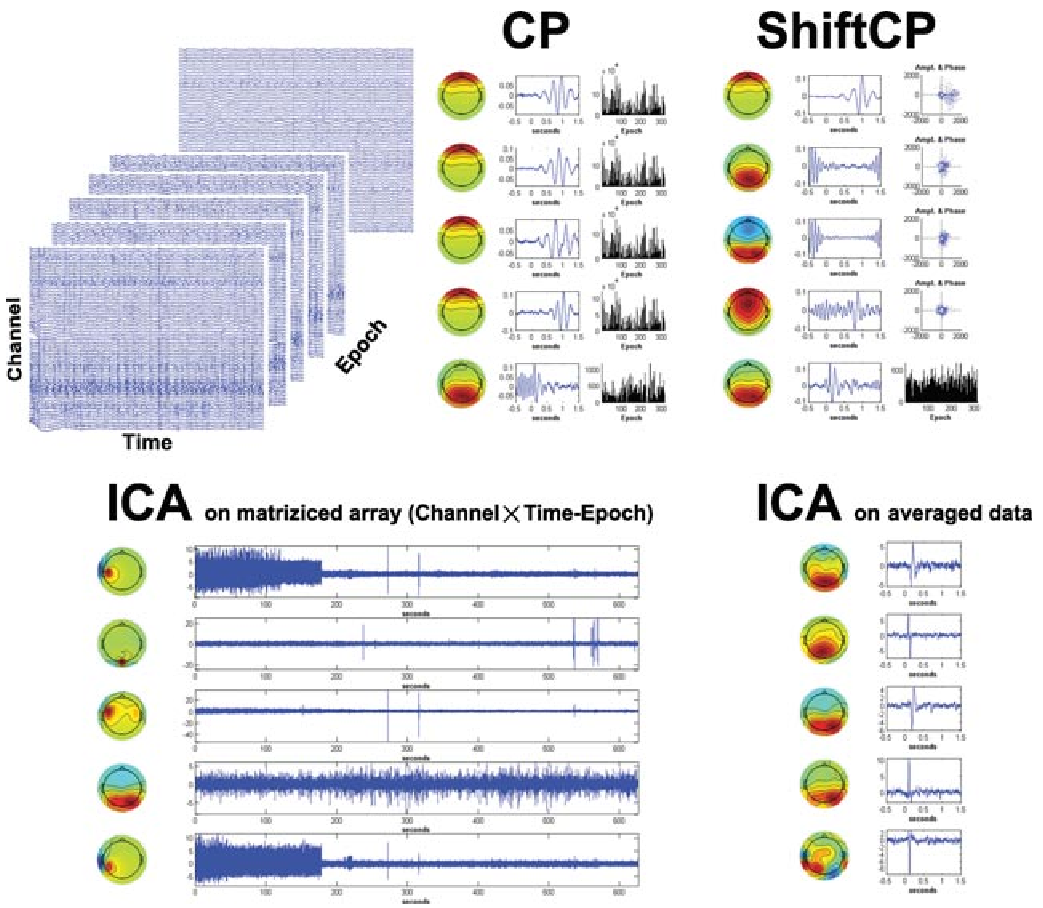

FIGURE 6 | Example of CP analysis and shiftCP analysis of Electroencephalography (EEG) data of $\mathcal{X}^{64\ \text{Channels}\times 1024\ \text{Time points}\times 313\ \text{Epochs/Trials}}$ described in Ref 6. Because of violation of trilinearity, a degenerate solution is extracted by the CP model given by the first four highly correlated components that account for the dynamics of the data through a strong degree of between component cancelation. However, when accounting for latency changes across the trials in four out of the five components, degeneracy no longer occur, whereas the most consistent spatial and temporal patterns of activation across the trials are successfully extracted (the amplitude and phase plot account for the trial-specific strength and delay of the various components). The corresponding two-way analysis here given by ICA in order to resolve the rotational ambiguity of two-way decomposition based on the fastICA algorithm (*http://www.cis.hut.fi/projects/ica/fastica/* using the non-linear function *tanh*(·)) no longer assume consistency across the trials. As a result, the matrix decomposition of *channel* × *time* − *epoch* is mainly driven by noisy artifacts, whereas the analysis of the trial averaged data also denoted the evoked potential (EP) to the bottom right mainly focus on accounting for the dynamics of the *P*100 − *N*200 − *P*300 complex of the EP. As a result, multilinear modeling enable direct extraction of the most consistent reproducible patterns across the trials.

array of wavelet-transformed EEG data given by $\mathcal{X}^{\text{Channel}\times\text{Frequency}\times\text{Time}\times\text{Subject}\times\text{Conditions}}$ was analyzed in Ref 8 based on nonnegative tensor factorization,[25,51] for an example of such data set, see also Figure 7 that includes the following three-way CP decomposition of the tutorial dataset 2 provided in Ref 50.

$$\mathcal{X}^{\text{Channel}\times\text{Time}-\text{Freq.}\times\text{Subj.}-\text{Cond.}} \approx \sum_{d=1}^{D} \boldsymbol{a}_d^{\text{Channel}} \circ \boldsymbol{b}_d^{\text{Time}-\text{Freq.}} \circ \boldsymbol{c}_d^{\text{Subj.}-\text{Cond.}}. \quad (21)$$

Multiway decomposition naturally admit the analysis of the consistent patterns and thereby also the most reproducible patterns of activation in the data. A consistency that is not imposed when analyzing the data by regular matrix analysis of the corresponding matricized array, see also Figures 6 and 7. In neuro-imaging data, there has been a tradition of averaging over repeats/trials in order to identify the event-related potential. However, this averaging assumes that the pattern of activation is equally present over the trials while the spatial correlation of activation is not taken into account. The benefit of multilinear modeling of this type of data by the CP model is that the activation is grouped spatially, whereas trial-dependent strength admits a weighted average over the trials such that noisy trials can be

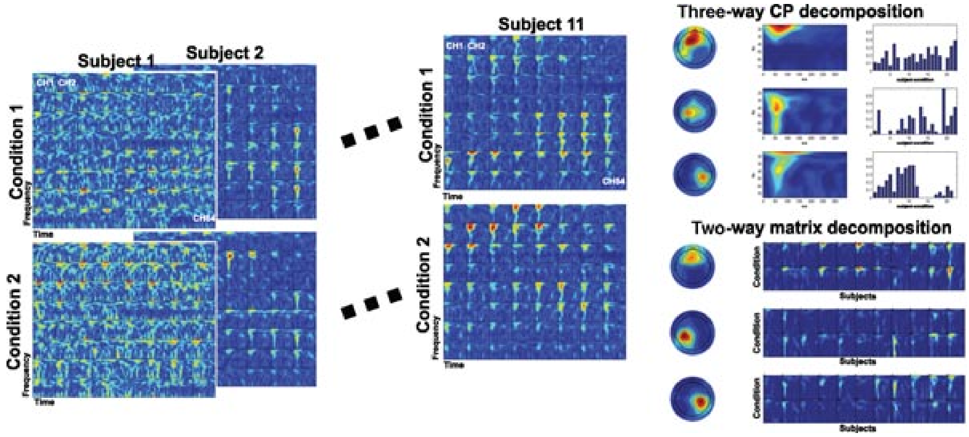

FIGURE 7 | Left panel: Tutorial dataset two of ERPWAVELAB[50] given by $\mathcal{X}^{64\text{ Channels}\times 61\text{ Frequency bins}\times 72\text{ Time points}\times 11\text{Subjects}\times 2\text{Conditions}}$. Right panel a three component nonnegativity constrained three-way CP decomposition of *Channel* × *Time* − *Frequency* × *Subject* − *Condition* and a three component nonnegative matrix factorization of *Channel* × *Time* − *Frequency* − *Subject* − *Condition*. The two models account for 60% and 76% of the variation in the data, respectively. The matrix factorization assume spatial consistency but individual time-frequency patterns of activation across the subjects and conditions, whereas the three-way CP analysis impose consistency in the time-frequency patterns across the subjects and conditions. As such, these most consistent patterns of activations are identified by the model.

down-weighted in the extracted estimates of the consistent event-related activations.

$$\mathcal{X}^{\text{Channel}\times\text{Time}\times\text{Trial}} \approx \sum_{d=1}^{D} \boldsymbol{a}_d^{\text{Channel}} \circ \boldsymbol{b}_d^{\text{Time}} \circ \boldsymbol{c}_d^{\text{Trial}} \quad (22)$$

Unfortunately, violation of multilinearity in the data can cause degeneracy in the CP model, see also Figure 6. To avoid CP degeneracy, artificial restrictions in the form of orthogonality have been imposed or alternatively the signals have been analyzed via purely additive models based on analysis of amplitudes in a spectral representation, see also Ref 6 and references therein. In Ref 6, these ad-hoc measures were found unsatisfactory. Rather than restricting the CP model, a *pseudo-multilinear* model using the unambiguous CP model combined with a time-shift accounting for explicit delays based on the shiftCP representation was proposed. In Figure 6, it can be seen that accounting for shift can indeed alleviate CP degeneracy while the consistent pattern of activations are identified, for details on the shiftCP approach, see also Ref 6 and references therein.

Signal Processing

Multilinear algebra has recently gained a large interest within the signal processing community largely due to its applications in higher-order statistics (HOS).[9–11,52] In the original work on independent component analysis (ICA) by Comon,[9] it was demonstrated how the blind source separation problem

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \quad (23)$$

such that $\mathbf{S}$ is statistically independent and $\mathbf{E}$ residual noise can be solved through the CP decomposition of some higher-order cumulants due to the important property that cumulants obey multilinearity.[9,52] The first-order cumulant corresponds to the mean and the second-order cumulant to the variance such that

$$E(\mathbf{X}) = \mathbf{A}E(\mathbf{S}) + E(\mathbf{E}) \quad (24)$$

$$Cov(\mathbf{X}) = \mathbf{A}Cov(\mathbf{S})\mathbf{A}^\top + Cov(\mathbf{E}) \quad (25)$$

Where $E(\cdot)$ denotes expectation and Cov the covariance. For a general Nth-order cumulant, we have

$$\mathcal{K}_{\mathbf{X}}^{(N)} = \mathcal{K}_{\mathcal{S}}^{(N)} \times_1 \mathbf{A} \times_2 \mathbf{A} \times \cdots \times_N \mathbf{A} + \mathcal{K}_{\mathbf{E}}^{(N)} \quad (26)$$

where $\mathcal{K}_{\mathbf{S}}^{(n)}$ is a diagonal matrix for independent $\mathbf{S}$. The ICA problem can potentially be uniquely solved by identifying $\mathbf{A}$ in the symmetric CP decomposition of any cumulants of order $N > 2$, which for the third- or fourth-order cumulant is given by

$$\mathcal{K}_{\mathbf{X}}^{(3)} \approx \mathcal{D} \times_1 \mathbf{A} \times_2 \mathbf{A} \times_3 \mathbf{A} \quad (27)$$

$$\mathcal{K}_{\mathbf{X}}^{(4)} \approx \mathcal{D} \times_1 \mathbf{A} \times_2 \mathbf{A} \times_3 \mathbf{A} \times_4 \mathbf{A}, \quad (28)$$

where $\mathcal{D}$ is a diagonal tensor. Generally speaking, it becomes harder to estimate cumulants from sample data as the order increases, i.e., longer datasets are required to obtain the same accuracy. Hence, in practice, the use of higher-order statistics is usually restricted to third- and fourth-order cumulants and because the third-order cumulants for symmetric distributions are zero, fourth-order cumulants are often used.[10]

The CP model has further proven useful for sensor array processing in wireless communication.[12,13] Here, the CP model provides powerful means for the exploitation of different types of diversity by admitting unique recovery. For the analysis of wireless communication, the following types of tensorial data have been analyzed

$$\mathcal{X}^{\text{Chip}\times\text{Symbol}\times\text{Antenna}}_{\text{DS-DCMA}} \approx \sum_{d}^{D^{\sharp Users}} \boldsymbol{a}_d^{\text{Chip}} \circ \boldsymbol{b}_d^{\text{Symbol}} \circ \boldsymbol{c}_d^{\text{Antenna}},$$

$$\mathcal{X}^{\text{Subarray}\times\text{Element}\times\text{Snapshot}}_{\text{MI-SAP}} \approx \sum_{d}^{D^{\sharp Sources}} \boldsymbol{a}_d^{\text{Subarray}} \circ \boldsymbol{b}_d^{\text{Element}} \circ \boldsymbol{c}_d^{\text{Snapshot}},$$

$$\mathcal{X}^{\text{FFT bin}\times\text{Symbol}\times\text{Antenna}}_{\text{MIMO-OFDM}} \approx \sum_{d}^{D^{\sharp\text{Trans. ante.}}} \boldsymbol{a}_d^{\text{FFT bin}} \circ \boldsymbol{b}_d^{\text{Symbol}} \circ \boldsymbol{c}_d^{\text{Antenna}}.$$

In direct-sequence code-division multiple access application (DS-DCMA), each user in theory contributes a rank-1 factor,[12] in multiple invariance sensor array processing application (MI-SAP), each source contributes a rank-1 factor,[13] whereas in multiuser multiple-input–multiple output orthogonal frequency-division multiplexing (MIMO-OFDM), each transmitting antenna contributes a rank-1 term.

Bioinformatics

Multiway modeling has recently found use within bioinformatics. Here, the HOSVD has been shown to enable the interpretation of cellular states and biological processes by defining the significance of each combination of extracted patterns, see also[19] and references therein. In Ref 19, the microarray data illustrated in Figure 8 was analyzed based on the following Tucker model

$$\mathcal{X}^{\text{Gene}\times\text{Time}\times\text{Condition}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}_l^{\text{Gene}} \circ \boldsymbol{b}_m^{\text{Time}} \circ \boldsymbol{c}_n^{\text{Condition}}. \qquad (29)$$

and it was demonstrated that HOSVD computationally can remove experimental artifacts from the global mRNA expression. In Ref 18, Tucker and CP analysis of a three-way array of $\mathcal{X}^{\text{Protein/Gene locus link}\times\text{Gene ontology category}\times\text{Osteogenic stimulant}}$ identified two distinct, stimulus-dependent sets of functionally related genes as they underwent osteogenic differentiation.

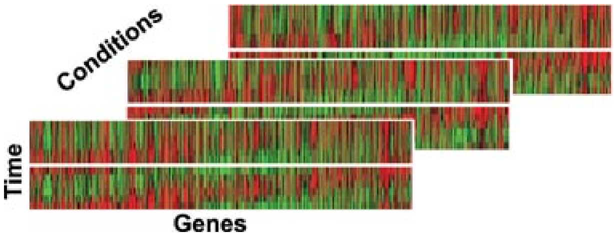

FIGURE 8 | Illustration of the three-way microarray data set used in the study of Ref 19.

In Ref 17, the Tucker model was applied to NMR time series data from studies of rats metabolic response to toxins. Three data sets were analyzed on the basis of NMR spectra of rat urine samples collected over several days after administration of a single dose of a model toxin in different doses, i.e.,

$$\mathcal{X}^{\text{Sample}\times\text{Spectra}\times\text{Time}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}_l^{\text{Sample}} \circ \boldsymbol{b}_m^{\text{Spectra}} \circ \boldsymbol{c}_n^{\text{Time}}. \qquad (30)$$

The Tucker analysis was here demonstrated to have the advantage of easily interpretable time profiles and extraction of metabolic perturbations with common time profiles only. The fields of bioinformatics and chemometrics heavily overlap. A good starting point for bioinformatics application of tensor decomposition is the review given in Ref 4.

Computer Vision

The use of the Tucker decompositions in computer vision was first proposed in the work on TensorFaces.[14] Here, facial image data from multiple subjects where each subject had multiple pictures taken under varying conditions was considered based on the Weizmann face database and recognition using TensorFaces proven to be significantly more accurate than standard PCA techniques. The analysis was based on the following Tucker compression

$$\mathcal{X}^{\text{Subj.}\times\text{View}\times\text{Illum.}\times\text{Expres.}\times\text{Pixel}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}^{\text{Subj.}} \circ \boldsymbol{b}^{\text{View}} \circ \boldsymbol{c}^{\text{Illum.}} \circ \boldsymbol{d}^{\text{Expres.}} \circ \boldsymbol{e}^{\text{Pixel}}. \qquad (31)$$

The data analyzed is illustrated in Figure 9. In Ref 15, a multilinear discriminant analysis (MDA) was proposed based on the Tucker representation and superior performance attained for a variety of image recognition tasks. In particular, it was demonstrated how multiple interrelated subspaces can collaborate to discriminate different classes and that the MDA algorithm can avoid the curse of dimensionality. A similar

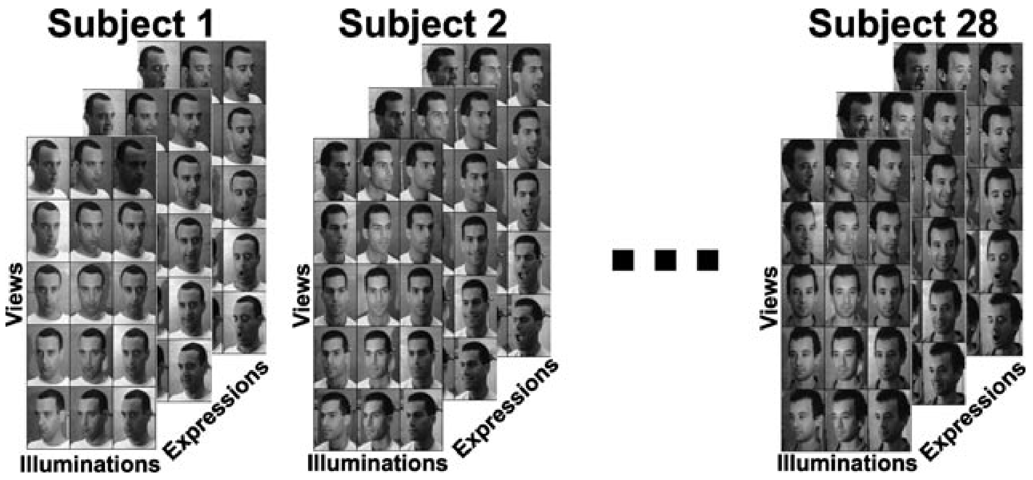

FIGURE 9 | Illustration of the Weizmann face database used in the analysis of TensorFaces.[14]

approach was proposed in Ref 53 inspired by features mimicking complex cells in V1 of the visual cortex. In Ref 16, the Tucker model has further proven useful for the identification of handwritten digits.

Web Mining

Chatroom communications arranged as a three-way array were analyzed in Ref 20 using tensor decomposition successfully capturing the underlying user group structure based on the following Tucker decomposition

$$\mathcal{X}^{\text{User}\times\text{Keyword}\times\text{Time}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}_l^{\text{Users}} \circ \boldsymbol{b}_m^{\text{Keywords}} \circ \boldsymbol{c}_n^{Time}. \qquad (32)$$

In Ref 21, the Enron e-mail collection was analyzed using DEDICOM based on e-mails between 184 users over 44 months, i.e.,

$$\mathcal{X}^{\text{User}\times\text{User}\times\text{Month}} \approx \sum_{d,d'} r_{d,d'} \boldsymbol{a}_d^{\text{User}} \circ \boldsymbol{a}_{d'}^{\text{User}} \circ \left(\boldsymbol{b}_d^{\text{Month}} * \boldsymbol{b}_{d'}^{\text{Month}}\right), \qquad (33)$$

where $*$ denotes element-wise multiplication. It was demonstrated that the decomposition had strong correspondence with known job classifications while revealing the patterns of communication between these roles while changes in the communication pattern over time, e.g., between top executives and the legal department became apparent. The Enron corpus contains messages between users with the structure (*to, from, text, time stamp*). The whole data can be represented as a fourth-order tensor of $\mathcal{X}^{\text{User}\times\text{User}\times\text{Time}\times\text{Terms}}$ and illustrates well how many data sets obtained from the web are multimodal in nature.

In Ref 22, multilinear algebra based on the CP model was used to analyze hyperlink graphs based on the anchor text of the hyperlinks between webpages, i.e.,

$$\mathcal{X}^{\text{Webpage}\times\text{Webpage}\times\text{Anchor text}} \approx \sum_{d} \boldsymbol{a}_d^{\text{Webpage}} \circ \boldsymbol{b}_d^{\text{Webpage}} \circ \boldsymbol{c}_d^{\text{Anchor text}}. \qquad (34)$$

and the model proven useful to automatically identify topics in the collection of webpages along with the associated authoritative web pages.

Furthermore, click through data were analyzed using the Tucker model in Ref 23 and shown to outperform the corresponding two-way approaches based on the decomposition

$$\mathcal{X}^{\text{User}\times\text{Query}\times\text{Webpage}} \approx \sum_{lmn} g_{l,m,n} \boldsymbol{a}_l^{\text{User}} \circ \boldsymbol{b}_m^{\text{Query}} \circ \boldsymbol{c}_n^{\text{Webpage}}. \qquad (35)$$

In Ref 54, the dynamic tensor analysis and streaming tensor analysis was proposed based on the Tucker model representation and it was demonstrated that the approach was useful for anomaly detection in network traffic as well as for multiway latent semantic indexing.

Finally, it is worth mentioning that the celebrated Netflix collaborative filtering challenge of predicting users ratings of movies (www.netflixprize.com) was not won until the data rather than analyzed solely as a (two-way) matrix $\boldsymbol{X}^{\text{Movie}\times\text{User}}$ of ratings was analyzed as a multiway array taking into account the temporal information, i.e., $\mathcal{X}^{\text{Movie}\times\text{User}\times\text{Time}}$.

The application of tensor factorization for web mining is rapidly growing and numerous studies have recently found that tensor decompositions are useful for learning the inherent relations between the modes of many types of web data born multimodal. A good account of some of the recent advances in tensor decomposition for web mining applications can be found in Ref 24.

CONCLUSION

Multiway analysis is rapidly growing in particular due to the storage capabilities and computational power of modern computers that admit analysis of large-scale multimodal data that arise in a multitude of scientific fields ranging from psychology, chemistry, neuroscience, signal processing, computer vision, and bioinformatics to the worldwide web. Although matrix decompositions have become key tools in practically all fields of science in order to comprehend and extract prominent features in data, it is our strong belief that also the described multiway decomposition methods will become key tools in order to investigate the many large-scale modern data sets that are born multimodal. A variety of examples of multiway applications have been given in this overview; however, the list of applications is in no way exhaustive and many types of new data applications are conceivable. In web mining, we saw how multiway analysis enabled to comprehend and identify associations in large corpora of data. Here, multiway analysis is also relevant in order to analyze multiple types of relations,[57] i.e., for author collaboration networks authors can co-author a paper, authors can cite other authors, authors can go to the same conferences, come from the same institution, and so on. Furthermore, these relations can change over time. In medicine, it is conceivable that multiway analysis will turn useful in order to mine patient journals, for instance, in the analysis of associations between patients, symptoms, diagnoses, and treatments and in market basket analysis to identify higher-order association between modes such as users, products, purchasing time, price, and geographic location. For all these data sets, tools borrowing on multiway analysis admit an analysis framework that explicitly takes the multimodal structures into account in order to identify the potential intrinsic relations between the modes of the data that might otherwise be missed. Furthermore, many types of data sets can benefit from designs that admit multiway analysis. As we saw taking measurements across samples of varying compound concentrations admitted unique recovery for fluorescence spectroscopy data but many other types of measurement data can benefit from a similar line of analysis. Often data contain repeated measures and rather than averaging measurements multiway analysis can introduce trial-dependent weights that can not only improve on the identification of the underlying signals by down-weighting noisy measurements but also potentially admit unique recovery as emphasized in the section on neuroscience applications. In the neuroscience and bioinformatics applications described, multiway analysis enabled extraction of the most consistent patterns of activation encompassing variability in strength and latency and as illustrated in the computer vision applications, multiway analysis form a promising framework for the identification of features that generalize well across various modes of variation. As a result, multiway/tensor analysis has many promising properties to offer for the analysis of large-scale multimodal data in general.

As we saw, tensors are not just matrices with additional subscripts. Tensors are objects with their own properties and tensor decomposition techniques enable the possibility of explicitly exploring structures formed by interaction between the modes. There is no doubt that analysis taking advantage of the multiway structure will help gain new knowledge of these many types of data and more adequately and effectively identify relationships between the modes of the data as well as consistent reproducible structures. However, care also has to be taken. Just because data have multiple modes does not necessarily imply that simple models such as the CP and Tucker models well account for the underlying dynamics in the data. For data compression and exploratory analysis, the basic models such as CP and Tucker can potentially facilitate an understanding of data that would otherwise be difficult to comprehend, whereas extensions of these basic factorizations has the potential for accommodating more complex structure and interaction in the data. Multiway decomposition is being applied to new fields every year and there is no doubt the future will bring many exiting applications and interesting extensions to the existing frameworks for analyzing data of more modalities than two.

ACKNOWLEDGEMENTS

The author would like to thank Lars Kai Hansen, Rasmus Bro, Tamara G. Kolda, and Lek-Heng Lim as well as two anonymous reviewers for very helpful comments. This overview article is

used for teaching multiway analysis at DTU Informatics at the Technical University of Denmark and teaching material and Matlab exercises are available from www.mortenmorup.dk.

REFERENCES

1. Kiers HA, Van Mechelen I. Three-way component analysis: principles and illustrative application. *Psychol Methods* 2001, 6:84–110.
2. Kroonenberg PM. *Applied Multiway Data Analysis*. New York: John Wiley & Sons, 2008.
3. Appellof CJ, Davidson ER. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Analyt Chem* 1981, 53:2053–2056.
4. Bro R. Review on multiway analysis in chemistry 2000–2005. *Critical Reviews in Analytical Chemistry* 2006, 36:279–293.
5. Andersen AH, Rayens WS. Structure-seeking multilinear methods for the analysis of fmri data. *Neuroimage* 2004, 22:728–739.
6. Mørup M, Hansen L, Arnfred S, Lim L-H, Madsen K. Shift-invariant multilinear decomposition of neuroimaging data. *NeuroImage* 2008, 42:1439–1450.
7. Miwakeichi F, Martnez-Montes E, Valds-Sosa PA, Nishiyama N, Mizuharam H, Yamaguchi Y. Decomposing eeg data into space time frequency components using parallel factor analysis. *Neuroimage* 2004, 22:1035–1045.
8. Mørup M, Hansen LK, Hermann CS, Parnas J, Arnfred SM. Parallel factor analysis as an exploratory tool for wavelet transformed event-related eeg. *Neuroimage* 2006, 29:938–947.
9. Comon P. Independent component analysis, a new concept? *Signal Process* 1994, 36:287–314.
10. De Lathauwer L, Vandewalle J. Dimensionality reduction in higher-order signal processing and rank-(r_1,r_2,...,r_n) reduction in multilinear algebra. *Lin Alg Appl* 2004, 391:31–55.
11. Comon P, Jutten C, Eds. *HANDBOOK OF BLIND SOURCE SEPARATION: Independent Component Analysis and Applications*. Elsevier, 2010.
12. Sidiropoulos ND, Member S, Giannakis GB, Bro R. Blind parafac receivers for ds-cdma systems. *IEEE Trans Signal Process* 2000, 48:810–823.
13. Sidiropoulos ND, Bro R, Giannakis GB. Parallel factor analysis in sensor array processing. *IEEE Trans Signal Process* 2000, 48:2377–2388.
14. Vasilescu MAO, Terzopoulos D. Multilinear analysis of image ensembles: Tensorfaces. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I* 2002, Springer-Verlag, 447–460.
15. Yan S, Xu D, Yang Q, Zhang L, Tang X, Zhang HJ. Multilinear discriminant analysis for face recognition. *Image Processing, IEEE Transactions on* 2007, 16:212–220.
16. Savas B, Eldén L. Handwritten digit classification using higher order singular value decomposition. *Pattern Recogn* 2007, 40:993–1003.
17. Dyrby M, Baunsgaard D, Bro R, Engelsen SB. Multiway chemometric analysis of the metabolic response to toxins monitored by nmr. *Chemom Intell Lab Systems* 2005, 76:79–89.
18. Yener B, Acar E, Aguis P, Bennett K, Vandenberg S, Plopper G. Multiway modeling and analysis in stem cell systems biology. *BMC Systems Biology* 2008, 2:63.
19. Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffley, JFX, Alter O. Global effects of dna replication and dna replication origin activity on eukaryotic gene expression. *Molecular Systems Biology* 2009, 5.
20. Acar E, Camtepe SA, Krishnamoorthy MS, Yener B. Modeling and multiway analysis of chatroom tensors. *Intelligence and Security Informatics, Lecture Notes in Computer Science* 2005, 3495:256–268.
21. Bader BW, Harshman RA, Kolda TG. Temporal analysis of semantic graphs using asalsan. In *ICDM* 2007, IEEE Computer Society, 33–42.
22. Kolda TG, Bader BW, Kenny JP. Higher-order web link analysis using multilinear algebra. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, 2005, 242–249.
23. Sun J-T, Zeng H-J, Liu H, Lu Y, Chen Z. Cubesvd: a novel approach to personalized web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*. ACM Press, New York, 2005, 382–390.
24. Kolda TG, Bader BW. Tensor decompositions and applications. *SIAM Review*.
25. Cichocki A, Zdunek R, Phan AH, Amari S-i. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multiway Data Analysis and Blind Source Separation*. New York: John Wiley & Sons, 2009.
26. Comon P, Luciani X, de Almeida ALF. Tensor decompositions, alternating least squares and other tales. *J Chemom* 2009, 23:393–405.
27. Tomasi G. *Practical and computational aspects in chemometric data analysis* PhD thesis, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark, May 2006.
28. Smilde A, Bro R, Geladi P. *Multiway Analysis: Applications in the Chemical Sciences*. New York: John Wiley & Sons, 2004.
29. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966, 31:279–311.

30. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 1970, 35:283–319.

31. Harshman RA. Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Working Papers in Phonetics* 1970, 16:1–84.

32. Lathauwer LD, Moor BD, Vandewalle J. Multilinear singular value decomposition. *SIAM J MATRIX ANAL APPL*, 2000, 21:1253–1278.

33. Hitchcock FL. Multiple invariants and generalized rank of a p-way matrix or tensor. *J Math Phys Camb* 1927, 39–70.

34. Kruskal JB. More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* 1976, 41:281–293.

35. Kruskal J. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl* 1977, 18:95–138.

36. Lim L-H, Comon P. Multiarray signal processing: tensor decomposition meets compressed sensing. *Comptes Rendus de l'Académie des sciences, to appear* 2010.

37. Sidiropoulos ND, Bro R. On the uniqueness of multilinear decomposition of *n*-way arrays. *J Chemom* 2000, 14:229–239.

38. Harshman RA, Lundy ME. Data preprocessing and the extended parafac model. *In: Law HG, Snyder, Jr. CW, Hattie JA, McDonald RP. (eds.), Research Methods for Multimode Data Analysis, Praeger, New York,* 1984, 216–281.

39. de Silva V, Lim L-H. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J Matrix Anal Appl* 2008, 30:1084–1127.

40. Acar E, Dunlavy DM, Kolda TG, Mørup M. Scalable tensor factorizations for incomplete data. arXiv:1005.2197v1, 2010.

41. Mørup M, Hansen LK. Automatic relevance determination for multiway models. *J Chemometrics* 2009, 23,352–363.

42. Bro R, Kiers HAL. A new efficient method for determining the number of components in parafac models. *J Chemom* 2003, 17:274–286.

43. Bro R, Kjeldahl K, Smilde AK, Kiers HAL. Cross-validation of component models: A critical look at current methods. *Anal Bioanal Chem* 2008, 390:1241–1251.

44. Acar E, Yener B. Unsupervised multiway data analysis: a literature survey. *IEEE Trans Knowl Data Eng* 2009, 21:6–20.

45. Bro R, Andersson CA. The *n*-way toolbox for matlab. *Chemom Intell Lab Systems* 2000, 52:1–4.

46. Bader BW, Kolda TG. Algorithm 862: Matlab tensor classes for fast algorithm prototyping. *ACM Trans Math Softw* 2006, 32:635–653.

47. Möcks J. Topographic components model for event-related potentials and some biophysical considerations. *IEEE Trans Biomed Eng* 1988, 35:482–484.

48. Cattell R. The three basic factor-analytic research designs – their interrelations and derivatives. *Psychol Bull* 1952, 49:499–520.

49. Murakami T, Kroonenberg PM. Three-mode models and individual differences in semantic differential data. *Multivariate Behav Res* 2003, 38:247–283.

50. Mørup M, Hansen L, Arnfred SM. Erpwavelab a toolbox for multi-channel analysis of time-frequency transformed event related potentials. *J Neurosci Methods* 2007, 161:361–368.

51. Cichocki A, Pha A-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences E92-A* 2009, 3:708–721.

52. Comon P. *Tensor Decomposition* Mathematics in Signal Processing V. Clarendon Press, Oxford, UK, 2002, 1–24.

53. Mu Y, Tao D, Li X, Murtagh F. Biologically inspired tensor features. *Cognit Comput* 2009, 1:327–341.

54. Sun J, Tao D, Faloutsos C. Beyond streams and graphs: dynamic tensor analysis. In *In KDD* 2006, 374–383.

55. De Lathauwer L. Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM J Matrix Anal Appl* 2008, 30:1033–1066.

56. Harshman R, Lundy M. Uniqueness proof for a family of models sharing features of tucker's three-mode factor analysis and parafac/candecomp. *Psychometrika* 1996, 61:133–154.

57. Kemp C, Tenenbaum JB, Griffiths TL, Yamada T, Ueda N. Learning systems of concepts with an infinite relational model. Proc. 21st National Conference on Artificial Intelligence (AAAI-06.)